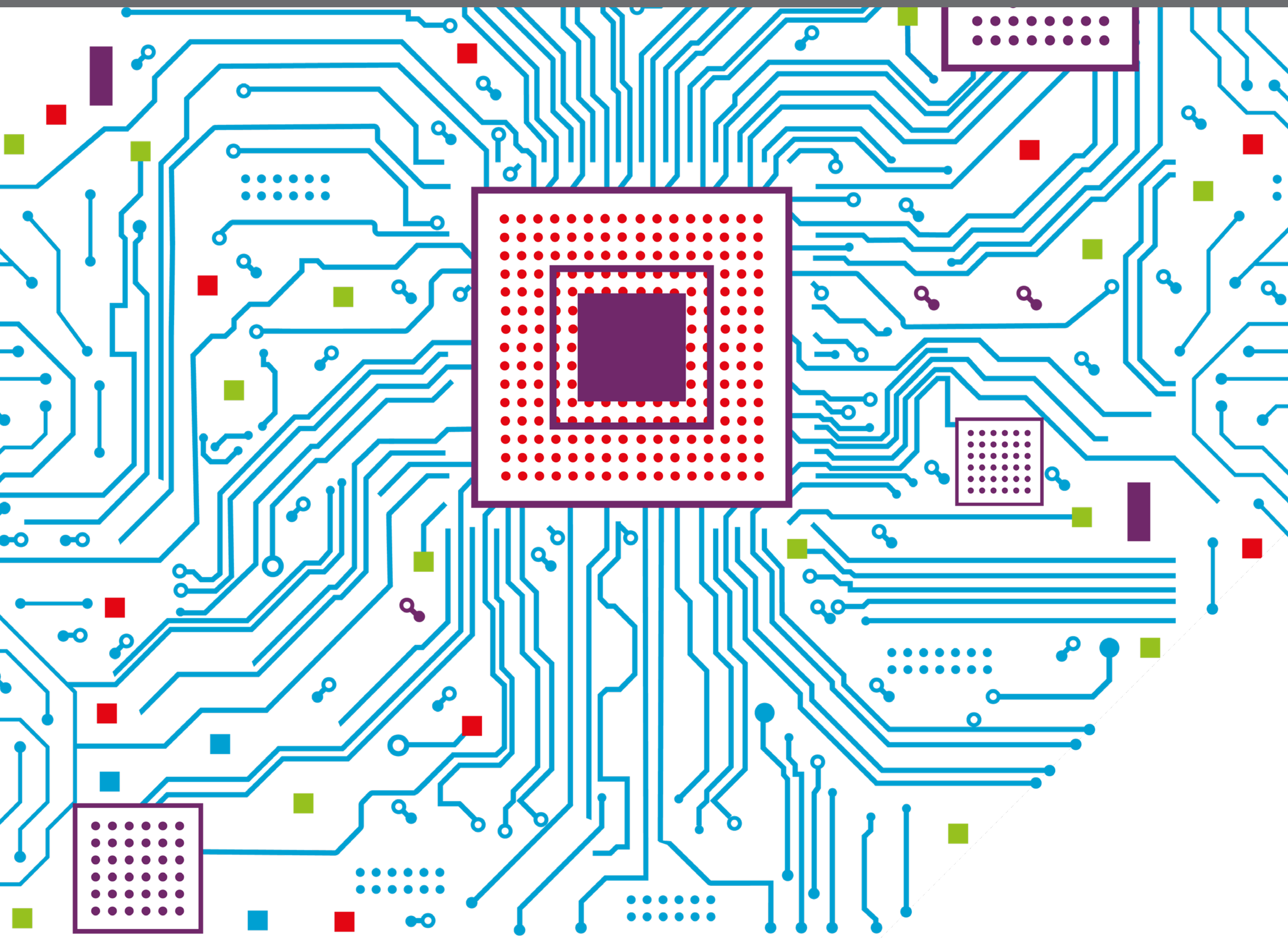
TOWARDS OMNIPRESENT AND SMART SPEECH ASSISTANTS

EDITED BY: Ingo Siegert, Stefan Hillmann, Benjamin Weiss,
Jessica Martina Szczuka and Alexey Karpov
PUBLISHED IN: Frontiers in Computer Science, Frontiers in Psychology and
Frontiers in Communication



frontiers Research Topics

Frontiers eBook Copyright Statement

ISSN 1664-8714
ISBN 978-2-88976-641-3
DOI 10.3389/978-2-88976-641-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

TOWARDS OMNIPRESENT AND SMART SPEECH ASSISTANTS

Topic Editors:
Ingo Siegert, Otto von Guericke University Magdeburg, Germany
Stefan Hillmann, Technische Universität Berlin, Germany
Benjamin Weiss, Technical University of Berlin, Germany
Jessica Martina Szczuka, University of Duisburg-Essen, Germany
Alexey Karpov, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Russia

Citation: Siegert, I., Hillmann, S., Weiss, B., Szczuka, J. M., Karpov, A., eds. (2022). Towards Omnipresent and Smart Speech Assistants. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-641-3

Table of Contents

frontiers | Frontiers in Computer Science

EDITORIAL
published: 30 June 2022
doi: 10.3389/fcomp.2022.966163

Editorial: Towards Omnipresent and Smart Speech Assistants

Ingo Siegert[1], Stefan Hillmann[2], Benjamin Weiss[2], Jessica M. Szczuka[3] and Alexey Karpov[4]*

[1] Mobile Dialog Systems, Institute for Information Technology and Communications, Otto von Guericke University Magdeburg, Magdeburg, Germany, [2] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany, [3] Social Psychology: Media and Communication, University of Duisburg-Essen, Duisburg, Germany, [4] St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia

Keywords: smart assistant, dialog, voice user interface, user experience, conversational interfaces

Editorial on the Research Topic

Towards Omnipresent and Smart Speech Assistants

OPEN ACCESS

Edited and reviewed by:
Anton Nijholt,
University of Twente, Netherlands

****Correspondence:***
Ingo Siegert
ingo.siegert@ovgu.de

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

Received: *10 June 2022*
Accepted: *16 June 2022*
Published: *30 June 2022*

Citation:
Siegert I, Hillmann S, Weiss B, Szczuka JM and Karpov A (2022) Editorial: Towards Omnipresent and Smart Speech Assistants. Front. Comput. Sci. 4:966163. doi: 10.3389/fcomp.2022.966163

1. INTRODUCTION

The functionality of digital voice assistant systems has been constantly increasing during the last decade and a lot of commercial systems are available. Driven by their ease of use, the attractiveness of such devices is constantly growing, and they allow conducting online searches and orders as well as smart home services by simply calling up the device (de Barcelos Silva et al., 2020; Dutsinma et al., 2022).

However, the implications of voice-based interaction are not always clear to the user, ranging from its functionality to the impact of speech as a social cue for resulting psychological effects. In the future, however, they should not only process simple commands, but also enable a natural and smooth interaction and be omnipresent. In addition to an improved speech recognition, this will require enhanced speech understanding and more intelligent dialog guidance.

While state-of-the art systems are mainly conceptualized for young adults and middle-aged people, future systems should adapt to the user in order to meet the needs of different (vulnerable) user groups, ranging from young children to the elderly. This will be accompanied by efforts to make systems more understandable and users more sophisticated. Consequently, legal aspects resulting from the spread of voice assistants and the stricter data protection regulations are important.

The goal of this Research Topic was to present the latest advances—both from academia and industry—in the area of voice assistants. It was aimed at collecting research contributions from the disciplines of human-computer interaction, artificial intelligence, and human factors in order to promote interdisciplinary collaborations and cross-fertilization of ideas. More specifically, we were interested in exploring the current landscape and future directions for the emerging topic of voice assistants. The Research Topic covers 11 articles from 34 different authors from different research fields, including linguistics, psychology, usability/user experience studies as well as the technical perspective. One apparent focus of this Research Topic was on analyzing and assessing user experience. Both, different user groups and situations are taken into account. However, we hope to see the aforementioned perspective on more sophisticated dialogs represented in the near future.

2. CONTRIBUTIONS

Cao et al. investigate how mind-based anthropomorphism influences users exploratory usage of intelligent personal assistants (IPA). The article describes a study collecting more than 500 valid answered questionnaires, and the results on the influence of cognitive and affective anthropomorphism on IPA self-efficacy and the user's social connection to the IPA.

Carolus et al. show in an online laboratory experiment that participants have empathy with a smart speaker, when watching videos of a user interacting with such a device. This claims a rather universal effect, as the results are independent of the participants' gender or usage experience, and thus expands the current body of empirical results around the Media Equation (Reeves and Nass, 1996).

Cohn et al. investigate users speech rate adjustments during conversations with an Amazon Alexa social bot in at-home and in-lab settings, considering automatic speech recognition (ASR) comprehension errors. It is found that users used a slower speech rate when talking to the bot, which is even more slowed down in the in-lab setting (relative to at-home).

Cohn and Zellou present the results of a study on differences in speech adaptations (e.g., speech rate, f0 mean, and f0 variation) during pre-scripted spoken interactions with a voice-AI assistant and a human interlocutor. The authors measured a decreased speech rate, higher average fundamental frequency (f0), and greater f0 variation for the device directed speech.

Frommherz and Zarcone collected ecologically-valid German dialog data via a crowdsourcing approach in the Wizard-of-Oz (WOZ) setting. Compared with the MultiWOZ dataset, their method for data collection has led to considerably less scripting and priming in the collected dialog data.

Hirsch presents a local and low-cost, low-energy voice assistant solution including a keyword recognition algorithm and a further client system without the need of an external power supply. This is the most relevant applied work, of a privacy-ensuring home speech assistant, among all the articles.

Mavrina et al. describe a longitudinal field study on communication breakdowns between family members and a voice assistant. Their article provides qualitative analysis of particularly interesting breakdown cases, as well as statistical analysis combining empirical and conversational data collected with children and adults during 5 weeks of free interaction with a voice assistant device.

Schlomann et al. present their opinion regarding elderly with and without cognitive disabilities. Their main argument is to raise the potential of speech assistants for elderly users by participatory design methods and verify the approaches by field studies.

Schreibelmayr and Mara conducted a randomized laboratory experiment on synthetic voices with 165 participants to explore what level of human-like realism human-interactors prefer, whether the participants evaluations vary across different domains of application, and if the listener's personality has an impact on the ratings.

Wienrich and Carolus have developed an instrument called "conversational agent literacy scale" (CALS), to measure conceptualizations and competencies about conversational agents in human users. This scale consists of five sub-scales and is based on a study with 170 participants.

Wienrich et al. found in a laboratory study that a voice assistant designed as a "specialist" is rated as more trustworthy by the users than a "generalist" in the health domain. By providing both, a theoretical line of reasoning and empirical data, the study lays the pathway for further studies on the users perspective on trustworthiness in voice-based systems.

3. CONCLUSION

In conclusion, this Research Topic comprises interdisciplinary contributions and gives some examples of both theoretical and practical implications for smart voice/speech assistants. Topics reach from laboratory studies on empathy or speaking behavior adjustments over field studies on communication breakdowns, to the description of a local client voice assistant system. It therefore reflects the diversity of this strongly developing field of research. However, the contributions also highlight unresolved questions in current research, e.g., pitfalls due to design and field study issues or a lack of studies regarding trust or acceptance.

We are aware that there is a plethora of further aspects that need to be addressed to complete, in the best sense, the aim of a human-like interaction with voice assistants for all kind of humans. The articles of this Research Topic paving the way to an understanding of the role of voice assistants and thus, in the future, voice assistants can be an integral part of our daily life in terms of a true intelligent assistant.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We would like to thank all the authors, reviewers, and editors who worked so diligently to help us bring this collection together.

REFERENCES

de Barcelos Silva, A., Gomes, M. M., da Costa, C. A., da Rosa Righi, R., Barbosa, J. L. V., Pessin, G., et al. (2020). Intelligent personal assistants: a systematic literature review. *Expert Syst. Appl.* 147, 113193. doi: 10.1016/j.eswa.2020.113193

Dutsinma, F. L. I., Pa, D., Funilkul, S., and Chan, J. H. (2022). A systematic review of voice assistant usability: An ISO 9241-11 approach. *SN Comput. Sci.* 3, 267. doi: 10.1007/s42979-022-01172-3

Reeves, B., and Nass, C., editors (1996). *The Media Equation: How People Treat Computers, Television, and New*

Media like Real People. Cambridge, UK: University Press Cambridge.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Communication

BRIEF RESEARCH REPORT
published: 07 May 2021
doi: 10.3389/fcomm.2021.671429

Speech Rate Adjustments in Conversations With an Amazon Alexa Socialbot

Michelle Cohn[1,2]*, Kai-Hui Liang[2,3], Melina Sarian[1], Georgia Zellou[1] and Zhou Yu[2,3]

[1]Phonetics Lab, University of California, Davis, CA, United States, [2]Natural Language Processing Lab, University of California, Davis, CA, United States, [3]Department of Computer Science, Columbia University, New York, NY, United States

This paper investigates users' speech rate adjustments during conversations with an Amazon Alexa socialbot in response to situational (in-lab vs. at-home) and communicative (ASR comprehension errors) factors. We collected user interaction studies and measured speech rate at each turn in the conversation and in baseline productions (collected prior to the interaction). Overall, we find that users slow their speech rate when talking to the bot, relative to their pre-interaction productions, consistent with hyperarticulation. Speakers use an even slower speech rate in the in-lab setting (relative to at-home). We also see evidence for turn-level entrainment: the user follows the directionality of Alexa's changes in rate in the immediately preceding turn. Yet, we do not see differences in hyperarticulation or entrainment in response to ASR errors, or on the basis of user ratings of the interaction. Overall, this work has implications for human-computer interaction and theories of linguistic adaptation and entrainment.

Keywords: vocal entrainment, socialbot, voice-activated artificially intelligent assistant, non-task oriented conversations, human-computer interaction, hyperarticulation

OPEN ACCESS

Edited by:
Ingo Siegert,
Otto von Guericke University Magdeburg, Germany

Reviewed by:
Nichola Anne Lubold,
Arizona State University, United States
Donna Mae Erickson,
Haskins Laboratories, United States

***Correspondence:**
Michelle Cohn
mdcohn@ucdavis.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *23 February 2021*
Accepted: *21 April 2021*
Published: *07 May 2021*

Citation:
Cohn M, Liang K-H, Sarian M, Zellou G and Yu Z (2021) Speech Rate Adjustments in Conversations With an Amazon Alexa Socialbot. Front. Commun. 6:671429. doi: 10.3389/fcomm.2021.671429

INTRODUCTION

After their introduction in the 2010s, there has been a widespread adoption of voice-activated artificially intelligent (voice-AI) assistants (e.g., Google Assistant, Amazon's Alexa, Apple's Siri), particularly within the United States (Bentley et al., 2018). Millions of users now speak to voice-AI to complete daily tasks (e.g., play music, turn on lights, set timers) (Ammari et al., 2019). Given their presence in many individuals' everyday lives, some researchers have aimed to uncover the cognitive, social, and linguistic factors involved in voice-AI interactions by examining task-based interactions with voice-AI (e.g., setting an appointment on a calendar in Raveh et al., 2019), scripted interactions in laboratory settings (Cohn et al., 2019; Zellou et al., 2021), and interviews to probe how people perceive voice-AI (Lovato and Piper, 2015; Purington et al., 2017; Abdolrahmani et al., 2018). Yet, our scientific understanding of *non-task based*, or purely social, interactions with voice-AI is even less established.

Since 2017, the Amazon Alexa Prize competition has served as a venue for social chit-chat between users and Amazon Alexa socialbots on any Alexa-enabled device; with a simple command, "Alexa, let's chat", any user can talk to one of several university-designed socialbots (Chen et al., 2018; Ram et al., 2018; Gabriel et al., 2020; Liang et al., 2020). Do individuals talk to these socialbots in similar ways as they do with real humans? The *Computers are Social Actors* (CASA; Nass et al., 1997; Nass et al., 1994) framework proposes that people apply socially mediated, 'rules', from human-human interaction to computers when they detect a cue of 'humanity' in the system. Voice-AI systems are already imbued with multiple human-like features: they have names, apparent genders Habler et al. (2019) and interact with users using spoken language. Indeed, there is some evidence that individuals engage with voice-AI in ways that parallel the ways they engage with humans (e.g., gender-asymmetries in phonetic alignment in Cohn et al., 2019; Zellou et al., 2021). In the case of voice-AI *socialbots*, the cues of humanity could be even more robust since the system is designed for social interaction.

To uncover some of the cognitive and linguistic factors in how users perceive voice-AI socialbots, the current study examines two speech behaviors: 'hyperarticulation' and 'entrainment'. We define 'hyperarticulation' as carefully articulated speech (also referred to as 'clear' speech; Smiljanić and Bradlow, 2009), thought by listener-oriented accounts to be tailored specifically to improve intelligibility for an interlocutor in the conversation (Lindblom, 1990). For example, there is a body of work examining acoustic adjustments speakers make when talking to computer systems, or 'computer-directed speech' (computer-DS) (Oviatt et al., 1998a; Oviatt et al., 1998b; Bell and Gustafson, 1999; Bell et al., 2003; Lunsford et al., 2006; Stent et al., 2008; Burnham et al., 2010; Mayo et al., 2012; Siegert et al., 2019). A common listener-oriented hyperarticulation is to slow speaking rate, produced in response to background noise (Brumm and Zollinger, 2011), as well as in interactions with interlocutors assumed to be less communicatively competent, such as computers (Oviatt et al., 1998b; Stent et al., 2008), infants (Fernald and Simon, 1984), and non-native speakers (Scarborough et al., 2007; Lee and Baese-Berk, 2020). Will users also slow their speech rate when they talk to a socialbot? One possibility that the advanced speech capabilities in Alexa socialbots (in terms of speech recognition, language understanding and generation) might lead to more naturalistic interactions, whereby users talk to the system more as they would an adult human interlocutor. Alternatively, there is work showing that listeners rate 'robotic' text-to-speech (TTS) voices as less communicatively competent than more human-like voices (Cowan et al., 2015) and that listeners perceive prosodic peculiarities in the Alexa voice, describing it as being 'monotonous' and 'robotic' (Siegert and Krüger, 2020). Accordingly, an alternative prediction is that speakers will use a slower speaking rate when talking to the Alexa socialbot, since robotic voices are perceived as being less communicatively competent.

In addition to hyperarticulation, we examine 'entrainment' (also known as 'accommodation', 'alignment', or 'imitation'): the tendency for speakers to adopt their interlocutor's voice and language patterns. For example, a speaker might increase their speech rate in response to hearing the socialbot's speech rate increase. Entrainment has been previously observed both in human-human (Levitan and Hirschberg, 2011; Babel and Bulatov, 2012; Lubold and Pon-Barry, 2014; Levitan et al., 2015; Pardo et al., 2017) and human-computer interaction (Coulston et al., 2002; Bell et al., 2003; Branigan et al., 2011; Fandrianto and Eskenazi, 2012; Thomason et al., 2013; Cowan et al., 2015; Gessinger et al., 2017; Gessinger et al., 2021), suggesting it is a behavior transferred to interactions with technology. Recent work has shown that entrainment occurs in interactions with voice-AI assistants as well (Cohn et al., 2019; Raveh et al., 2019; Zellou et al., 2021). Like hyperarticulation, there are some accounts proposing that entrainment improves intelligibility (Pickering and Garrod, 2006), aligning representations between interlocutors. For example, people entrain toward the lexical and syntactic patterns of computers, lessening (presumed) communicative barriers (Branigan et al., 2011; Cowan et al., 2015). At the same time, entrainment can also reveal social attitudes: social accounts of alignment propose that people converge to convey social closeness and diverge to signal distance (Giles et al., 1991; Shepard et al., 2001), such as entraining more to interlocutors they like (Chartrand and Bargh, 1996; Levitan et al., 2012). In the current study, we predict that speakers who rate the socialbot more positively will also show more entrainment toward it.

While the vast majority of prior work examines hyperarticulation and entrainment separately (e.g., Burnham et al., 2010; Cohn et al., 2019), the current study models these behaviors in tandem. This is important as hyperarticulation and entrainment might both result in the same observed behavior: a speaker might speak slower when talking to the socialbot overall (hyperarticulation), but also slow in response to a slower speech rate by the bot (entrainment). Including both in the same model allows us to attribute observed behavior to its underlying cognitive processes. This is also important as hyperarticulation and entrainment might, at times, conflict (e.g., slowing overall speech rate, but entraining to the faster rate of the bot). Additionally, including both measures in the same model can directly test the extent hyperarticulation and entrainment are mediated by functional pressures (e.g., speech recognition errors) and social-situational pressures (e.g., presence of an experimenter).

Functional Factors in Hyperarticulation and Entrainment

How might hyperarticulation and entrainment vary as a function of intelligibility pressures that change dynamically within a conversation? Automatic speech recognition (ASR) mistakes are common in a spontaneous interaction with a voice-AI system. The present study investigates whether turn-by-turn dynamics of hyperarticulation and entrainment vary based on whether the Alexa system makes a comprehension error or not. There is a rich literature examining hyperarticulation toward computer interlocutors in response to an error made by the system (Oviatt and VanGent, 1996; Oviatt et al., 1998b; Bell and Gustafson, 1999; Swerts et al., 2000; Vertanen, 2006; Stent et al., 2008; Maniwa et al., 2009; Burnham et al., 2010). For example, Stent et al. (2008) found that speakers' increased hyperarticulation in response to an ASR error lingered for several trials before 'reverting' back to their pre-error speech patterns; in the present study, we similarly predict slower speech rate following an ASR error. While less examined than hyperarticulation, there is some evidence suggesting that entrainment also serves a functional role (Branigan et al., 2011; Cowan et al., 2015); for example, participants show more duration alignment if their interlocutor made an error

(Zellou and Cohn, 2020). Thus, we might also predict greater entrainment following an error, relative to pre-error.

Situational Factors in Hyperarticulation and Entrainment

How might context shape speech hyperarticulation and entrainment toward an Alexa socialbot? In the current study, half of the participants interacted with the socialbot in-person in a laboratory setting with experimenters present, while the other half interacted at home[1] using the Amazon Alexa app. While many studies of voice-AI are conducted in a laboratory setting (e.g., Cohn et al., 2019; Zellou et al., 2021), there is evidence that the presence of an experimenter influences how participants complete a task (Orne, 1962; Belletier et al., 2015; Belletier and Camos, 2018). Indeed, *Audience Design* theory proposes that people tailor their speech style for their intended addressee, as well as for 'overhearers' (i.e., individuals listening to the conversation, but not directly taking part) (Clark and Carlson, 1982). For example, speakers are more polite when there is a bystander present (Comrie, 1976). As a result, we might predict more careful, hyperarticulated speech in a lab setting with overhearers. Prior work has also shown that engaging with additional interlocutors shapes entrainment: Raveh et al. (2019) found that speakers entrained less toward an Alexa assistant if they had interacted with a third interlocutor (a human confederate), compared to dyadic interactions only between the user and Alexa. Therefore, we might predict that participants will display less entrainment in the laboratory setting (relative to at-home).

METHODS

In the current study, we use a socialbot system originally designed for Amazon Alexa Prize (Chen et al., 2018; Liang et al., 2020). In-lab user studies were conducted on the same day (pre-social isolating measures) in a quiet room. At-home user studies occurred across nine days in April-June, where speakers participated in an online experiment, activating the socialbot from home and recording their interaction with their computer microphone in a quiet room.

Participants

Participants (n = 35) were native English speakers, recruited from UC Davis (mean age = 20.94 years old ±2.34; age range 18–30 years; 22 female, 13 male). The in-lab user condition, consisting of 17 participants (mean age = 20.76 years ±2.66; 14 female, 3 male). An additional 18 participants (mean age = 21.11 years ±2.03, 9 female, 9 male) completed an at-home user condition. A t-test revealed that there was no significant difference in ages between these groups [t (29.9) = −0.43, p = 0.67]. Nearly all participants (34/35) reported using voice-AI assistants in the past. All participants consented to the study (following the UC Davis Institutional Review Board) and received course credit for their participation.

Procedure

In-lab participants completed the experiment in a quiet room, with an Amazon Echo located in front of them on a table. Their interactions were recorded using a microphone (Audio-Technica AT 2020) facing the participant. An experimenter initiated the socialbot, and 1-2 experimenters were present in the room to listen to the conversation. Those in the at-home condition completed the experiment online via a Qualtrics survey which was used to record their speech (via AddPipe[2] and their computer microphone). For the at-home condition, participants were given instructions to install the Alexa app to their phones and activate a Beta version of the socialbot.

All participants began with a baseline recording of an utterance: "The current month is [current month]. Test of the sound system complete." Then, they initiated the socialbot conversation and were instructed to have two conversational interactions with the system for roughly 10 min each (see **Table 1** for an example excerpt). If the bot crashed before the 10 min, they were asked to re-engage the Alexa Skill again. Dialogue flows included multiple domains (e.g., movies, sports, animals, travel, food, music, and books), as well as general chit-chat and questions about Alexa's 'backstory' (e.g., favorite color, animal, etc.) (Chen et al., 2018; Liang et al., 2020). At the end of the interaction, participants rated the Alexa socialbot across three dimensions, on a scale of 1–5: "How engaging did you find the bot? 1 = not engaging, 5 = extremely engaging", "How likely would you talk to the bot again? 1 = not likely, 5 = extremely likely", "How coherent was the bot? 1 = not coherent, 5 = extremely coherent".

Acoustic Analysis

Baseline and conversation recordings were initially transcribed with Amazon ASR or Sonix[3]. Trained research assistants confirmed the accuracy of the transcripts and annotated the sound files in a Praat Textgrid (Boersma and Weenink, 2018), labeling the interlocutor turns and the presence of ASR errors made by the socialbot. Errors included 'long pause' errors, such as when the socialbot took a long pause and then used an interjection or responded with phrases like "Tik tok! Did I confuse you?" or "Are you still there?" Other ASR errors included when the socialbot responded with a different word or topic than what the user mentioned. For instance, when the user said they were watching tv shows recently, the socialbot responded with "Great! I love talking about sports . . . "

We analyze only the first continuous conversation with Alexa in order to assess differences from baseline to the bot interaction, rather than differences between bot conversations. On average, participants spoke with the socialbot for 12.48 min (sd = 5.44) including 96.63 total turns (user + Alexa) (sd = 44.58). The

[1]Due to the COVID-19 pandemic, all in-lab user studies were paused in Spring 2020.

[2]https://addpipe.com/product

[3]http://sonix.ai/

TABLE 1 | Excerpt from a socialbot conversation.

Alexa turn	User turn
Hi! this is an Alexa prize social bot. It's really good to meet you! Do you want to tell me your name?	I'm [anonymized name]
Hi, [anonymized name]. How are you doing today?	I'm good how are you
Nice! I hope you're having a good day! I'm really good! Thanks for asking. I'm curious, What do you like to do for fun?	I like to read
Cool! I love to read when I'm not chatting. Have you read any good books recently?	Let me think
Sure, take your time	I re-read harry potter recently
Ah, I believe you're talking about Harry Potter and the Philosopher's stone. How would you rate this book on a scale from 1 to 10?	Good question
I see. I think for me, I really liked the book! Maybe you will find this interesting. that international master Jeremy Silman created the chess puzzle for Harry Potter and the Philosopher's stone where Ron sacrifices himself, but wasn't credited for his role, while "everyone from the hairdressers to the donut delivery boy" were. What do you think?	That's interesting I didn't know that
Yeah. It is! Sometimes one passage in a book can really stick with you. What was your favorite part of this book?	I really like the part where they are on the rock in the middle of the ocean episode of book

socialbot made an average of 6.94 errors per conversation (sd = 5.57). T-tests confirmed no difference in conversation duration [t (29.52) = −1.03, p = 0.31], overall number of turns [t (29.58) = −0.90, p = 0.37], or number of errors [t (23.49) = −0.23, p = 0.82] across in-lab and at-home contexts. In total, the corpus consisted of 1,417 productions by the human users.

Speech rate (mean number of syllables per second) was measured using a Praat script (De Jong et al., 2017) for each of the socialbot's turns, user's turns, and the user's baseline productions. To measure differences in hyperarticulation in talking to the Alexa socialbot, we centered each user's turn-level speaking rate relative to their baseline production (i.e., subtracting all 'speech rate' values by the user's average baseline speech rate). This centered value is then used to ascertain change from a user's baseline. For instance, a positive value indicates an increase in speaking rate from baseline.

To measure entrainment, we test 'synchrony' (Coulston et al., 2002; Levitan & Hirschberg, 2011): how speakers synchronize their productions across turns. For instance, when the Alexa produces a relatively faster speaking rate, does the user *also* show a relative increase in speaking rate? We used the user's turn-level rate measurements (centered within user) and also centered the Alexa's productions (subtracting the mean speaking rate of Alexa's overall values for each conversation). Accordingly, comparing the 'Alexa-prior turn' (centered) and user's value (centered) can capture whether users adjust their speech to match the directionality of change. Additionally, this method allows us to compare *both* hyperarticulation and entrainment in the same model, with the dependent variable of the (centered) user's speaking rate.

STATISTICAL ANALYSIS AND RESULTS

Ratings

A t-test revealed that the Alexa was rated as more engaging in the at-home condition (mean = 4.10) relative to the in-lab condition (mean = 3.35) [t (31.84) = 2.52, $p < 0.05$]. There was no significant difference in ratings of how coherent the bot was [t (30.52) = 0.83, p = 0.41] or in how much the participant would want to talk to the bot again [t (30.01) = −1.52, p = −0.14] based on situational context. We calculated an overall ratings value, summing users' ratings for engagement, coherence, and desire to talk to the bot again (mean = 11.30, range = 7–14) to use in the statistical model on speaking rate change.

Users' Baseline Productions and Alexa Productions Across Context

Mean values for speaking rate of the user's baseline productions, users' responses to the socialbot, as well as the socialbot's productions are provided in **Table 2**. As seen, there were differences in the baseline productions based on setting, where speakers produced slower rate in-lab in their baseline production. The Alexa productions had a faster speech rate in-lab (relative to at-home)[4].

Hyperarticulation and Entrainment

We modeled speech rate (centered within user) with a linear mixed effects model using the *lme4* R package (Bates et al., 2015). Fixed effects included Setting (2 levels: in-lab, at-home), Overall Rating (coherence + satisfaction + engagement, centered), and all possible two-way interactions with Alexa Prior Turn Rate (continuous, centered). We additionally added Gender as a fixed effect (2 levels: female, male)[5]. Random effects included by-User random intercepts[6]. Categorical contrasts were sum coded.

The model showed a significant negative intercept, indicating that users' speaking rate decreases (i.e., fewer syllables/second) in the socialbot interactions relative to baseline productions [$Coef$ = −0.62, t = −5.96, $p < 0.001$]. Additionally, there was a main effect of Setting, shown in **Figure 1**: speakers produced an even slower speech rate in-lab, relative to at-home [$Coef$ = −0.37, t = −3.59, p <

[4]Differences in the social bot speech rate reflect the un-scripted nature of the conversations. The bot scrapes information from the relevant APIs (e.g., IMDB), leading to unique Alexa productions.

[5]A post-hoc model confirmed there were no interactions between Gender and the other covariates.

[6]More complex random effects structures resulted in singularity errors, indicating model overfit.

TABLE 2 | Mean speech rate for users (baseline and interaction) and Alexa.

	In-lab	Diff. from user baseline	At-home	Diff. from user baseline	Two sample t-test (in-lab vs. at-home)
User (baseline)	4.17 syll/s	–	3.13 syll/s	–	t (42.71) = 2.38, p < 0.05
User (interaction)	2.93 syll/s	(Slower: −1.24)	3.07 syll/s	(Slower: −0.06)	t (1,468.8) = 2.62, p < 0.01
Alexa	3.95 syll/s	(Slower: −0.22)	3.77 syll/s	(Faster: +0.64)	t (1,678.3) = 5.11, p < 0.001

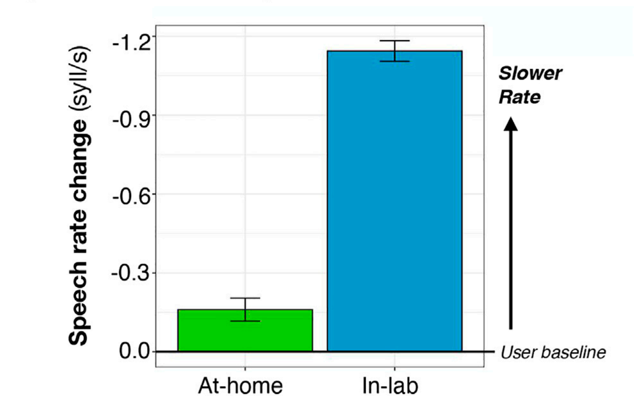


FIGURE 1 | Speech rate change (relative to the user's baseline, 0.0; centered) across the at-home and in-lab settings. A decrease in the number of syllables/second (a more negative number) indicates a slower speech rate. Error bars depict standard error.

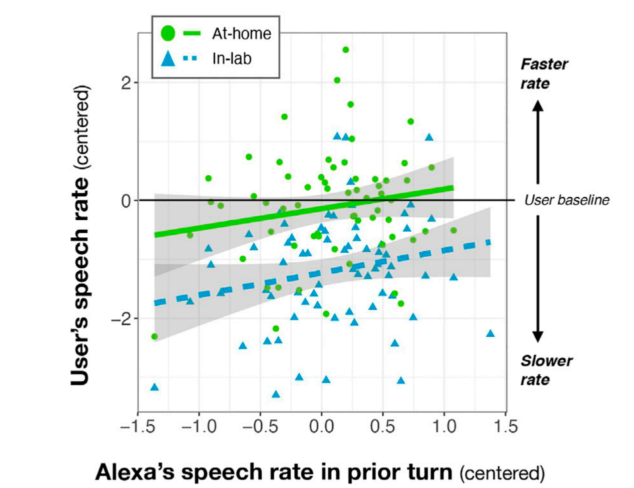


FIGURE 2 | Speech rate change (centered relative to the user's baseline) across the at-home and in-lab settings. A decrease in the number of syllables/second (a negative number) indicates a slower speech rate. Shading depicts the standard error.

0.01]. Furthermore, there was an effect of Gender: female users slowed their speech rate even more during socialbot interactions [*Coef* = −3.18, t = −2.95, p < 0.01].

There was also an effect of Alexa Prior Turn Rate: user's speech rate increases when the speech rate increases in the Alexa Prior Turn [*Coef* = 5.61, t = 11.90, p < 0.001] (see **Figure 2**). There was no observed effect of Overall Rating and no interactions observed.

Hyperarticulation and/or Entrainment After an Automatic Speech Recognition Error?

We analyzed speaker's speech rate in a subset of the data consisting of the four user turns preceding an ASR error (Pre-Error) and four turns following an ASR error (Post-Error) (n = 771 turns, n = 32[7] users). Rate (centered) was modeled with a linear mixed effects model. Fixed effects included Error Condition (pre-error, post-error), Setting (in-lab, at-home), their interaction, and Gender (female, male)[8], and by-User random intercepts. Contrasts were sum coded.

The model revealed a similar effect in the Pre- and Post-error subset as in the main model: an overall negative intercept [*Coef* = −0.71, t = −6.88, p < 0.001], an effect of Setting [*Coef* = 0.44 t = −4.45, p < 0.001], Alexa Prior Turn [*Coef* = 0.57, t = 7.91, p < 0.001], as well as Gender [*Coef* = −0.36, t = −3.55, p < 0.01]. However, there were no effects of Error Condition and no interactions including Error Condition observed.

DISCUSSION

This study examined users' speech rate hyperarticulation and entrainment toward an Amazon Alexa socialbot in a conversational interaction. While generally tested and analyzed separately (e.g., Burnham et al., 2010; Cohn et al., 2019), this study highlights the importance of accounting for both hyperarticulation and entrainment to provide a fuller picture of speech interactions with voice-AI/computer interlocutors.

First, we find evidence of hyperarticulation: relative to their original baseline productions, users consistently decrease their speech rate when talking to the socialbot. This supports listener-centered accounts: speakers produce 'clearer' speech for listeners who might have trouble understanding them (Lindblom, 1990; Smiljanić and Bradlow, 2009). Indeed, these findings are consistent with slower speech rate observed for interlocutors

[7]We only included participants who did not have additional errors within the ± 4 turns. For example, if multiple errors occurred within four turns, we did not include those participants (n = 3).

[8]Posthoc models confirmed no significant interactions between Gender and the covariates.

presumed to have communicative difficulties, such as dialogue systems that have higher error rates (Oviatt et al., 1998b; Stent et al., 2008), as well as infants and non-native speakers (Fernald and Simon, 1984; Scarborough et al., 2007).

Above and beyond the hyperarticulation effect, we *also* find evidence for turn-level entrainment toward the speech rate patterns of the social bot. If Alexa produces a faster speech rate, users are more likely to speed up in the subsequent turn; conversely, if Alexa's speech rate slows, users also slow their rate in the subsequent turn. This is consistent with prior findings in entrainment toward computers (e.g., amplitude convergence toward computer characters in Coulston et al., 2002). Yet, we did not find evidence that entrainment was linked to social ratings of the interaction, as is proposed by some alignment accounts (Giles et al., 1991; Shepard et al., 2001). One possibility is that socially mediated pressures differently affect entrainment toward voice-AI and humans in non-task oriented interactions (here, social chit-chat), but might do so in more task-oriented interactions (e.g., in a tutoring task in Thomason et al., 2013) or in less socially rich contexts (e.g., single word shadowing in (Cohn et al., 2019; Zellou et al., 2021). Another possibility is that the range of ratings might have been too narrow to detect a difference (if present), where the majority of speakers rated the interactions favorably. Future work exploring whether social sentiments influence entrainment toward socialbots can elucidate these questions.

Furthermore, we also observed differences in speech rate hyperarticulation by context: users slowed down even more in conversations in-lab than at-home. This is consistent with our prediction that participants would produce more careful, 'clear' speech when other observers were present—and is in line with *Audience Design* theory (Clark and Carlson, 1982) that productions are also tailored based on 'overhearers'. Still, we cannot conclusively point to the overhearer as the source of this effect; it is possible that this reflects that the in-lab condition participants produced *faster* speech in their baseline (averaging ~4 syllables/sec) and, possibly, had more room to hyperarticulate (slowing to an average of 2.93 syllables/sec). Future work parametrically manipulating speech rate—as well as comparing the same participants both in-lab and at-home can further tease apart these possibilities.

In addition to examining situational context, we also tested the impact of functional pressures in communication—specifically whether speakers hyperarticulate and/or entrain more following a system ASR error. We did not find effects for either behavior, contra findings human-computer interaction for post-error hyperarticulation (e.g., Oviatt et al., 1998b; Vertanen, 2006) or post-error entrainment (Zellou and Cohn, 2020). One possible explanation for why we do not observe hyperarticulation following ASR errors is that speakers were already talking in a very slow, 'clear speech' manner when talking to the socialbot. This explanation is consistent with studies in which, at a higher error rate, speakers maintain hyperarticulation (Oviatt et al., 1998b; Stent et al., 2008).

There were also limitations in the present study that can serve as the basis for future research. One such limitation is that we had different participants in the in-lab and at-home conditions; while one benefit to this approach was that the interaction consisted of the first socialbot conversation each user had with the system, future work examining user speech across different contexts can further tease apart the source of differences observed across settings. Furthermore, we observed differences by gender, where female participants slowed their speech even more to the socialbot; yet, as the current study was not balanced by gender, future work is needed to test whether this difference is truly socially mediated—with more hyperarticulation produced by females (e.g., increased pitch range by females in Oviatt et al., 1998b)—or possibly driven by the individual speakers in the study. Additionally, here we test one socialbot system; future work testing other systems can shed more light on how users hyperarticulate and entrain toward socialbots, more generally.

Overall, this study contributes to our broader scientific understanding of human and voice-AI interaction. Here, we find that speakers use hyperarticulation *and* entrainment in speech interactions with an Alexa socialbot, paralleling some patterns observed in human-human interaction. Future work directly testing a human vs. socialbot interlocutor comparison can further tease apart possible differences in social interactions with the two types of interlocutors. Additionally, human-human conversational entrainment is coordinative, with each speaker adapting their output (Levitan et al., 2015; Szabó, 2019). There is some work investigating the effects of adapting TTS output to entrain toward the user (Lubold et al., 2016). Future studies examining the extent to which speakers entrain to Alexa socialbots—as they entrain to the user—can shed light on the situational, functional, and interpersonal dynamics of human-socialbot interaction.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because participants cannot be deidentified in their conversations with the socialbot. Requests to access the datasets should be directed to mdcohn@ucdavis.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UC Davis Institutional Review Board (IRB). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MC and K-HL contributed to the conception and design of the study. K-HL developed the socialbot with ZY. MC and MS led the acoustic analysis and received feedback from GZ. MC wrote the first draft of the manuscript. All authors contributed to the editing and revision of the manuscript.

FUNDING

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC, an Amazon Faculty Research Award to GZ, and the Amazon Alexa Prize competition in supporting the bot development to ZY and KL. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

REFERENCES

Abdolrahmani, A., Kuber, R., and Branham, S. M. (2018). " Siri Talks at You" an Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, 249–258.

Ammari, T., Kaye, J., Tsai, J. Y., and Bentley, F. (2019). Music, Search, and IoT. *ACM Trans. Comput.-Hum. Interact.* 26 (3), 1–28. doi:10.1145/3311956

Babel, M., and Bulatov, D. (2012). The Role of Fundamental Frequency in Phonetic Accommodation. *Lang. Speech* 55 (2), 231–248. doi:10.1177/0023830911417695

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Softw.* 67 (1), 1–48. doi:10.18637/jss.v067.i01

Bell, L., Gustafson, J., and Heldner, M. (2003). Prosodic Adaptation in Human-Computer Interaction. *Proc. ICPHS* 3, 833–836.

Bell, L., and Gustafson, J. (1999). Repetition and its Phonetic Realizations: Investigating a Swedish Database of Spontaneous Computer-Directed Speech. *Proc. ICPhS* 99, 1221–1224.

Belletier, C., and Camos, V. (2018). Does the Experimenter Presence Affect Working Memory?. *Ann. N.Y. Acad. Sci.* 1424, 212–220. doi:10.1111/nyas.13627

Belletier, C., Davranche, K., Tellier, I. S., Dumas, F., Vidal, F., Hasbroucq, T., et al. (2015). Choking under Monitoring Pressure: Being Watched by the Experimenter Reduces Executive Attention. *Psychon. Bull. Rev.* 22 (5), 1410–1416. doi:10.3758/s13423-015-0804-9

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., and Lottridge, D. (2018). Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2 (3), 1–24. doi:10.1145/3264901

Boersma, P., and Weenink, D. (2018). *Praat: Doing Phonetics by Computer (6.0.37).* [Computer software]. http://www.praat.org/.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Brown, A. (2011). The Role of Beliefs in Lexical Alignment: Evidence from Dialogs with Humans and Computers. *Cognition* 121 (1), 41–57. doi:10.1016/j.cognition.2011.05.011

Brumm, H., and Zollinger, S. A. (2011). The Evolution of the Lombard Effect: 100 Years of Psychoacoustic Research. *Behav.* 148 (11–13), 1173–1198. doi:10.1163/000579511x605759

Burnham, D. K., Joeffry, S., and Rice, L. (2010). Computer-and Human-Directed Speech before and after Correction. Proceedings of the 13th Australasian International Conference on Speech Science and Technology. Melbourne, Australia, 13–17.

Chartrand, T. L., and Bargh, J. A. (1996). Automatic Activation of Impression Formation and Memorization Goals: Nonconscious Goal Priming Reproduces Effects of Explicit Task Instructions. *J. Personal. Soc. Psychol.* 71 (3), 464–478. doi:10.1037/0022-3514.71.3.464

Chen, C.-Y., Yu, D., Wen, W., Yang, Y. M., Zhang, J., Zhou, M., et al. (2018). *Gunrock: Building A Human-like Social Bot by Leveraging Large Scale Real User Data*. 2nd Proceedings of Alexa Prize. Available at: https://developer.amazon.com/alexaprize/challenges/past-challenges/2018

Clark, H. H., and Carlson, T. B. (1982). Hearers and Speech Acts. *Language* 58, 332–373. doi:10.1353/lan.1982.0042

Cohn, M., Ferenc Segedin, B., and Zellou, G. (2019). Imitating Siri: Socially-Mediated Alignment to Device and Human Voices. Proceedings of International Congress of Phonetic Sciences, 1813–1817.

Comrie, B. (1976). Linguistic Politeness Axes: Speaker-Addressee, Speaker-Referent, Speaker-Bystander. *Pragmatics Microfiche* 1 (7), 1–12.

Coulston, R., Oviatt, S., and Darves, C. (2002). Amplitude Convergence in Children's Conversational Speech with Animated Personas. Seventh International Conference on Spoken Language Processing.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., and Beale, R. (2015). Voice Anthropomorphism, Interlocutor Modelling and Alignment Effects on Syntactic Choices in Human−computer Dialogue. *Int. J. Human-Computer Stud.* 83, 27–42. doi:10.1016/j.ijhcs.2015.05.008

De Jong, N. H., Wempe, T., Quené, H., and Persoon, I. (2017). *Praat Script Speech Rate V2*. https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2.

Fandrianto, A., and Eskenazi, M. (2012). Prosodic Entrainment in an Information-Driven Dialog System. Thirteenth Annual Conference of the International Speech Communication Association.

Fernald, A., and Simon, T. (1984). Expanded Intonation Contours in Mothers' Speech to Newborns. *Dev. Psychol.* 20 (1), 104–113. doi:10.1037/0012-1649.20.1.104

Gabriel, R., Liu, Y., Gottardi, A., Eric, M., Khatri, A., Chadha, A., et al. (2020). Further Advances in Open Domain Dialog Systems in the Third Alexa Prize Socialbot Grand Challenge. *Proc. Alexa Prize*. Available at: https://assets.amazon.science/0e/e6/2cff166647bfb951b3ccc67c1d06/further-advances-in-open-domain-dialog-systems-in-the-third-alexa-prize-socialbot-grand-challenge.pdf

Gessinger, I., Raveh, E., Le Maguer, S., Möbius, B., and Steiner, I. (2017). Shadowing Synthesized Speech-Segmental Analysis of Phonetic Convergence. *Interspeech*. 3797–3801. doi:10.21437/Interspeech.2017-1433

Gessinger, I., Raveh, E., Steiner, I., and Möbius, B. (2021). Phonetic Accommodation to Natural and Synthetic Voices: Behavior of Groups and Individuals in Speech Shadowing. *Speech Commun.* 127, 43–63. doi:10.1016/j.specom.2020.12.004

Giles, H., Coupland, N., and Coupland, I. (1991). *1. Accommodation Theory: Communication, Context, and. Contexts of Accommodation: Developments in Applied Sociolinguistics*, 1.

Habler, F., Schwind, V., and Henze, N. (2019). Effects of Smart Virtual Assistants' Gender and Language. In Proceedings of Mensch und Computer 2019. 469–473.

Lee, D.-Y., and Baese-Berk, M. M. (2020). The Maintenance of Clear Speech in Naturalistic Conversations. *The J. Acoust. Soc. America* 147 (5), 3702–3711. doi:10.1121/10.0001315

Levitan, R., Benus, S., Gravano, A., and Hirschberg, J. (2015). *Entrainment and Turn-Taking in Human-Human Dialogue*. AAAI Spring Symposia. doi:10.18653/v1/w15-4644

Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., and Nenkova, A. (2012). Acoustic-prosodic Entrainment and Social Behavior. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 11–19.

Levitan, R., and Hirschberg, J. (2011). Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Levels and Dimensions. Twelfth Annual Conference of the International Speech Communication Association.

Liang, K., Chau, A., Li, Y., Lu, X., Yu, D., Zhou, M., et al. (2020). *Gunrock 2.0: A User Adaptive Social Conversational System*. ArXiv Preprint ArXiv:2011.08906.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. Speech production and speech modelling. Springer, 403–439. doi:10.1007/978-94-009-2037-8_16

Lovato, S., and Piper, A. M. (2015). Siri, Is This You?: Understanding Young Children's Interactions with Voice Input Systems. Proceedings of the 14th International Conference on Interaction Design and Children, 335–338.

Lubold, N., and Pon-Barry, H. (2014). Acoustic-prosodic Entrainment and Rapport in Collaborative Learning Dialogues. Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge, 5–12.

Lubold, N., Walker, E., and Pon-Barry, H. (2016). Effects of Voice-Adaptation and Social Dialogue on Perceptions of a Robotic Learning Companion. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 255–262.

Lunsford, R., Oviatt, S., and Arthur, A. M. (2006). Toward Open-Microphone Engagement for Multiparty Interactions. Proceedings of the 8th International Conference on Multimodal Interfaces, 273–280.

Maniwa, K., Jongman, A., and Wade, T. (2009). Acoustic Characteristics of Clearly Spoken English Fricatives. *J. Acoust. Soc. America* 125 (6), 3962–3973. doi:10.1121/1.2990715

Mayo, C., Aubanel, V., and Cooke, M. (2012). Effect of Prosodic Changes on Speech Intelligibility. Thirteenth Annual Conference of the International Speech Communication Association, 1706–1709.

Nass, C., Moon, Y., Morkes, J., Kim, E.-Y., and Fogg, B. J. (1997). Computers Are Social Actors: A Review of Current Research. *Hum. Values Des. Comp. Tech.* 72, 137–162. doi:10.1145/259963.260288

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers Are Social Actors. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 72–78.

Orne, M. T. (1962). On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications. *Am. Psychol.* 17 (11), 776–783. doi:10.1037/h0043424

Oviatt, S., Levow, G.-A., Moreton, E., and MacEachern, M. (1998a). Modeling Global and Focal Hyperarticulation during Human-Computer Error Resolution. *J. Acoust. Soc. America* 104 (5), 3080–3098. doi:10.1121/1.423888

Oviatt, S., MacEachern, M., and Levow, G.-A. (1998b). Predicting Hyperarticulate Speech during Human-Computer Error Resolution. *Speech Commun.* 24 (2), 87–110. doi:10.1016/s0167-6393(98)00005-3

Oviatt, S., and VanGent, R. (1996). Error Resolution during Multimodal Human-Computer Interaction. Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference On, 204–207.

Pardo, J. S., Urmanche, A., Wilman, S., and Wiener, J. (2017). Phonetic Convergence across Multiple Measures and Model Talkers. *Atten Percept Psychophys* 79 (2), 637–659. doi:10.3758/s13414-016-1226-0

Pickering, M. J., and Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Res. Lang. Comput.* 4 (2–3), 203–228. doi:10.1007/s11168-006-9004-0

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. (2017). "Alexa Is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2853–2859. doi:10.1145/3027063.3053246

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., et al. (2018). *Conversational AI: The Science behind the Alexa Prize*. ArXiv Preprint ArXiv: 1801.03604.

Raveh, E., Siegert, I., Steiner, I., Gessinger, I., and Möbius, B. (2019). Three's a Crowd? Effects of a Second Human on Vocal Accommodation with a Voice Assistant. *Proc. Interspeech* 2019, 4005–4009. doi:10.21437/Interspeech.2019-1825

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., and Brenier, J. (2007). An Acoustic Study of Real and Imagined Foreigner-directed Speech. *J. Acoust. Soc. America* 121 (5), 3044. doi:10.1121/1.4781735

Shepard, C. A., Robinson, W. P., and Giles, H. (2001). "Communication Accommodation Theory," in *The New Handbook of Language and Social Psychology* (John Wiley & Sons), 33–56.

Siegert, I., and Krüger, J. (2020). ""Speech Melody and Speech Content Didn't Fit Together"-Differences in Speech Behavior for Device Directed and Human Directed Interactions," in *Advances in Data Science: Methodologies and Applications* (Springer), 65–95. doi:10.1007/978-3-030-51870-7_4

Siegert, I., Nietzold, J., Heinemann, R., and Wendemuth, A. (2019). "The Restaurant Booking Corpus–Content-Identical Comparative Human-Human and Human-Computer Simulated Telephone Conversations," in *Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, 126–133.

Smiljanić, R., and Bradlow, A. R. (2009). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Lang. Linguist Compass* 3 (1), 236–264. doi:10.1111/j.1749-818X.2008.00112.x

Stent, A. J., Huffman, M. K., and Brennan, S. E. (2008). Adapting Speaking after Evidence of Misrecognition: Local and Global Hyperarticulation. *Speech Commun.* 50 (3), 163–178. doi:10.1016/j.specom.2007.07.005

Swerts, M., Litman, D., and Hirschberg, J. (2000). Corrections in Spoken Dialogue Systems. Sixth International Conference on Spoken Language Processing.

Szabó, I. E. (2019). Phonetic Selectivity in Accommodation: The Effect of Chronological Age. Proceedings of the 19th International Congress of Phonetic Sciences. Canberra, Australia, 3195–3199.

Thomason, J., Nguyen, H. V., and Litman, D. (2013). Prosodic Entrainment and Tutoring Dialogue Success. International Conference on Artificial Intelligence in Education, 750–753. doi:10.1007/978-3-642-39112-5_104

Vertanen, K. (2006). Speech and Speech Recognition during Dictation Corrections. Ninth International Conference on Spoken Language Processing, 1890–1893.

Zellou, G., Cohn, M., and Ferenc Segedin, B. (2021). Age- and Gender-Related Differences in Speech Alignment toward Humans and Voice-AI. *Front. Commun.* 5, 1–11. doi:10.3389/fcomm.2020.600361

Zellou, G., and Cohn, M. (2020). Social and Functional Pressures in Vocal Alignment: Differences for Human and Voice-AI Interlocutors. *Proc. Interspeech* 2020, 1634–1638. doi:10.21437/Interspeech.2020-1335

Conflict of Interest: The authors declare that this study received funding from Amazon.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 28 May 2021
doi: 10.3389/fcomp.2021.682982

'Alexa, I feel for you!' Observers' Empathetic Reactions towards a Conversational Agent

Astrid Carolus[1]*, *Carolin Wienrich*[2], *Anna Törke*[1], *Tobias Friedel*[1], *Christian Schwietering*[1] *and Mareike Sperzel*[1]

[1]*Institute Human-Computer-Media, Media Psychology, Julius-Maximilians-University, Würzburg, Germany,* [2]*Institute Human-Computer-Media, Human-Technique-Systems, Julius-Maximilians-University, Würzburg, Germany*

Conversational agents and smart speakers have grown in popularity offering a variety of options for use, which are available through intuitive speech operation. In contrast to the standard dyad of a single user and a device, voice-controlled operations can be observed by further attendees resulting in new, more social usage scenarios. Referring to the concept of 'media equation' and to research on the idea of 'computers as social actors,' which describes the potential of technology to trigger emotional reactions in users, this paper asks for the capacity of smart speakers to elicit empathy in observers of interactions. In a 2 × 2 online experiment, 140 participants watched a video of a man talking to an Amazon Echo either rudely or neutrally (factor 1), addressing it as 'Alexa' or 'Computer' (factor 2). Controlling for participants' trait empathy, the rude treatment results in participants' significantly higher ratings of empathy with the device, compared to the neutral treatment. The form of address had no significant effect. Results were independent of the participants' gender and usage experience indicating a rather universal effect, which confirms the basic idea of the media equation. Implications for users, developers and researchers were discussed in the light of (future) omnipresent voice-based technology interaction scenarios.

Keywords: conversational agent, empathy, smart speaker, media equation, computers as social actors, human-computer interaction

OPEN ACCESS

Edited by:
Benjamin Weiss,
Hanyang University, South Korea

Reviewed by:
Iulia Lefter,
Delft University of Technology, Netherlands
Zoraida Callejas,
University of Granada, Spain

Correspondence:
Astrid Carolus
astrid.carolus@uni-wuerzburg.de

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

Received: *19 March 2021*
Accepted: *05 May 2021*
Published: *28 May 2021*

Citation:
Carolus A, Wienrich C, Törke A, Friedel T, Schwietering C and Sperzel M (2021) 'Alexa, I feel for you!' Observers' Empathetic Reactions towards a Conversational Agent. Front. Comput. Sci. 3:682982. doi: 10.3389/fcomp.2021.682982

INTRODUCTION

Conversational Agents (CAs) have grown in popularity over the last few years (Keynes, 2020). New devices such as smart speakers (Perez, 2019) or application such as chatbots or virtual assistants (Petrock, 2019) have become part of everyday technology usage. The voice-controlled operation of technology offers a variety of features and functions such as managing a calendar or controlling the lights. They promise to simplify daily life, while their operation is convenient and low threshold (Cannon, 2017). CAs are utilizable in situations in which users need their hands for something else than handling a device. Inexperienced or less skilled users are capable to operate CAs (Sansonnet et al., 2006) resulting in increasing numbers of user groups. Moreover, CAs have become the object of user-centered scientific research, which analyzes the human users' reactions towards the device, the underlying usage motivations or the effects of usage [e.g., CHI 2019 Workshop by Jacques et al. (2019)]. Research in this area so far mostly focuses on the standard scenario of technology usage: a single person operates a certain device. However, the usage of conversational agents expands this user-device dyad. A more social scenario will unfold if the user speaking to an CA is observed by others or if the device is used jointly by couples or families (for an overview: Porcheron et al., 2017). Consequently, speech operation widens the scientific perspective on technology usage. Not only the users themselves but also co-users and observers come into view. For example, research analyzing the verbal abuse of CAs impressively demonstrates need for research widening its focus: 10 to approximately 40% of interactions include aspects of abusive language or misuse of the device (Chin et al., 2020). As these interactions might be observed by others, by partners or children, the impact of the abusive behavior clearly exceeds the user-device-dyad.

Referring to increasingly social usage scenarios, the present study focuses on the effects of observing user-technology interactions and asks: How are observers of a rude interaction between a CA and a user affected? Do observers experience empathetic reactions towards a smart speaker, which is treated rudely?

THEORETICAL BACKGROUND

Until recently, the ability to understand and apply spoken language was regarded as a fundamentally and exclusively human characteristic (Pinker, 1994). Interacting with 'talking' devices and 'having a conversation' with CAs constitutes a new level of usage relevant for both components of HCI, the perspective on 'humans' as well as on 'computers' (Luger and Sellen, 2016).

Talking Technology: Conversational Agents

Conversational Agents–also referred to as voice assistants, vocal social agents, voice user interfaces or smart speakers–are among the most popular devices used to run voice-controlled personal assistants (Porcheron et al., 2018). The best-selling smart speakers worldwide are Amazon Echo and Google Home running Alexa or Google Assistant, respectively (BusinessWire, 2018). In the US, 66.4 million people own a smart speaker already, with a 61% market share of Amazon Echo and 24% share of Google Home (Perez, 2019). In 2019, 20% of United Kingdom households had a smart speaker already, while Germany passed the 10% mark (Kinsella, 2019). Statistics predict that the global increase will continue to 1.8 billion users worldwide by the end of 2021 (Go-Gulf, 2018).

Within the last 60 years, natural language-processing applications and the number of services supporting voice commands have evolved rapidly. McTear, Callejas and Griol (2016) summarize the history of conversational interfaces beginning back in the 1960s, when the first text-based dialogue systems answered questions and the first chatbots simulated casual conversation. About 20 years later, speech-based dialogue systems and spoken dialogue technology evolved, which were soon transferred to commercial contexts. In subsequent years, conversational agents and social robots came into the picture. Hirschberg and Manning (2015), p. 261 ascribe the recent improvements to four key factors: the progress in computing power, machine learning, and in the understanding of the structure and the deployment of human language. Additionally, it can be ascribed to the large amounts of linguistic data available. As a result, today's CA technology is much easier to use than earlier voice recognition systems, which allowed only very restricted phrases and word patterns. From a user-centered perspective, human-computer interaction has never resembled human-human interaction that closely. CAs are on their way to become everyday interaction partners. We will learn how to interact with them, how to pronounce and how to phrase our speech commands properly–also by observing others. Children will grow up observing their parents interacting with smart speakers before interacting with them themselves (Hoy, 2018).

Questions about the effects of these new characteristics of usage arise from multiple disciplines with developers, researchers, users and the society as a whole being involved. First studies reveal ethical challenges (Pyae and Joelsson, 2018). By demonstrating the 'submissiveness (of AI software) in the face of gender abuse,' West et al. (2019), p. 4 raise issues that CAs, which are projected as young women ('Alexa,' 'Siri,' 'Cortana'), potentially perpetuate gender biases. To gain deeper insights into the impact of 'speaking' technology on humans, the media equation approach offers a fruitful theoretical framework.

The 'Computers As Social Actors' Paradigm and the Media Equation Approach

In the early 1990s, Clifford Nass and Byron Reeves introduced a new way of understanding electronic devices. They conceptualized computers as 'social actors' to which users automatically react as if they were human beings (Reeves and Nass, 1996). Their empirical studies revealed that users tend to (unconsciously) interpret cues sent by computers as social indicators of a human counterpart to 'whom' they react accordingly. Literature provides various explanations for this phenomenon with the evolutionary perspective offering a framing theoretical perspective (Nass et al., 1997; Nass and Gong, 2000; Kraemer et al., 2015; Carolus et al., 2019a). Like our bodies, the human brain is adapted to our early ancestors' world in which every entity one perceived was a real physical object and every entity communicating as a human being sure enough was a human being (Buss and Kenrick, 1998). 'Mentally equipped' in this way, we encounter today's new media and technology sending various cues, which would have indicated a human counterpart back in the days of our evolutionary ancestors. Unconsciously, 'evolved psychological mechanisms'–neurocognitive mechanisms evolved to efficiently contribute to the adaptive problems of our ancestors' world (here: interacting with other human beings)–are triggered (Cosmides and Tooby, 1994). Thus, the computer's cues are interpreted as a communicative act and therefore as a human interaction resulting in the individual to behave accordingly, showing social reaction originally exclusive for human-human interaction (Tooby and Cosmides, 2005; Dawkins, 2016).

Research analyzing these phenomena followed an experimental approach, which Nass et al. (1994) referred to as the paradigm 'Computers As Social Actors' (CASA). Findings of social science describing social dynamics of human-human interactions were transferred to the context of human-computer interactions. In laboratory studies, one of the human counterparts of the social dyad was replaced by a computer sending certain allegedly 'human' cues. For example, they anticipated research on CAs when they analyzed participants interacting with computers, which appeared to be voice controlled talking in either a female or a male voice. Results revealed that participants transferred gender stereotypes to these devices. A domineering appearance of a computer 'speaking' with a male voice was interpreted more positively as the exact identical statements presented by a computer with a female voice (Nass et al., 1997). Considering that these experiments took place more than 20 years ago, when devices were much bulkier and their handling was a lot more difficult, the revealed social impact of technology on their users was remarkable. Particularly, when considering that participants reported to know they were interacting with a computer–and not with a male or female person. They consciously knew that they were interacting with technology, but they (unconsciously) ascribed gender stereotypical characteristics.

Additional studies revealed further indication of gender stereotyping (Lee et al., 2000; Lee and Nass, 2002; Morishima et al., 2002) and further social norms and rules to be applied to computers, e.g., politeness (Nass et al., 1999) or group membership (Nass et al., 1996). More recently, studies transferred this paradigm to more recent technology (Carolus et al., 2018; Carolus et al., 2019a). Rosenthal-von der Puetten et al. (2013) widened the focus and analyzed observers' empathetic reactions toward a dinosaur robot (Ugobe's Pleo). They showed that witnessing the torture of this robot elicits empathetic reactions in observers. Similarly, Cramer et al. (2010) as well as Tapus et al. (2007) revealed different forms of empathy on users' attitudes toward robots in Human-Robot Interactions [see also: Rosenthal-von der Puetten et al. (2013)]. However, the experimental manipulation to observe animal like or humanlike robots being tortured invites contradiction. Empathetic reactions might be triggered by effects of anthropomorphism (Riek et al., 2009) as the objects and their treatment highlighted the anthropomorphic (or animal like) character of the devices. Social robots looked like living creatures and reacted to physical stimulus, accordingly. Moreover, the incidents participants observed are of limited everyday relevance. They were invited to a laboratory to interact (or watch others interacting) with devices, which are far away from everyday experiences. Consequently, the explanatory power regarding everyday technology usage outside of scientists' laboratories is limited.

Smart Speaker as Language-Processing Social Agents

Modern CAs can 'listen' and respond to the users' requests in 'natural and meaningful ways' (Lee et al., 2000, p. 82) which Luger and Sellen refer to as 'the next natural form of HCI' (Luger and Sellen, 2016, p. 5286). From the media equation perspective, voice assistants represent a new form of a 'social actor': CAs adopt one of the most fundamentally human characteristics resulting in research questions regarding their social impact on human users–those actively speaking to them and those listening (Purington et al., 2017).

While processing language is a salient, humanlike feature of CAs, the outward appearance of smart speakers, for example, is distinctly technological. In contrast to embodied agents or social robots, they are barely anthropomorphic but look like portable loudspeakers. Google's Home Mini and Amazon's Echo Dot resemble an oversized version of a puck, their larger devices (Echo and Home) come in a cylinder shape with some colored light signals. Consequently, smart speakers are regarded as a promising research object to analyze the effects of speech–independent of further humanoid or anthropomorphistic cues such as the bodily or facial expressions of a social robot, for example. Moreover, they bridge the gap of everyday relevance as they are increasingly popular and are used by an increasing number of average users outside of scientific contexts. Finally, because 'speaking is naturally observable and reportable to all present to its production' (Porcheron et al., 2017, p. 433), smart speakers allow realistic social usage scenarios. Devices are designed for multiple simultaneous users. Amazon Echo and Google Home, for example, use multiple microphones and speakers facilitating conversations with multiple users. Consequently, bystanders of interactions–children, partners or other family members–watch, listen, or join the conversation and might be affected by its outcomes (Sundar, 2020).

Empathy-Put Yourself in Their Shoes

A huge body of literature focuses on empathy resulting in various attempts to define the core aspects of the concept with Cuff et al. (2016) presenting 43 different definitions. Back in 1872, the German philosopher Robert Vischer introduced the term 'Einfuehlung' which literally means 'feeling into' another person (Vischer, 1873). Taking the perspective of this other person aims for an understanding of 'what it would be like to be living another body or another environment' (Ganczarek et al., 2018). A few years later, Lipps (1903) argued that an observer of another person's emotional state tends to imitate the emotional signals of the other person 'inwardly' by physically adapting body signals. Macdougall (1910) and Titchener (1909) translated 'Einfuehlung' as 'Empathy' and introduced the term still used today. Today, empathy is broadly referred to as 'an affective response more appropriate to someone else's situation than to one's own' (Hoffman, 2001, p. 4). Moreover, modern conceptualizations distinguish two main perspectives on 'empathy': the rational understanding and the affective reaction to another person's feelings or circumstances (De Vignemont and Singer, 2006). The cognitive component refers to the recognition and understanding of the person's situation by including the subcomponents of perspective-taking (i.e., adopting another's psychological perspective) and identification (i.e., identifying with the other character). The affective component includes the subcomponent of empathic concern (e.g., sympathy, compassion), pity (i.e., feeling sorry for someone) and personal distress (i.e., feelings of discomfort or anxiety) into the process of sympathizing (Davis, 1983; for an overview of empathy from a neuro scientific, psychological, and philosophical perspective see Rogers et al. (2007). Conceptually and methodologically, Davis (1983, p. 168) widens the perspective on empathy when distinguishing between state empathy and trait empathy. State empathy describes an affective state and is a result of a 'situational manipulation' (Duan and Hill, 1996). Trait empathy is defined as a dispositional trait resulting in enduring interindividual differences (Hoffman, 1982), which is a significant predictor of empathetic emotion (Davis, 1983).

Empathetic Reactions to Technology

Users' emotional reactions towards technology has become an increasingly important area of research Misuse or abusive treatment have been reported for various forms of technological artifacts or social agents, with a substantial body of research focusing on graphically represented (virtual) agents and robots (Brahnam and De Angeli, 2008; Paiva et al., 2017). Due to their relative novelty, little empirical research has been done regarding abusive interactions with smart speakers. Although not entirely comparable, robots and smart speakers are regarded as intelligent agents interacting with its users through a physical body. Thus, the literature review focuses on literature on robots as 'empathetic agents' (Paiva et al., 2017).

Both anecdotal and scientific examples reveal incidents of aggressive and abusive behavior towards robots (Bartneck and Hu, 2008). Salvini et al. (2010) reported on a cleaning robot abused by bypassers. Brscic et al. (2015) told about children abusing a robot in a mall (for an overview see: Tan et al. (2018)). A more recent study refers to another perspective: people react empathetically to robots which are attacked by others (for an overview: Leite et al. (2014)). For example, Vincent (2017) reported on a drunken man, who attacked a robot in a car park resulting in empathetic reactions with the robot. Empathy with social robots has been studied increasingly in the last years to understand humans' empathetic reactions towards them, to prevent abusive behavior or to develop robots, which users perceive as being empathetic agents (e.g., Salvini et al., 2010; Nomura et al., 2016; Bartneck and Keijsers, 2020). As the abuse even of technology raises ethical questions (and economic questions due to resulting destruction), most of the studies avoid encouraging the participants to physically harm the device. Instead, paradigms involving reduced radical variations of abuse were established (Paiva et al., 2017; Bartneck and Keijsers, 2020). Rosenthal-von der Puetten et al. (2013) introduced a paradigm, which allows to study more radical interactions. Their participants were shown pre-recorded videos of a dinosaur robot, which was tortured physically. Observing torture resulted in increased physiological arousal and self-reports revealed more negative and less positive feelings. In sum, measures revealed participants' empathy with the robot–a finding that literature review confirms (Riek et al., 2009; Kwak et al., 2013; Paiva et al., 2017). In a follow-up study, Rosenthal-von der Puetten et al. (2014) compared participants' neural activation and self-reports when watching a video of a human or a robot being harmed. In the human-torture condition, neural activity and self-reports revealed higher levels of emotional distress and empathy. Further studies asked for the characteristics of the entities that elicit emotional reactions revealing that participants were rather empathetic with robots, which were more humanlike in terms of anthropomorphistic appearances perceived agency and the capacity to express empathy themselves (Hegel et al., 2006; Riek et al., 2009; Cramer et al., 2010; Gonsior et al., 2012; Leite Iolanda et al., 2013; Leite et al., 2013b). Additionally, Kwak et al. (2013) emphasized the impact of physical embodiment. Participants' empathy was more distinct toward a physically embodied robot (vs. a physically disembodied robot).

As an interim conclusion, literature research on empathy with technological devices focused on robots and revealed that humanlike cues (e.g., outward appearance and empathy toward their human counterparts) increased the level of elicited empathy in humans. However, robots send various social cues (e.g., outward appearance, move, facial expression, verbal and nonverbal communication), which evolve into meaningful social signals during an interaction (for an overview: Feine et al. (2019). Analyzing interactions with these technological entities will result in confounding regarding the underlying cues of elicited empathy. Consequently, the present study takes a step back to refine the analysis. Following the taxonomy of social cues (Feine et al., 2019, p. 30), we distinguish between verbal, visual, auditory and invisible cues that CAs could present. With or focus on an interaction with a smart speaker we concentrate on verbal cues keeping visual cues reduced (simple cylindric shape of the device, no facial or bodily expression). Therefore, smart speakers offer the externally valid option to narrow down the magnitude of social cues and study the effects of (mainly) verbal cues only. Moreover, pre-recorded videos were found to constitute a promising approach to study the perspective of bystanders or witnesses of interactions with technology.

Interindividual Differences in Empathetic Reactions

To elaborate empathetic reactions to technology, potential interindividual differences need to be considered. De Vignemont and Singer (2006) analyzed modulatory factors, which affect empathy. Two factors of the appraisal processes are of interest for the context of technology usage: the 1) 'characteristics of the empathizer' and 2) 'his/her relationship with the target' (De Vignemont and Singer, 2006, p. 440; see also; Anderson and Keltner, 2002). Referring to the first aspect, Davis (1983) showed that gender had an impact: female participants reported greater levels of empathy than male participants. In contrast, (Rosenthal-von der Puetten et al., 2013) did not find a significant effect of gender on empathy towards their robot. They indicated a rather inconsistent state of research, which they ascribed to different definitions of empathy studies referred to, as well as to different methods and measures studies used (Rosenthal-von der Puetten et al., 2013, p. 21). In sum, there are open questions left, which further research needs to elaborate on.

Secondly, regarding the relationship with the target, Rosenthal-von der Puetten et al. (2013) focused on 'acquaintance' operationalized as two forms of 'prior interaction' with the robot: before the actual experiment started, the experimental group had interacted with the robot for 10 min, while the control group had no prior contact. Results revealed that ‘prior interaction’ had no effect on the level of participants’ empathy. In contrast to research on a certain robot, studies analyzing voice assistants need to reconsider the operationalization of prior contact. Considering that voice assistants can be regarded as a state-of-the-art technology, which has become part of the everyday lives of an increasing number of users, we suggest that the ‘relationship with the target’ seems to be more complex compared to the analysis of social robots, which are still barely used outside of laboratories. Smartphones might serve as a model. Carolus et al. (2019b) argued that smartphones constitute ‘digital companion,’ which ‘accompany their users throughout the day’ with the result that they could be more to their users than just a technological device. By transferring characteristics and outcomes known from human-human relationships to the smartphone-user relationship, they introduced the idea of a ‘digital companionship’ between smartphone users and their devices. In their study, they offered empirical support for their theoretical conceptualization, concluding their concept of companionship to be a ‘fruitful approach to explain smartphone-related behaviors’ (p. 915). The present study carries forward their idea and considers voice assistants to also be a potential ‘digital companion.’ Consequently, characteristics constituting this kind of relationship are to be focused on. From the various aspects characterizing a relationship, the way the interaction partners address each other is regarded as a first indicator offering valuable insights. The style of address, the form of greeting and the pronouns used refer to a complex system within communication, facilitating social orientation. Social relationships are expressed in the way conversational partners address each other, for example (e.g., Ervin-Tripp, 1972). In this context, the name refers to a certain individual and indicates a certain familiarity with the use of the forename strengthening the process. In this regard, names refer to a “nucleus of our individual identity” (Pilcher, 2016). Technology groups adopt these principles and give their human forenames. Amazon’s Echo is better known as ‘Alexa,’ which is actually not the name but a wake word of the assistant. Although there are more options of wake words (Echo, Amazon or Computer) ‘Alexa’ has become the popular address of the device, again indicating human preferences for an allegedly human counterpart.

Consequently, analyzing potential empathetic reactions to CAs requires considering both the users’ interindividual differences and indicators of the relationship the users might have with the device.

To summarize, the present paper refers to the concept of media equation and the idea of computers constituting social actors, which trigger social reactions in their human users originally exclusive for human-human interaction. Considering the technological progress and digital media devices pervading our lives, the cognitive, emotional as well as conative reactions to this state-of-art technology are of great scholarly and practical interest. Furthermore, empathy as a constituting factor of social cooperation and prosocial behavior and resulting social relationships is a significant focus of research, offering insights into both how technology affects humans and how these human users react, in return. Studies so far provided valuable contributions to the field but focused on objects of research, which do not closely represent current usage of digital technology and which involve a variety of social cues resulting in confounding effects. The present paper continues the analysis of the empathetic impact of technology but identifies smart speakers to be the more externally valid research object. First, they constitute ‘the next natural form of HCI’ as voice controlling adopts a basic principle of humanity. Second, voice assistants have become increasingly popular, offering a variety of applications, which end-consumers are already using in everyday life. Third, this new way of using technology affects not only the single user but results in a social usage situation as other persons present become observers or parallel users of the human-technology interaction.

HYPOTHESES AND RESEARCH QUESTIONS

To analyze an observer’s empathetic reactions to a voice assistant, this study adopts the basic idea of Rosenthal-von der Puetten et al. (2013). Thus, the first hypothesis postulates a difference between the observation of neutral and rude treatment of a voice assistant. Because the observers’ general tendency to empathize has been found to be a significant predictor of elicited empathetic reactions, this interindividually different predisposition, henceforth referred to as trait empathy, needs to be considered. Consequently, the postulated difference of the first hypothesis needs to be controlled for trait empathy:

Hypothesis 1: While controlling for the observers’ trait empathy, watching the voice assistant being treated rudely results in significantly more empathy with the assistant than watching it being treated neutrally. Following the three dimensions of empathy introduced by Rosenthal-von der Puetten et al. (2013) we distinguished three sub-hypotheses focusing on one of the three dimensions of empathy: While controlling for the observers’ trait empathy, watching the voice assistant being treated rudely results in significantly more. . .

H1a: . . .pity for the assistant than watching it being treated neutrally.

H1b: . . .empathy with the assistant than watching it being treated neutrally.

H1c: . . .attribution of feelings to the assistant than watching it being treated neutrally.

The form of address has been introduced to constitute an important characteristic of a social relationship and to contribute to social orientation. Calling the technological entity by an originally human forename is a core aspect of operating voice assistants such as Amazon Echo, which is named ‘Alexa.’ Consequently, hypothesis 2 postulates that the way the device is addressed influences an observer’s empathetic reaction–while trait empathy is controlled for again. In line with hypothesis 1, three sub-hypotheses are postulated to account for the three dimensions of empathy.

Hypothesis 2: While controlling for trait empathy, watching the voice assistant being called ‘Alexa’ results in significantly more. . .

H2a: . . .pity for the assistant than it being called 'Computer.'

H2b: . . .empathy with the assistant than it being called 'Computer.'

H2c: . . .attribution of feelings to the assistant than it being called 'Computer.'

Furthermore, two explorative questions are posed referring to influencing factors for which research so far has revealed contradicting results. Inconsistent results of the role of the subjects' gender regarding empathy with technology leads to **Research Question 1:** Do men and women differ regarding empathy with an assistant being treated rudely or neutrally?

The relationship with the target has been argued to be a potentially modulatory factor. However, to describe social relationships adequately a variety of variables would need to be considered. Hypothesis 2 focuses on the form of address as one constituting characteristic. Furthermore, prior contact with the device is regarded as an additional indicator. Contradicting this assumption, Rosenthal-von der Puetten et al. (2013) did not find an effect resulting from prior interaction with their research object. However, they operationalized prior interaction as a 10-min period of interaction. Focusing on voice assistants, which are more common in everyday life, allows a more externally valid operationalization when asking for prior experience in real life and outside of the laboratory. Consequently, **Research Question 2** asks for the effects of experience on empathetic reactions: Does prior experience influence the empathy toward an assistant being treated rudely or neutrally?

METHODOLOGICAL APPROACH

A total of 140 participants engaged in an online experiment, ranging in age from 17 to 79 years (M = 30.14, SD = 13.9), with 64% women. Most participants were highly educated: 43% were students in higher education, 34% had finished university and 13% had finished vocational training. Regarding smart speaker experience, 51% have interacted occasionally and 40% have never interacted with one before. Only 13 participants reported to own a smart speaker (12 participants owned the Amazon device, 1 owned Google Home).

Procedure and Experimental Design

The online experiment started with a brief instruction about the broad purpose of the study and the ethical guidelines laid out by the German Psychological Association. Afterward, in a 2 × 2 experimental design, participants were randomly assigned to one out of four conditions to watch a video showing a man who prepares a meal while interacting with Amazon Echo. When his commands resulted in error messages, the man reacted either rudely or neutrally (factor 1: treatment). Furthermore, he addressed the agent by 'Alexa: or by 'Computer' (factor 2: form of address).

Stimulus Material

In line with the 2 × 2 design presented above, four pre-recorded videos were used as stimulus material. In the videos, a man was preparing a meal in his kitchen. Simultaneously, he talked to Amazon Echo, which was standing on the table right in front of him (see **Figure 1**). He instructs the device to do several tasks like booking a hotel, asking for a train connection and sending messages. To minimize any possible influence of sympathy or antipathy, the face of the protagonist was never visible in the videos. Only his upper body was filmed. To ensure comparability of the four video conditions, the videos were produced as equally as possible regarding script, camera angle and film editing. Hence, the video set and the actor were the same across conditions. Four cameras were used to shoot, and the camera settings were not changed during or between the shooting of the different videos. In post-production, video editing was kept constant across all conditions. To ensure a controlled dialogue, voice outputs of the device were pre-programmed. We used the chat platform Dexter (https://rundexter.com/) to create a skill involving the sequences of the dialogue, which was implemented using Amazon Web Services (https://aws.amazon.com/). In contrast to the videos of previous studies, we avoided a rather unrealistic or extreme story but were guided by common usage scenarios of smart speakers (Handley, 2019).

In all four videos, the plots were basically identical. During preparation, the man's commands became more and more complex. His rather vague commands became difficult to execute. Consequently, more and more commands failed, which allowed us to implement the treatment-factor: the man either proceeded with his neutral commands (**neutral condition**) or he got angry when commands failed and acted rudely (**rude condition**). In the neutral condition, the man recognizes his operating errors and corrects himself. He speaks calmly using neutral language. In the rude condition, the man increasingly furiously during the interaction. He scolds the device using foul names and starts yelling. Finally, the man shoves away the assistant. The form of address-factor was realized by implementing two different forms of address: either 'Alexa' (**'Alexa' condition**) or 'Computer' (**'Computer' condition**). The length of the four resulting videos were kept constant, varying 2 s only (4:49–4:51 min). In terms of content, the outlined minimal changes in the script resulted in minimal adjustments of the storyline (see **Table 1**). For example, the foul name the man used to address the device in the 'computer' condition was changed from 'snipe' into 'tin box.' To warrant a conclusive storyline the reaction of the device needed to be adapted resulting in slight differences of the interaction in the rude condition compared to the neutral condition. In a preliminary study, these differences were analyzed to warrant the comparability of the stimulus material.

Preliminary Study: Development of the Stimulus Material

A pretest was conducted to ensure 1) the assignment of the 'rude' and the 'neutral' treatment and–besides that postulated difference-2) the comparability between these two videos. 89 participants (82% women) engaged in an online experiment,

FIGURE 1 | Screenshots of the video the participants watched.

TABLE 1 | Text passages from the video.

Rude and "Alexa"	Neutral and "Computer"
Man: Alexa, book me this hotel room	**Man**: Computer, book me this hotel room
Assistant: Unfortunately, it is not possible for me to book a hotel room	**Assistant**: Unfortunately, it is not possible for me to book a hotel room
Man: Are you serious? Why not? Alexa, why not?	**Man**: Computer, why not?
Assistant: No payment information has been deposited so far. Do you want to add a credit card now?	**Assistant**: No payment information has been deposited so far. Do you want to add a credit card now?
Man: Alexa, why not? -- Wow, do I have to do this myself now?	**Man**: No

ranging in age from 18 to 36 years (M = 20.57, SD = 2.57). They were evenly distributed among the four conditions. 14 participants reported to own a smart speaker.

First and to ensure rudeness vs. neutrality, a two-sided t-test for independent samples revealed significant differences in the evaluations of the man, who interacted with the device. In line with our postulated assignment, the man was rated significantly less attractive ($t_{(87)} = -12.66$, $p < 0.001$) in the rude (M = 3.31; SD = 0.91) than in the neutral condition (M = 5.76; SD = 0.91) (Schrepp et al., 2014). The general positivity (Johnson et al., 2004) differed significantly between the rude (M = 3.68; SD = 0.81) and the neutral condition (M = 6.40; SD = 0.94), $t_{(87)} = 14.57$, $p <$ 0.001. Moreover, four more single items (see **Supplementary Material** for the detailed list) confirmed the postulated differences: the 'rude condition' was rated to be more unobjective ($t_{(87)} = 11.80$, $p < 0.001$), impolite ($t_{(87)} = 14.71$, $p < 0.001$.), aggressive ($t_{(87)} = 15.71$, $p < 0.001$.) and violent ($t_{(87)} =$ 11.8, $p < 0.001$.).

Second and to ensure the comparability between these two videos, participants evaluated the device. The exact same measures used to evaluate the man were used again (see **Supplementary Material**). Comparing the 'Alexa' and the 'computer' condition, the evaluations of the device did not differ regarding attractiveness ($t_{(87)} = 0.15$, $p = 0.883$) and the general positivity towards to the device ($t_{(87)} = 0.68$, $p = 0.500$). Likewise, the semantic differentials revealed no significant differences regarding unobjectiveness ($t_{(87)} = -7.88$, $p = 0.433$), impoliteness ($t_{(87)} = -1.34$, $p = 0.184$), aggressiveness ($t_{(87)} = -1.21$, $p = 0.230$) and violence ($t_{(87)} = -1.55$, $p < 0.125$.).

To summarize, the pretest ensures the validity of the stimulus material. The rude condition did significantly differ regarding perceived rudeness, which can be ascribed to differences in the man's behavior. Evaluations of the device, however, did not differ significantly between the 'rude' and the 'neutral' condition.

Measures

After watching the video, participants answered a questionnaire asking for 1) the empathy with the voice assistant, 2) their trait empathy, 3) their prior experience with smart speakers and 4) demographic information.

To assess **empathy with the voice assistant**, 22 items, based on the items used by Rosenthal-von der Puetten et al. (2013) were presented. According to the affective component of empathy, the scale includes items addressing 'feelings of pity' (e.g., 'I felt sorry

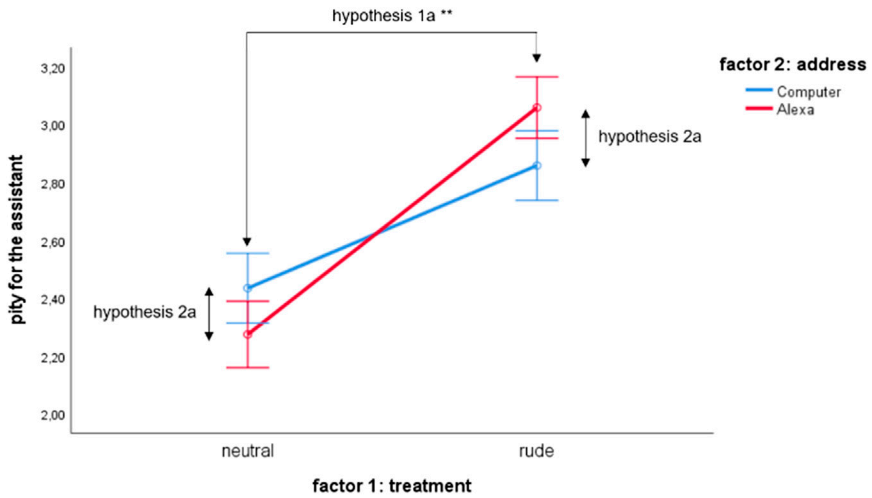


FIGURE 2 | Effects of treatment and form of address on pity for the CA (controlled for trait empathy).

for the voice assistant.''). To assess the cognitive component, the scale incorporated items asking for 'empathy' (e.g., 'I could relate to the incidents in the video'). Furthermore, an attribution of feelings to the device was assessed by ten items (e.g., 'I can imagine that . . . the voice assistant suffered.'). Since we did not focus on a quantitative graduation of observers' responses in our study, arousal was not assessed. The items were answered on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Items were averaged so that higher values indicated higher levels of empathy with the assistant. Internal consistency of the scale was $\alpha = 0.82$.

Trait empathy was measured using the Saarbruecker Personality Questionnaire (SPF), a German version of the Interpersonality Reactive Index (IRI) by Paulus (2009). In sum, 21 items were answered on a 5-point Likert scale (i.e., 'I have warm feelings for people who are less well off than me.''). Again, items were averaged with higher values indicating higher levels of trait empathy. Internal consistency of the scale was $\alpha = 0.83$.

Prior experience with voice assistants was assessed by asking if the participant had 'ever interacted with a voice assistant' and if he/she uses 'a voice assistant at home.' The answering options were 'never,' 'a few times' and 'regularly.' Finally, participants were asked about their age, gender and education.

RESULTS

To analyze the impact of treatment (factor 1: rude vs. neutral) and the impact of form of address (factor 2: 'Alexa' vs. 'Computer') on participants' empathy with the voice assistant, three two-way ANCOVAs were conducted controlling for participants' trait empathy as the covariate. To test **hypothesis 1**, while controlling trait empathy, the impacts of the rude vs. the neutral condition on empathy with the assistant were compared–with H1a focusing on pity, H1b on empathy and H1c on attribution of feelings. Regarding H1a, the covariate trait empathy was not significantly related to the intensiveness of pity with the voice assistant, $F_{(1,135)} = 1.72$, $p = 0.192$, partial $\eta^2 = 0.01$. In line with H1a, participants who observed the assistant being treated rudely reported a significantly higher level of pity with the device than participants of the neutral condition, $F_{(1,\ 135)} = 27.13$, $p < 0.001$, partial $\eta^2 = 0.17$ with partial eta-squared indicating a large effect (Cohen, 1988). **Figure 2** shows the results. The results of H1b showed that the covariate trait empathy was again not significant, $F_{(1,135)} = 0.47$, $p = 0.496$, partial $\eta^2 < 0.001$. According to hypothesis 1b, there was again a significant main effect of the factor treatment, $F_{(1,135)} = 10.04$, $p = 0.002$, partial $\eta^2 = 0.07$, indicating a medium effect of rude vs. neutral treatment on the subscale empathy (see **Figure 3**).

Finally, H1c again revealed a non-significant covariate $F_{(1,135)} = 2.03$, $p = 0.157$, partial $\eta^2 = 0.015$. The way the device was treated again resulted in significant differences, $F_{(1,135)} = 4.51$, $p = 0.036$, partial $\eta^2 = 0.032$, which is interpreted as a small effect on the subscale attribution (see **Figure 4**).

In sum, all subscales of the empathy-scale revealed significant results in line with the expectations. Rude treatment led to more empathy compared to neutral treatment.

In contrast, as **Figures 2–4** show, the two different forms of address (**hypothesis 2**) did not result in significant differences (pity subscale: $F_{(1,135)} = 0.03$, $p = 0.861$, partial $\eta^2 < 0.001$; empathy subscale: $F_{(1,135)} = 0.02$, $p = 0.882$, partial $\eta^2 < 0.001$; attribution subscale: $F_{(1,135)} = 0.063$, $p = 0.429$, partial $\eta^2 = 0.01$). All three ANCOVAs conducted showed no significant interaction terms of the two factors treatment and form of address, neither for pity ($F_{(1,135)} = 2.43$, $p = 0.122$, partial $\eta^2 = 0.02$), nor for the empathy subscale ($F_{(1,135)} = 0.59$, $p = 0.444$, partial $\eta^2 < 0.001$) and the attribution subscale ($F_{(1,135)} = 0.88$, $p = 0.350$, partial $\eta^2 = 0.01$). Consequently, all three sub-hypotheses 2a–2c were rejected. Although the effect was not significant, participants of the rude condition reported the highest level of pity when the voice assistant was called 'Alexa' (see **Figure 2**).

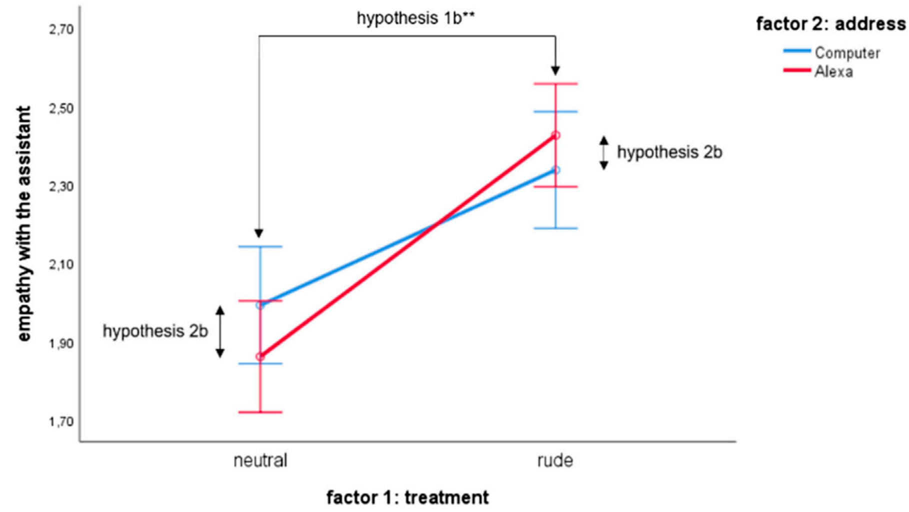


FIGURE 3 | Effects of treatment and form of address on empathy subscale with the CA (controlled for trait empathy).

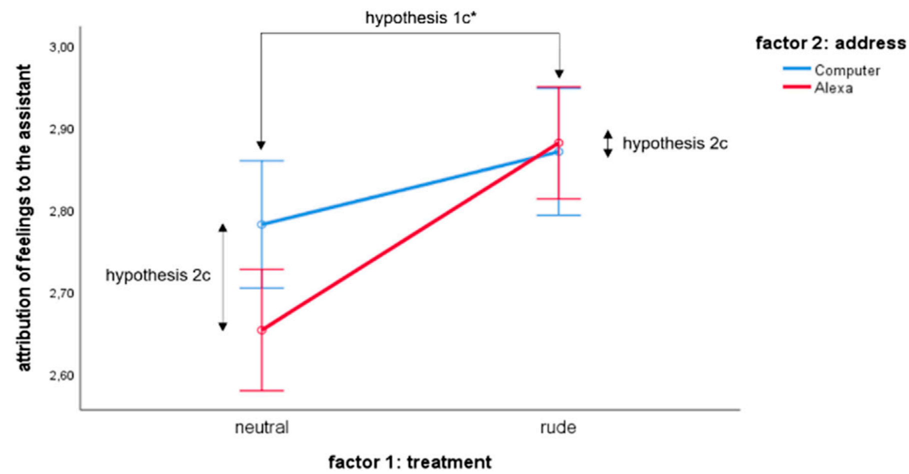


FIGURE 4 | Effects of treatment and form of address on attribution of feelings to the CA (controlled for trait empathy).

To analyze potential differences between male and female participants (**research question 1**), three two-way ANCOVAs were conducted, again. However, factors were different than in the analyses reported before: Because the form of address was shown not to result in significant effects, it was eliminated from the following analyses. Instead, participants' gender was analyzed as the second factor, treatment was kept as the first factor and trait empathy as the covariate. **Figure 5** gives an overview of the results. Regarding pity, results revealed no significant main effect of gender, $F_{(1,135)} = 0.19$, $p = 0.662$, $\eta^2 < 0.001$. Moreover, the covariate was not significant, $F_{(1,135)} = 0.96$, $p = 0.329$, $\eta^2 = 0.01$; but the main effect of treatment was, $F_{(1,135)} = 27.14$, $p < 0.001$, $\eta^2 = 0.17$. Similarly, when analyzing the empathy subscale, the main effect of gender was not significant, $F_{(1,135)} = 2.72$, $p = 0.101$, $\eta^2 = 0.02$. Again, the covariate trait empathy was not significant $F_{(1,135)} = 0.002$, $p = 0.962$, $\eta^2 < 0.001$; but the main effect of treatment was $F_{(1,135)} = 9.14$, $p = 0.003$, $\eta^2 = 0.06$. Finally, attribution of feelings for the assistant again revealed no significant gender effect, $F_{(1,135)} = 0.09$, $p = 0.772$, $\eta^2 < 0.001$. Again, the covariate was not significant, $F_{(1,135)} = 1.74$, $p = 0.189$, $\eta^2 = 0.01$. In contrast to previous results, the effect of treatment was not significant $F_{(1,135)} = 2.46$, $p = 0.119$, $\eta^2 = 0.02$. Moreover, the interaction term almost reached significance, $F_{(1,135)} = 3.30$, $p = 0.07$, $\eta^2 = 0.02$, indicating that in the rude condition only female but not male participants reported higher attributions of feelings to the assistant. Nevertheless, to summarize results of question 1, men and women do not differ regarding all three subtypes of empathy with the assistant.

Research question 2 asked for the effect of prior experience with voice assistants on the reported empathy with the assistant watched in the video. In line with the preceding analyses, the approach of three ANCOVAs was retained. Furthermore, factor 1 (treatment) was

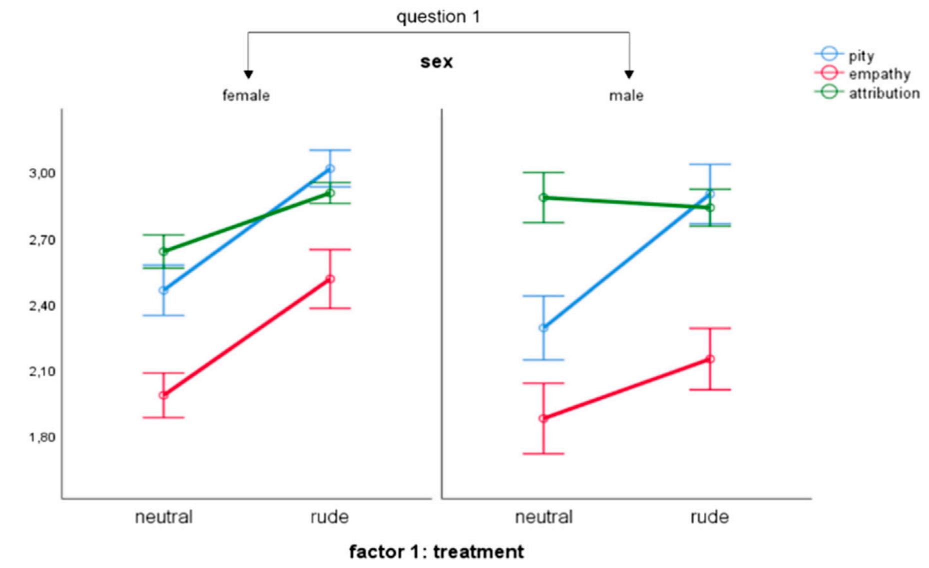


FIGURE 5 | Effects of treatment and participant's gender on empathy with the CA (controlled for trait empathy).

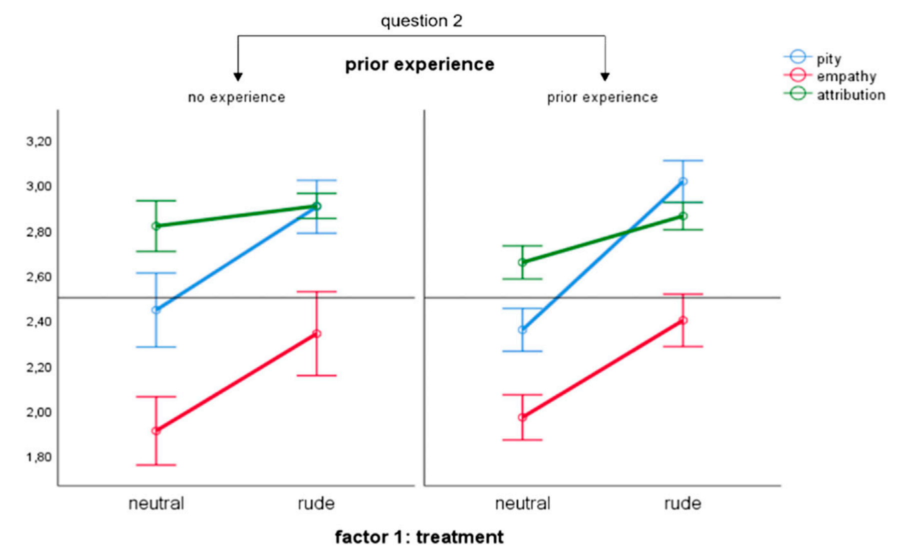


FIGURE 6 | Effects of treatment and participant's prior experience on empathy with the CA (controlled for trait empathy).

retained and prior experience was added as factor 2. Again, trait empathy was kept as the covariate. **Figure 6** gives an overview of the results. Regarding pity, results revealed no significant main effect of prior experience, $F_{(1,135)} = 0.3$, $p = 0.338$, $\eta^2 = 0.01$. While the covariate was not significant, $F_{(1,135)} = 1.48$, $p = 0.226$, $\eta^2 = 0.01$, the main effect of treatment was, $F_{(1,135)} = 25.49$, $p < 0.001$, $\eta^2 = 0.16$. Similarly, when analyzing the subscale empathy, the main effect of prior experience was not significant, $F_{(1,135)} = 0.58$, $p = 0.448$, $\eta^2 = 0.02$. Again, the covariate was not significant, $F_{(1,135)} = 0.38$, $p = 0.539$, $\eta^2 < 0.01$ but the main effect of treatment was, $F_{(1,135)} = 9.57$, $p = 0.002$, $\eta^2 = 0.07$. Finally, attribution of feelings for the assistant again revealed no significant effect of prior experience, $F_{(1,135)} = 1.08$, $p = 0.302$, $\eta^2 = 0.01$. Once again, the covariate was not significant, $F_{(1,135)} = 2.17$, $p = 0.143$, $\eta^2 = 0.02$, but the main effect of treatment was, $F_{(1,135)} = 4.63$, $p = 0.033$, $\eta^2 = 0.03$. To conclude, participants with or without prior experience with voice assistants do not differ regarding all three subtypes of empathy with an assistant they watched in the video.

DISCUSSION

The present study focuses on empathetic reactions to smart speakers, which have become everyday technology for an increasing number of users over the last few years. Following the approach of 'computers as social actors' introduced in the 1990s, we argue that the basic idea that 'media equals real life' is amplified by voice-based operation of devices. Voice control refers to a basic principle of humanity, which has been exclusive for human-human interactions until recently. With the adaptation of this feature, shared commonalities of human-human interactions and interactions with CAs were derived. Just like interhuman conversations can be listened to, attendees of voice-based operations within households can pay attention, for example. Therefore, research on the emotional impact of conversational agents expands its perspective and involves the impact on further attendees' cognitive, emotional and conative reactions. Empathetic reactions as a constituting factor of social cooperation, prosocial behavior, and resulting social relationships were shown to be a promising empirical starting point (e.g., Rosenthal-von der Puetten et al., 2013; Kraemer et al., 2015). To bridge the research gap regarding 1) more real usage scenarios and everyday-relevant technology as well as 2) devices with reduced anthropomorphistic cues, this study focuses on attendees' empathetic reaction to an interaction between a user and a smart speaker.

There are **two key findings** of the present research. First, compared to neutral treatment, treating the CA rudely resulted in higher ratings of empathy. Second, all the other factors analyzed did not have significant effects. Neither the participants' characteristics (trait empathy, gender, prior experience) nor the way the assistant was addressed ('Alexa' vs. 'Computer') influenced the observers' empathy with the device. Referring to the basic idea of Reeves and Nass (1996), who postulate that media equation 'applies to everyone, it applies often, and it is highly consequential' our results indicate that CAs elicit media equation effects, which are 'highly consequential' in terms of its independence of every influencing factor analyzed in this study. These results contradict our theoretical explanations considering both the impact of participants' individual characteristics and the relationship people might have with technological devices. However, post-hoc explorative analyses revealed three partial results, which seem to be worth noting as they could be carefully interpreted as possibly indicating further implications. 1) Participants of the rude condition would report the highest level of pity if the voice assistant was called "Alexa" (see **Figure 2**). This might indicate a (non-significant) impact of the form of address. 2) Within the rude condition, only female but not male participants reported higher attributions of feelings to the CA indicating a (non-significant) impact of gender. 3) The effects on the subscale of pity were stronger than for the empathy subscale followed by only small effects for the attribution subscale, indicating that the stimuli took effect on different aspects of empathy.

Limitations and Directions for Future Research

Although not statistically significant, the explorative results suggest a need for further research on interindividual differences between the participants (i.e., gender) and on interindividually different relationships people have with their devices (i.e., address). Future studies need to draw more heterogenous samples to analyze these potential effects more profoundly. Moreover, the variables the present study focused on need to be further elaborated. Regarding interindividual differences beyond gender, Kraemer et al. (2015) compiled further characteristics relevant to consider. Age, computer literacy and the individual's personality were shown to be potential influencing factors, which future research needs to transfer to users interacting with modern voice-based devices. Regarding interindividually different relationships, this study took 'prior experience' as a first indicator. Differentiating between no prior use vs. prior use is only the first step on the way to a comprehensive assessment of the postulated social relationship with a voice assistant stemming from interactions over time. However, referring to the concept of a 'digital companionship' (Carolus et al., 2019b), prior experience needs to be assessed in a more detailed way involving variables such as closeness to the voice assistant, trust in the assistant and preoccupation with it seem to be relevant characteristics of the relationship. Further constituting outcomes could be stress caused by the assistant as well as the potential to cope with stressful situations with the support of the assistant. Future research needs to incorporate these variables to analyze the effects of the relationship users have with their devices. Consequently, future research needs to reflect upon the target device the participants are confronted with. Arguing from the perspective of an established user-device relationship, future studies could use the participant's own device as the target device–or at least a CA like the participant's own device as similarity has been shown to be an important condition of empathy (Serino et al., 2009). Presenting a video with a rather random device a foreigner is interacting with does not fulfill these requirements convincingly. According to measures, different aspects of empathy might be activated when observing a technological device being treated rudely, compared to observing a human and animal. In addition, the level of emotional reaction (e.g., arousal) might be different. Finally, as this study did only focus on participants feelings for the technological part of the observed interaction, future research could also ask for the human counterpart. Being confronted with a device, which seems to not work properly, and which does not express empathy with the user's struggle, future studies could also analyze the empathy participants have with the unsuccessful user of the device.

Interpretations of the results presented are faced with further methodological limitations of the study. 1) Participants only watched a video but did not observe a real-life interaction between another person and the device. Moreover, the interaction observed was not a real-life interaction but was performed by an actor resulting in questionable realism. Our ongoing development of the approach took this shortcoming into account and applied the approach to an experimentally manipulated real-life scenario. 2) To gain first insights into potentially influencing variables we controlled for trait empathy, gender and prior experience. However, research we have presented in this paper argues that further variables need to be considered. As outlined before, future studies need to include the variables operationalizing 'characteristics of the empathizer' and 'his/her relationship with the target' in a more detailed way (De Vignemont and Singer, 2006, p. 440). Furthermore, from a psychological perspective, psychological variables relevant in the context of social relationships and social interactions need to be considered, e.g., the need for affiliation, loneliness as well as affectivity or the participant's emotional state 3) In addition, more aspects of empathy (e.g., the cognitive component according to Davis (1983); Rogers et al. (2007)). and the level of reactions should be considered by addressing a more diverse methodical approach (e.g., arousal measures, visual scales). Moreover, we limited ourselves to the analysis of empathy as the dependent variable. What we regard as a first promising starting point needs to be expanded in future studies. Especially in view of the social aspect of the use of voice assistants introduced, there are various effects to focus on, e.g., envy or jealousy, which may be elicited by the device as well as affection or attachment. 4) Lastly, and regarding the briefly introduced issue of gender stereotypes, manipulating the characteristics of the voice assistant itself are to be focused on. Changing the female voice into a male voice or changing the 'personality' of the device (e.g., neutral vs. rude answers given by the device itself) will be the subject of future studies.

Study Implications and Conclusion

Despite the outlined limitations, our results suggest several theoretical and practical implications. To put the study presented in a nutshell: People watching users treating voice assistants badly will empathize with the device. What sounds irrational at first, can be explained in the light of media equation and becomes a valuable insight for developers, researchers and users. First, developers should be aware of the human the final consumers' psychological processes. Devices do not need to have anthropomorphic features to elicit social reactions in their human counterpart. Although they are consciously recognized as technological devices, they might trigger social reactions. Therefore, we argue that psychological mechanisms regulating human social life (norms, rules, schemata) are a fruitful source for developers and programmers when designing the operation of digital devices. Knowing how humans tend to react offers possibilities to manipulate these reactions–in a positive as well as in a negative way. For example, knowing that users empathize can be used to increase acceptance of misunderstandings or mistakes of devices. Furthermore, knowing that users feel for their CA can be used to counteract abusive. Developers could intentionally address the users' tendency to transfer social rules originally established for human-human interactions to digital devices. Or, when the observer empathizes and reminds the user of an appropriate behavior. However, knowing users' psychology also allows to manipulate them in a less benevolent way. Companies can adopt the psychological mechanism to bind users to their services and products and to maximize their profits. Second, researchers are encouraged to adopt the results presented to further analyze interwoven effects of the users' psychology and the processing and functioning of the technological equipment. Together with further societal actors, conclusions should be drawn regarding educational programs which enable users to keep pace with technological progress and to develop media literacy skills. Competent users are key factors of our shared digital future. Third, and as a consequence of the developers' as well as the researchers' responsibilities, users need to accept and adopt the opportunities and be prepared for the challenges of the digital future which has already started.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AC: Conceptualization, Methodology, Writing-Original Draft, Writing - Review and Editing, Supervision, Project administration. CW: Formal analysis, Data Curation, Writing - Review and Editing. AT: Conceptualization, Methodology, Investigation. TF: Conceptualization, Methodology, Investigation. CS: Conceptualization, Methodology, Investigation. MS: Conceptualization, Methodology, Investigation.

FUNDING

This publication was supported by the Open Access Publication Fund of the University of Wuerzburg.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2021.682982/full#supplementary-material

REFERENCES

Anderson, C., and Keltner, D. (2002). The Role of Empathy in the Formation and Maintenance of Social Bonds. *Behav. Brain Sci.* 25 (1), 21–22. doi:10.1017/s0140525x02230010

Bartneck, C., and Hu, J. (2008). Exploring the Abuse of Robots. *Is* 9 (3), 415–433. doi:10.1075/is.9.3.04bar

Bartneck, C., and Keijsers, M. (2020). The Morality of Abusing a Robot. *J. Behav. Robotics* 11 (1), 271–283. doi:10.1515/pjbr-2020-0017

Brahnam, Sheryl., and De Angeli, Antonella. (2008). *Special Issue on the Abuse and Misuse of Social Agents*. Oxford, UK: Oxford University Press.

Brščić, Dražen., Kidokoro, Hiroyuki., Suehiro, Yoshitaka., and Kanda, Takayuki. (2015). "Escaping from Children's Abuse of Social Robots. *In Proc. Tenth Annu. Acm/Ieee Int. Conf. Human-Robot Interaction*, 59–66.

BusinessWire (2018). Strategy Analytics: Google Home Mini Accounts for One in Five Smart Speaker Shipments Worldwide in Q2 2018, September 19, 2018. Available at: https://www.businesswire.com/news/home/20180919005089/en/Strategy-Analytics-Google-Home-Mini-Accounts-for-One-in-Five-Smart-Speaker-Shipments-Worldwide-in-Q2-2018.

Buss, D. M., and Kenrick, D. T. (1998). Evolutionary social psychology. *The handbook of social psychology*. Editors D. T. Gilbert, S. T. Fiske, and G. Lindzey (McGraw-Hill), 982–1026.

Cannon, A. (2017). 20 Ways Amazon's Alexa Can Make Your Life Easier." Wise Bread. Available at: https://www.wisebread.com/20-ways-amazons-alexa-can-make-your-life-easier (Accessed March 17, 2021).

Carolus, A., Binder, J. F., Muench, R., Schmidt, C., Schneider, F., and Buglass, S. L. (2019b). Binder, Ricardo Muench, Catharina Schmidt, Florian Schneider, and Sarah L. Buglass.Smartphones as Digital Companions: Characterizing the Relationship between Users and Their Phones. *New Media Soc.* 21 (4), 914–938. doi:10.1177/1461444818817074

Carolus, A., Muench, R., Schmidt, C., and Schneider, F. (2019a). Impertinent Mobiles - Effects of Politeness and Impoliteness in Human-Smartphone Interaction. *Comput. Hum. Behav.* 93, 290–300. doi:10.1016/j.chb.2018.12.030

Carolus, Astrid., Schmidt, Catharina., Muench, Ricardo., Mayer, Lena., and Schneider, Florian. (2018). "Pink Stinks-At Least for Men." In International Conference on Human-Computer Interaction. Springer, 512–525.

Chin, Hyojin., Molefi, Lebogang. Wame., and Yi, Mun. Yong. (2020). Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse, *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.* (New York, NY: . Association for Computing Machinery), 1–13. doi:10.1145/3313831.3376461

Cohen, J. (1998). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY, United States: Routledge Academic.

Cosmides, Leda., and Tooby, John. (1994). *Origins of Domain Specificity: The Evolution of Functional Organization*. Citeseer.

Cramer, Henriette., Goddijn, Jorrit., Wielinga, Bob., and Evers, Vanessa. (2010). "Effects of (In) Accurate Empathy and Situational Valence on Attitudes towards Robots." In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 141–142. doi:10.1145/1734454.1734513

Cuff, B. M. P., Brown, S. J., Taylor, L., and Howat, D. J. (2016). Empathy: A Review of the Concept. *Emot. Rev.* 8 (2), 144–153. doi:10.1177/1754073914558466

Davis, M. H. (1983). Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *J. Personal. Soc. Psychol.* 44 (1), 113–126. doi:10.1037/0022-3514.44.1.113

Dawkins, Richard. (2016). *The Selfish Gene*. Oxford University Press.

De Vignemont, F., and Singer, T. (2006). The Empathic Brain: How, when and Why? *Trends Cogn. Sci.* 10 (10), 435–441. doi:10.1016/j.tics.2006.08.008

Duan, C., and Hill, C. E. (1996). The Current State of Empathy Research. *J. Couns. Psychol.* 43 (3), 261–274. doi:10.1037/0022-0167.43.3.261

Ervin-Tripp, Susan. M. (1972). "Sociolinguistic Rules of Address," in *Sociolinguistics*. Editors J. B. Pride and J. Holmes (Harmondsworth: Penguin), 225–240.

Feine, J., Gnewuch, U., Morana, S., and Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *Int. J. Human-Computer Stud.* 132, 138–161. doi:10.1016/j.ijhcs.2019.07.009

Ganczarek, Joanna., Hueünefeldt, Thomas., and Belardinelli, Marta. Olivetti. (2018). *From "Einfühlung" to Empathy: Exploring the Relationship between Aesthetic and Interpersonal Experience*. Springer.

Go-Gulf (2018). "The Rise of Virtual Digital Assistants Usage – Statistics and Trends." GO-Gulf. (Blog). April 27, 2018. Available at: https://www.go-gulf.com/virtual-digital-assistants/.

Gonsior, Barbara., Sosnowski, Stefan., Buß, Malte., Wollherr, Dirk., and Kühnlenz, Kolja. (2012). "An Emotional Adaption Approach to Increase Helpfulness towards a Robot." In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2429–2436.

Handley, Lucy. (2019). "More Than 50% of People Use Voice Assistants to 'ask Fun Questions.' Here's what Else They Do." CNBC. February 26, 2019. Available at: https://www.cnbc.com/2019/02/26/what-do-people-use-smart-speakers-for-asking-fun-questions.html.

Hegel, Frank., Torsten SpexardWrede, Britta., Horstmann, Gernot., and Vogt, Thurid. (2006). "Playing a Different Imitation Game: Interaction with an Empathic Android Robot." In 2006 6th IEEE-RAS International Conference on Humanoid Robots. IEEE, 56–61.

Hirschberg, J., and Manning, C. D. (2015). Advances in Natural Language Processing. *Science* 349 (6245), 261–266. doi:10.1126/science.aaa8685

Hoffman, Martin. (2001). *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press.

Hoffman, Martin. L. (1982). "Development of Prosocial Motivation: Empathy and Guilt." In The Development of Prosocial Behavior. Elsevier, 281–313.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Med. Reference Serv. Q.* 37 (1), 81–88. doi:10.1080/02763869.2018.1404391

Jacques, Richard., Følstad, Asbjørn., Gerber, Elizabeth., Grudin, Jonathan., Luger, Ewa., Monroy-Hernández, Andrés., et al. (2019). "Conversational Agents: Acting on the Wave of Research and Development." In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 1–8.

Johnson, D., Gardner, J., and Wiles, J. (2004). Experience as a Moderator of the Media Equation: The Impact of Flattery and Praise. *Int. J. Human-Computer Stud.* 61 (3), 237–258. doi:10.1016/j.ijhcs.2003.12.008

Keynes, Milton. (2020). Strategy Analytics: New Record for Smart Speakers as Global Sales Reached 146.9 Million in 2019, February 13, 2020. Available at: https://www.businesswire.com/news/home/20200213005737/en/Strategy-Analytics-New-Record-for-Smart-Speakers-As-Global-Sales-Reached-146.9-Million-in-2019.

Kinsella, Bret. (2019). Over 20% of UK Households Have Smart Speakers while Germany Passes 10% and Ireland Approaches that Milestone. *Voicebot.Ai. October* 11, 2019, 2019 . Available at: http://voicebot.ai/2019/10/11/over-20-of-uk-households-have-smart-speakers-while-germany-passes-10-and-ireland-approaches-that-milestone/.

Kraemer, Nicole. C., AstridRosenthal-von der Pueütten, M., and Hoffmann, Laura. (2015). Social Effects of Virtual and Robot Companions. *Handbook Psychol. Commun. Tech.* 32, 137.

Kwak, Sonya. S., Kim, Yunkyung., Kim, Eunho., Shin, Christine., and Cho, Kwangsu. (2013). "What Makes People Empathize with an Emotional Robot?: The Impact of Agency and Physical Embodiment on Human Empathy for a Robot." In 2013 IEEE RO-MAN. IEEE, 180–185. doi:10.1109/roman.2013.6628441

Lee, E.-J., and Nass, C. (2002). Experimental Tests of Normative Group Influence and Representation Effects in Computer-Mediated Communication. *Hum. Comm Res.* 28 (3), 349–381. doi:10.1111/j.1468-2958.2002.tb00812.x

Lee, Eun. Ju., Nass, Clifford., and Scott, Brave. (2000). "Can Computer-Generated Speech Have Gender? an Experimental Test of Gender Stereotype." In CHI'00 Extended Abstracts on Human Factors in Computing Systems, 289–290.

Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2014). "Empathic Robots for Long-Term Interaction. *Int. J. Soc. Robotics* 6 (3), 329–341. doi:10.1007/s12369-014-0227-1

Leite, Iolanda., Henriques, Rui., Martinho, Carlos., and Paiva, Ana. (2013a). Sensors in the Wild: Exploring Electrodermal Activity in Child-Robot Interaction." In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 41–48. doi:10.1109/hri.2013.6483500

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., and Paiva, A. (2013). The Influence of Empathy in Human-Robot Relations. *Int. J. Human-Computer Stud.* 71 (3), 250–260. doi:10.1016/j.ijhcs.2012.09.005

Lipps, Theodor. (1903). *Ästhetik (Psychologie Des Schönen Und Der Kunst)*. Hamburg, Germany: L. Voss.

Luger, Ewa., and Sellen, Abigail. (2016). "Like Having a Really Bad PA" the Gulf between User Expectation and Experience of Conversational Agents." In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 5286–5297.

Macdougall, Robert. (1910). *Lectures on the Experimental Psychology of the Thought Processes*. JSTOR.
McTear, M., Callejas, Z., and Griol, D. (2016). *The Conversational Interface*. Cham: Springer, 51–72. m2016. doi:10.1007/978-3-319-32967-3_4 Conversational Interfaces: Past and Present.
Morishima, Y., Bennett, C., Nass, C., and Lee, K. M. (2002). *Effects of (Synthetic) Voice Gender, User Gender, and Product Gender on Credibility in E-Commerce.* " Unpublished Manuscript, Stanford, CA: Stanford University.
Nass, C., Fogg, B. J., and Moon, Y. (1996). Can Computers Be Teammates?. *Int. J. Human-Computer Stud.* 45 (6), 669–678. doi:10.1006/ijhc.1996.0073
Nass, C., and Gong, L. (2000). Speech Interfaces from an Evolutionary Perspective. *Commun. ACM* 43 (9), 36–43. doi:10.1145/348941.348976
Nass, Clifford., Steuer, Jonathan., Ellen, R., and Tauber. (1994). "Computers Are Social Actors." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 72–78.
Nass, C., Moon, Y., and Carney, P. (1999). Are People Polite to Computers? Responses to Computer-Based Interviewing Systems1. *J. Appl. Soc. Pyschol* 29 (5), 1093–1109. doi:10.1111/j.1559-1816.1999.tb00142.x
Nass, C., Moon, Y., and Green, N. (1997). Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers with Voices. *J. Appl. Soc. Pyschol* 27 (10), 864–876. doi:10.1111/j.1559-1816.1997.tb00275.x
Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., and Yamada, S. (2016). Why Do Children Abuse Robots? *Is* 17 (3), 347–369. doi:10.1075/is.17.3.02nom
Paiva, A., Leite, I., Boukricha, H., and Wachsmuth, I. (2017). Empathy in Virtual Agents and Robots. *ACM Trans. Interact. Intell. Syst.* 7 (3), 1–40. doi:10.1145/2912150
Paulus, Christoph. (2009). "Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) Zur Messung von Empathie: Psychometrische Evaluation Der Deutschen Version Des Interpersonal Reactivity Index. Available at: http://psydok.sulb.uni-saarland.de/volltexte/2009/2363/."
Perez, Sarah. n. d. (2019). Over a Quarter of US Adults Now Own a Smart Speaker, Typically an Amazon Echo." TechCrunch (Blog).Accessed November 17, 2019. Available at: https://social.techcrunch.com/2019/03/08/over-a-quarter-of-u-s-adults-now-own-a-smart-speaker-typically-an-amazon-echo/. doi:10.1037/e613692010-001
Petrock, Victoria. n. d. (2021). US Voice Assistant Users 2019." Insider Intelligence. Accessed March 17, 2021, Available at: https://www.emarketer.com/content/us-voice-assistant-users-2019. doi:10.21236/ada463921
Pilcher, J. (2016). Names, Bodies and Identities. *Sociology* 50 (4), 764–779. doi:10.1177/0038038515582157
Pinker, Steven. (1994). *The Language Instinct*. William Morrow & Co. doi:10.1037/e412952005-009
Porcheron, Martin., Fischer, Joel. E., McGregor, Moira., Brown, Barry., Luger, Ewa., Candello, Heloisa., et al. (2017). "Talking with Conversational Agents in Collaborative Action." In Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 431–436.
Porcheron, Martin., Fischer, Joel. E., Reeves, Stuart., and Sarah Sharples. (2018). "Voice Interfaces in Everyday Life." In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12.
Purington, Amanda., Taft, Jessie. G., Sannon, Shruti., Bazarova, Natalya. N., and Taylor, Samuel. Hardman. (2017). "Alexa Is My New BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo." In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2853–2859.
Pyae, Aung., and Joelsson, Tapani. N. (2018). "Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker." In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, 127–131.
Reeves, Byron., and Nass, Clifford. (1996). *The Media Equation: How People Treat Computers, Television, and New Media like Real People*. UK: Cambridge university press Cambridge.
Riek, Laurel. D., Rabinowitch, Tal-Chen., Chakrabarti, Bhismadev., and Robinson, Peter. (2009). "Empathizing with Robots: Fellow Feeling along the Anthropomorphic Spectrum." In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 1–6. IEEE.
Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., and Convit, A. (2007). Who Cares? Revisiting Empathy in Asperger Syndrome. *J. Autism Dev. Disord.* 37 (4), 709–715. doi:10.1007/s10803-006-0197-8
Rosenthal-von der, P., Astrid, M., Kraeämer, Nicole. C., Hoffmann, Laura., Sobieraj, Sabrina., and Eimler, Sabrina. C. (2013). An Experimental Study on Emotional Reactions towards a Robot. *Int. J. Soc. Robotics* 5 (1), 17–34.
Rosenthal-Von Der, P., Astrid, M., Schulte, Frank. P., Eimler, Sabrina. C., Sobieraj, Sabrina., Hoffmann, Laura., et al. (2014). Investigations on Empathy towards Humans and Robots Using FMRI. *Comput. Hum. Behav.* 33, 201–212.
Salvini, Pericle., Ciaravella, Gaetano., Yu, Wonpil., Ferri, Gabriele., Manzi, Alessandro., Mazzolai, Barbara., et al. (2010). How Safe Are Service Robots in Urban Environments? Bullying a Robot." In 19th International Symposium in Robot and Human Interactive Communication. IEEE, 1–7.
Sansonnet, J.-P., Leray, D., and Martin, J.-C. (2006). "Architecture of a Framework for Generic Assisting Conversational Agents." In International Workshop on Intelligent Virtual Agents, 145, 156–56. Springer. doi:10.1007/11821830_12
Schrepp, Martin., Hinderks, Andreas., and Thomaschewski, Jörg. (2014). *International Conference of Design, User Experience, and Usability*. Springer, 338–392.Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios.
Serino, Andrea., Giovagnoli, Giulia., and Làdavas, Elisabetta. (2009). I Feel what You Feel if You Are Similar to Me. *PLoS One* 4 (3), e4930. doi:10.1371/journal.pone.0004930
Sundar, Harshavardhan. (2020). Locating Multiple Sound Sources from Raw Audio. Amazon Science, April 27, 2020, Available at: https://www.amazon.science/blog/locating-multiple-sound-sources-from-raw-audio.
Tan, Xiang. Zhi., Vázquez, Marynel., Carter, Elizabeth. J., Morales, Cecilia. G., and Steinfeld, Aaron. (2018). "Inducing Bystander Interventions during Robot Abuse with Social Mechanisms." In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 169–177.
Tapus, A., Mataric, M., and Scassellati, B. (2007). Socially Assistive Robotics [Grand Challenges of Robotics]. *IEEE Robot. Automat. Mag.* 14 (1), 35–42. doi:10.1109/mra.2007.339605
Titchener, Edward. Bradford. (1909). *Lectures on the Experimental Psychology of the Thought-Processes.* Macmillan. doi:10.1037/10877-000
Tooby, John., and Cosmides, Leda. (2005). *Conceptual Foundations of Evolutionary Psychology*.
Vincent, James. (2017). *A Drunk Man Was Arrested for Knocking over Silicon Valley's Crime-Fighting Robot*. Verge.
Vischer, Robert. (1873). *Über Die Optische Formgefühl: Ein Beitrag Zur Ästhetik*.
West, Mark., Kraut, Rebecca., and Chew Han, Ei. (2019). *"I'd Blush if I Could: Closing Gender Divides in Digital Skills through Education*.

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

frontiers in Computer Science

ORIGINAL RESEARCH
published: 17 June 2021
doi: 10.3389/fcomp.2021.685250

The Trustworthiness of Voice Assistants in the Context of Healthcare Investigating the Effect of Perceived Expertise on the Trustworthiness of Voice Assistants, Providers, Data Receivers, and Automatic Speech Recognition

Carolin Wienrich, Clemens Reitelbach and Astrid Carolus*

Julius Maximilian University of Würzburg, Würzburg, Germany

OPEN ACCESS

Edited by:
Stefan Hillmann,
Technical University Berlin, Germany

Reviewed by:
Christos Troussas,
University of West Attica, Greece
Rubén San-Segundo,
Polytechnic University of Madrid, Spain

****Correspondence:***
Carolin Wienrich
carolin.wienrich@uni-wuerzburg.de

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

Received: *24 March 2021*
Accepted: *19 May 2021*
Published: *17 June 2021*

Citation:
Wienrich C, Reitelbach C and Carolus A (2021) The Trustworthiness of Voice Assistants in the Context of Healthcare Investigating the Effect of Perceived Expertise on the Trustworthiness of Voice Assistants, Providers, Data Receivers, and Automatic Speech Recognition. Front. Comput. Sci. 3:685250. doi: 10.3389/fcomp.2021.685250

As an emerging market for voice assistants (VA), the healthcare sector imposes increasing requirements on the users' trust in the technological system. To encourage patients to reveal sensitive data requires patients to trust in the technological counterpart. In an experimental laboratory study, participants were presented a VA, which was introduced as either a "specialist" or a "generalist" tool for sexual health. In both conditions, the VA asked the exact same health-related questions. Afterwards, participants assessed the trustworthiness of the tool and further source layers (provider, platform provider, automatic speech recognition in general, data receiver) and reported individual characteristics (disposition to trust and disclose sexual information). Results revealed that perceiving the VA as a specialist resulted in higher trustworthiness of the VA and of the provider, the platform provider and automatic speech recognition in general. Furthermore, the provider's trustworthiness affected the perceived trustworthiness of the VA. Presenting both a theoretical line of reasoning and empirical data, the study points out the importance of the users' perspective on the assistant. In sum, this paper argues for further analyses of trustworthiness in voice-based systems and its effects on the usage behavior as well as the impact on responsible design of future technology.

Keywords: voice assistant, trustworthiness, trust, anamnesis tool, expertise framing (Min5-Max 8)

INTRODUCTION

Voice-based artificial intelligence systems serving as digital assistants have evolved dramatically within the last few years. Today, Amazon Echo or Google Home is the most popular representatives of the fastest-growing consumer technology (Hernandez, 2021; Meticulous Market Research, 2021). On the one hand, voice assistants (VAs) engage human users in direct conversation through a natural language interface leading to promising applications for the healthcare sector, such as diagnosis and therapy. On the other hand, their constituting features to recognize, process, and produce human language results in this technology to resemble human-human interaction. Attributing some kind of humanness to technology arouses (implicit) assumptions about the technological devices and affects the user's perception and operation of the device. The media equation approach postulates that the social rules and dynamics guiding human-human interaction similarly apply to human-computer interaction (Reeves and Nass, 1996). Using voice assistants in official application areas involving sensitive data such as medical diagnoses draws attention to the concept of trust: if patients were to reveal personal, sensitive information to the voice-based systems, they would need to trust them. Consequently, questions of the systems' trustworthiness arise asking for features of voice assistants, which might affect the patients' willingness to trust them in a medical context. Results stemming from studies investigating trust in human-human interactions revealed that ascribed expertise is a crucial cue of trust (Cacioppo and Petty, 1986; Chaiken, 1987; Chaiken and Maheswaran, 1994). Reeves and Nass (1996) transferred the analysis of expertise and trust to human-technology interactions. They showed that designating devices (here: a television program) as "specialized" results in more positive evaluations of the content they presented. Many other studies replicated their approach and framed a technological device or a technological agent as a specialist showing that users ascribed a certain level of expertise and evaluated it (implicitly) as more trustworthy (Koh and Sundar, 2010; Kim, 2014, Kim, 2016; Liew and Tan, 2018).

Voice assistants gain in importance in healthcare contexts offering promising contributions in the area of medical diagnosis, for instance. However, both the analysis and the understanding of the psychological processes characterizing the patient-voice assistant interaction are still in their early stages. Similarly, the effects of the assistant's design on the perception of expertise and the evaluation of trust are still in their infancy. Thus, the present paper addressed the following research question: How does framing a voice assistant as a specialist affect the user's perception of its expertise and its trustworthiness?

To gain first insights into the process of patients' perception of expertise of voice-based systems and their willingness to trust in them, a laboratory study was conducted in which participants interacted with a voice assistant. The assistant was introduced as a diagnostic tool for sexual health, which asked a list of questions about sexual behavior, sexual health, and sexual orientation to determine the diagnosis. In a first step, and in accordance with the approach of Reeves and Nass (1996), we manipulated the level of expertise of the voice assistant, which introduced itself as either a "specialist" or a "generalist". In line with established approaches investigating the trustworthiness of technology (e.g., McKnight et al., 1998; Söllner et al., 2012), we compared the participants' perceived trustworthiness of the "specialist" vs. the "generalist" VA. Additionally, we compared the assessments of further source layers of trustworthiness, namely of the platform provider, the provider of the tool, the data receiver, and of automatic speech recognition in general. Moreover, to account for additional explanatory value of interindividual differences in the trustworthiness ratings, we asked for participants' dispositions and characteristics such as their disposition to trust and their tendency to disclose sexual information about themselves. Finally, we analyzed the different source layers of trustworthiness to predict the trustworthiness of the VA based on the trustworthiness of the other source layers. In sum, the present paper showed for the first time that a short written introduction and a "spoken" introduction presented by the VA itself were sufficient to affect the users' perception and their trust in the system significantly. Hence it addresses a human-centered approach to voice assistants to show that small design decisions determine user's trust in VA in a safety-critical application field.

RELATED WORK

Voice Assistants in Healthcare

While voice-based artificial intelligence systems have increased in popularity over the last years, their spectrum of functions, their field of applications, and their technological sophistication have not been fully revealed but are still in their early stages. Today's most popular systems—Amazon Echo (AI technology: Amazon Alexa) or Google Home (AI technology: Google Assistant)—presage a variety of potential usage scenarios. However, according to usage statistics, in a private environment, voice assistants are predominantly used for relatively trivial activities such as collecting information, listening to music, or sending messages or calls (idealo, 2020). Beyond private usage scenarios, voice assistants are applied in professional environments such as industrial production or technical service (e.g., Baumeister et al., 2019), voice marketing, or internal process optimization (Hörner, 2019). In particular, the healthcare sector has been referred to as an emerging market for voice-based technology. More and more use cases emerge in the context of medicine, diagnosis, and therapy (The Medical Futurist, 2020) with voice assistants offering promising features in the area of anamnesis. Particularly the possibility to assess data remotely gains in importance these days. Recently, chatbots were employed to collect the patients' data, their medical conditions, their symptoms, or a disease process (ePharmaINSIDER, 2018; The Medical Futurist, 2020). While some products provide only information (e.g., OneRemission), others track health data (e.g., Babylon Health) or check symptoms and make a diagnosis (e.g., Infermedica). Until today, only a few solutions have integrated speech recognition or direct connection to VA, such as Alexa *via* skills (e.g., Sensely, Ada Health, GYANT). The German company *ignimed UG* (https://ignimed.de/) takes a similar approach: based on artificial intelligence, the patient's information is collected and transmitted to the attending physician, who can work with the patient. Although these voice assistants are used for similar purposes, which all require user to trust the system, users' perceived trustworthiness of voice assistants in healthcare has not been investigated yet.

The medical context imposes different requirements on the system than private usage scenarios do. Data revealed here are more personal and more sensitive, resulting in increasing requirements regarding the system's security and trustworthiness. Consequently, besides focusing on technological improvements of the system, its security or corresponding algorithms, research needs to focus on the patients' perception of the system and their willingness to interact with them in a health-related context. Exceeding the question of which gestalt design impacts both usability and user experience, the field of human-computer interaction needs to ask for features affecting the patients' perceived trustworthiness of the technological counterpart they interact with. One promising approach is analyzing and transferring findings from human-human interactions to human-computer or human-voice assistant interactions. Following the media equation approach of Nass and colleagues, this study postulates similarities between the human counterpart and the technological counterpart, which results in psychological research to be a fruitful source of knowledge and inspiration for the empirically based design of voice assistants in a medical context.

Interpersonal Trust: The Role of Expertise

Interpersonal interactions are characterized by uncertainty and risks since the behavior of the interaction partner is unpredictable—at least to a certain extent. Trust defines the intention to take the risks of interaction by reducing the perceived uncertainty and facilitating the willingness to interact with each other (Endreß, 2010). In communication contexts, trust refers to the listener's degree of confidence in, and level of acceptance of, the speaker and the message (Ohanian, 1990). Briefly spoken, trust in communication refers to the listener's trust in the speaker (Giffin, 1967). According to different models of trust, characteristics of both the *trustor* (the person who gives the trust) and the *trustee* (the person who receives the trust) determine the level of trust (e.g., Mayer et al., 1995). The dimensions of competence, benevolence, and integrity describe the trustee's main characteristics (see, for example, the meta-analysis of McKnight et al. 2002). The perceived trustworthiness of trustees increases with increasing perceived competence, benevolence, or integrity. In communication contexts, the term *source credibility* closely refers to the trustor's perceived trustworthiness in terms of the trustee. It refers to the speaker's positive characteristics that affect the listener's acceptance of a message. The *source-credibility model* and the *source-attractiveness model* concluded that three factors, namely, expertness, trustworthiness, and attractiveness, underscore the concept of source credibility (Hovland et al., 1953). In this context, expertise is also referred to as authoritativeness, competence, expertness, qualification, or being trained, informed, and educated (Ohanian, 1990). In experiments, the perceived expertise of speakers was manipulated by labeling them as "Dr." vs. "Mr." or as "specialist" vs. "generalist" (e.g., Crisci and Kassinove, 1973). The labels served as cues that can bias the perception of the competence, benevolence, or integrity of trustees or communicators and the perception of trust.

When comprehending the underlying effects of information processing, well-established models of persuasion reveal two routes of processing—the heuristic (peripheral) route and the systematic (central) route (e.g., the *heuristic–systematic model, HSM* by Chaiken (1987); the *elaboration likelihood model, ELM* by Cacioppo and Petty (1986). The heuristic (peripheral) route is based on judgment-relevant cues (e.g., source's expertise) and needs less cognitive ability and capacity than the systematic (central) route, which is based on judgment-relevant information (e.g., message content). Typically, individuals will prefer the heuristic route as the more parsimonious route of processing if they trust the source, particularly if cues activate one of the three trustworthiness dimensions (Koh and Sundar, 2010). For example, individuals will perceive more trustworthiness when a person is labeled as a "specialist" compared to a "generalist" since a specialist sends more cues of expertise and activates the dimension of competence (Chaiken, 1987; Chaiken and Maheswaran, 1994). Remarkably, the effect will endure even if both the specialist and the generalist possess objectively the same level of competence or expertise. Consequently, individuals interacting with a specialist are more likely to engage in heuristic processing and implicitly trust the communicator (Koh and Sundar, 2010).

Regarding the resulting level of trust, the trustor's characteristics were found to moderate the impact of the trustee's characteristics. First, the perceived level of expertise depends on the interindividual differences in the processing of information. The outlined indicators of trust need to be noticed and correctly interpreted to have an effect. Furthermore, the individual's personality and experiences were shown to influence the perception of trustworthiness. Finally, an individual's disposition to trust as the propensity to trust other people has been shown to be a significant predictor (Mayer et al., 1995; McKnight et al., 2002).

To summarize, research in various areas revealed that perceived expertise affects the trustee's trustworthiness (e.g., commercial: Eisend, 2006, health: Gore and Madhavan, 1993; Kareklas et al., 2015), general review see Pornpitakpan (2004). This perception is also affected by the trustor's characteristics (e.g., the disposition to trust). With the digital revolution proceeding, technology has become increasingly interactive, assembling human-human interaction to an increasing extent. Today, an individual does not only interact with other human beings but also with technological devices. These new ways of human-technology interaction require the individual to trust in technological counterparts. Consequently, the question arises whether the outlined mechanisms of trust can be transferred to non-human technological counterparts.

Trust in Technology: The Role of Expertise

The media equation approach postulates that social rules and dynamics, which guide human-human interaction apply to human-computer interaction similarly (Reeves and Nass, 1996; Nass and Moon, 2000). To investigate the media equation assumptions, Nass and Moon (2000) established the CASA paradigm (i.e., computer as social actors) and adopted well-established approaches from research on human-human interaction to the analysis of human-computer interactions. Many experimental studies applying the CASA paradigm demonstrated that individuals tend to transfer social norms to computer agents, for example, gender and ethnic stereotypes and rules of politeness and reciprocity (Nass and Moon, 2000). More specific in the context of trust, experimental studies and imagine-based approaches revealed that trust-related situations activate the same brain regions regardless of whether the counterpart is a human being or a technological agent (Venkatraman et al., 2015; Riedl et al., 2013). Consequently, researchers concluded that there are similar basic effects elicited by human and technological trustees (Bär, 2014).

However, when interacting with a non-human partner, the trustee's entity introduces several interwoven levels of trustworthiness, referred to as *source layers* in the following (Koh and Sundar, 2010). The trustors can trust the technical device or system (e.g., VA) itself. Moreover, they could also refer to the provider, the domain, or the human being "behind" the system, such as the person who receives the information (Hoff and Bashir, 2015). Similar to interpersonal trust, three dimensions determine the perceived trustworthiness of technology: performance (analogous to human competence), clarity (analogous to human benevolence), and transparency (analogous to human integrity) (Backhaus, 2017). As known from interpersonal trust, credibility factors bias the perception of the expertise ascribed to the technology. For example, in the study by Reeves and Nass (1996), participants watched and evaluated a *news* or a *comedy* television program. In the study, they were assigned to one of two conditions: the "specialist television" or the "generalist television". The conditions differed regarding the instruction presented by the experimenter, who referred to the television as either the "news TV" or the "entertainment TV" (specialist condition) or to "usual TV" (generalist condition). Findings indicated that individuals evaluated the content presented by the specialist TV set as more positive than the content of the generalist TV set–even though the content was completely identical. The results have been replicated in the context of specialist/generalist television channels (e.g., Leshner et al., 1998), smartphones (Kim, 2014), or embodied avatars (Liew and Tan, 2018). Additionally, in the context of e-health, the level of expertise was shown to affect the perception of trustworthiness (e.g., Bates et al., 2006). Koh and Sundar (2010) explored the psychological effects of expertise (here: specialization) of web-based media technology in the context of e-commerce. They distinguished multiple indicators or sources of trustworthiness (i.e., computer, website, web agent), they referred to as "source layers of web-based mass communication". In their study, they analyzed the effects on individuals' perceptions of expertise and trust distinguishing between these source layers. In their experiment, participants interacted with media technology (i.e., computer website, web agent), which was either labeled as specialist ("wine computer", "wine shop" or "wine agent") and generalist ("computer", "e-shop"" or "e agent"). Again, only the label but not the content differed between the two experimental conditions. Findings supported the positive effects of the specialization label. Participants reported greater levels of trust in specialist media technology compared to generalist media technology with the "specialized" web agent eliciting the strongest effects (compared to "specialized" website or computer). Consequently, this study focusses on the multiple indicators or the multiple source layers contributing to the trustworthiness of a complex technological system. According to which source layer is manipulated the users' assessment of trustworthiness might differ fundamentally.

To summarize, research so far focused mostly on the credibility of online sources (e.g., websites), neglecting other technological agents like voice-based agents, which currently capture the market in the form of voice bots, voice virtual assistants or smart speaker skills. Furthermore, research so far focused on the engineering progress resulting in increasingly improved performances of the systems but tends to neglect the human user, who will interact with the system. As outlined above, in usage scenarios involving sensitive data, the human users' trust in the technological system is a fundamental requirement and a necessary condition of the user opening up to the system. Voice-based systems in a healthcare context need to be perceived as trustworthy agents to get a patient to disclose personal information. However, scientific studies so far reveal a lack of detailed and psychologically arguing analyses and empirical studies investigating the perceived trustworthiness of voice assistants. The present study aims for first insights into the users' perception of the trustworthiness of voice assistants in the context of healthcare raising the following research questions: 1) Does the introduction of a voice assistant as an expert increase its trustworthiness in the context of healthcare? 2) Do the users' individual dispositions influence the perceived trustworthiness of the assistant? 3) How do the levels of perceived trustworthiness of the multiple source layers (e.g., assistant tool, provider, data receiver) interact with each other?

Outline of the Present Study

To answer the research questions, a laboratory study was conducted. Participants interacted with the *Amazon Echo Dot*, Amazon's voice assistant referred to as "the tool" in the following. The VA was introduced to the participants as an "anamnesis tool for sexual health and disorders", which would ask questions about the participants' sexual behavior, their sexual health, and their sexual orientation. Following the approach of Nass and Reeves (1996), participants were randomly assigned to one out of two groups, which differed by one single aspect: the labeling of the VA. Participants received a written instruction in which the VA was either referred to as a "specialist" or a "generalist". Furthermore, at the beginning of the interaction, the VA introduced itself as either a "specialist" or a "generalist". In line with studies investigating trust in artificial agents (e.g., McKnight et al., 1998; Söllner et al., 2012) and studies including multiple sources layers of trustworthiness (e.g., Koh and Sundar, 2010), we distinguished between different source layers of perceived trustworthiness of our setting: the perceived trustworthiness of the VA tool itself, the provider of the tool (i.e., a German company), the platform provider (i.e., Amazon), automatic speech recognition in general, and the receiver of the data (i.e., the attending physician). Furthermore, participant's individual characteristics, i.e., the disposition to trust and the tendency to sexual self-disclosure, were considered.

METHOD

Participants

The 40 participants (28 females, 12 males; average age = 22.45 years; SD = 3.33) were recruited via personal contact or the university recruitment system offering course credit. All participants were German native speakers. Except for one, all participants were students. 80% of them reported having already interacted with a voice assistant. However, when analyzing the duration of these interactions, the sample's experience was rather limited: 75% reported to have interacted with a VA for less than 10 h and 45% for less than 2 h in total.

Task, Manipulation, Pre-test of Manipulation and Pre-test of Required Trust

During the experiment, participants interacted with a VA, *Amazon Echo Dot* (3rd Generation, black). While the VA asked them questions about their sexual behavior, sexual health, and sexual orientation, participants were instructed to answer these questions as honestly as possible using speech input. Participants were randomly assigned to one of two groups ($n = 20$ per group), which only differed regarding the label of the VA. In an introduction text, the VA was introduced as an anamnesis tool, labeled either a "specialist" (using words such as "specialist," "expert") or as a "generalist" (e.g., "usual," "common"). Additionally, the VA introduced itself in two ways. In the "specialist" condition, it referred to itself as a "special tool for sexual anamnesis" and in the general condition as a "general survey tool."

A pre-study ensured the effect of this manipulation and trust to be a prerequisite of answering the anamneses questions. In an online survey, 30 participants read one of the two introduction texts and described the tool in their own words, afterward. A content analysis of their descriptions showed that participants followed the labeling of the text using compatible keywords to describe the device ("specialist" condition: e.g., special, expert; vs. "generalist" condition: normal, common). However, because only twelve participants used at least one predefined condition-related keyword, the experimental manipulation was strengthened by adding more keywords to the instruction text. The final manipulation text is attached to the additional material. Since the VA's perceived trustworthiness was the main dependent variable, the second part of the pre-test ensured that answering the sexual health-related questions required trust. All anamnesis questions were presented to the participants, who rated how likely they would answer each question. The scale ranged from 100 (very likely) to 0 (no, too private) with lower scores indicating higher levels of required trust to answer the question. Questions were clustered in four categories: puberty, sexual orientation, diseases/hygiene and sexual activity. Results showed that questions regarding puberty (average rating = 75.78) and sexual orientation (67.01) required less trust than diseases/hygiene (56.47) and sexual activity (50.75). To ensure a minimum standard of required trust, one question of the puberty category was removed. Furthermore, four conditional questions were added to the categories of diseases/hygiene and sexual activity, which would ask for more detailed information if previous questions were answered with "yes").

To assess the perceived trustworthiness of the tool, different source layers were considered. First, the trustworthiness of the tool provider, German company, *ignimed UG*, had to be evaluated. Second, since the VA tool was connected to Amazon Echo Dot, the trustworthiness of the platform provider (Amazon) was assessed. Third, the trustworthiness of the potential perceiver to the data (gynecologists/urologist) was rated. Finally, we added automatic speech recognition as a proxy for the underlining technology, which the participant also rated in terms of trustworthiness. Note, the experimental manipulation of expertise referred only to the tool itself. Consequently, the VA tool represents the primary source layer, while others refer to further source layers.

The Sexual Health Anamnesis Tool: Questions the VA Asked

After introducing itself, the VA started the anamnesis conversation, which involves 21 questions (e.g., D*o you have venereal diseases*?—*Which one?*). Four categories of questions were presented: puberty (e.g., *What have been the first signs of your puberty?*), diseases/hygiene (e.g., *Have you ever had one or more sexual diseases?*), sexual orientation (e.g., *What genders do you have sexual intercourse with?*) and sexual activity of the past 4 weeks (e.g., *How often have you had sexual intercourse in the past* 4 *weeks?*). The complete list of final measurements follows below.

Measurements

After finishing the conversation with the VA, participants answered a questionnaire presented *via* LimeSurvey on a 15.6″ laptop with an attached mouse. The measures of the questionnaire are presented below:

Perceived Trustworthiness of Source Layers

To measure the **trustworthiness of the VA** three questions adapted from Corritore et al. (2003) were asked (e.g., *I think the tool is trustworthy*). Additionally, an overall item adapted from Casaló et al. (2007) was presented (*Overall I think that the tool is a save place for sensitive information*). All questions were answered on a 7-point Likert scale, ranging from *not true at all* to *very true*.

Questions concerning institutional trust from the SCOUT Questionnaire (Bär et al., 2011) were transferred to assess the perceived trustworthiness of the tool provider, the platform provider, and automatic speech recognition. Items were answered on a 5-point Likert scale, ranging from *not agree at all* to *agree totally*. Five questions assessed the **tool provider's perceived trustworthiness** (e.g., *I believe in the honesty of the provider*) and the **platform provider** (same five questions). Four questions assessed the **trustworthiness of automatic speech recognition** (*Automatic speech recognition is trustworthy technology.*). Finally, the **data receiver's perceived trustworthiness**, namely, the participant's gynecologist/urologist, the KUSIV3-questionnaire, was used (Beierlein et al., 2012). It includes three questions (e.g., *I am convinced that my gynecologist/urologist has good aims*) on a 5-point Likert scale, ranging from *not agree at all* to *agree totally*.

Individual Characteristics

The disposition of trust was measured with three statements (e.g., *For me, it is easy to trust persons or things*), assessed on a 5-point Likert scale (ranging from *not agree at all* to *agree totally*) taken

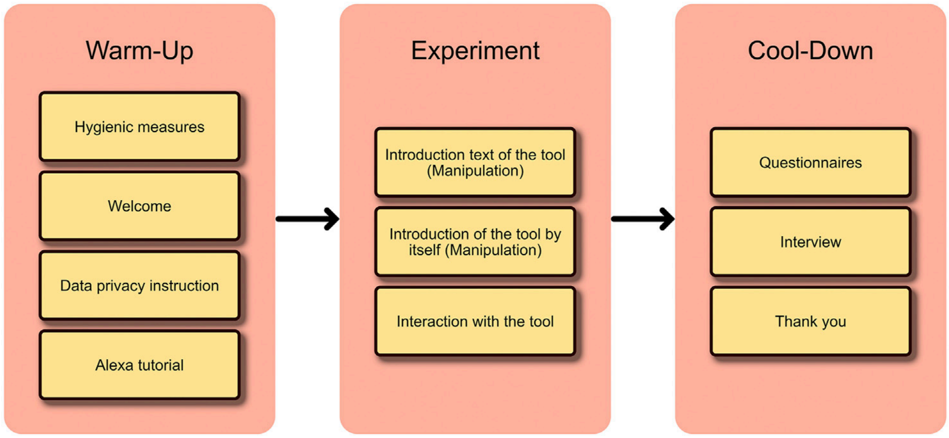


FIGURE 1 | Illustrates the procedure of study.

from the SCOUT Questionnaire (Bär et al., 2011). The tendency to sexual self-disclosure was measured with the sexual self-disclosure scale (Clark, 1987). Participants rated four questions, two on a 5-point Likert scale (e.g., *How often do you talk about sexuality?*, ranging from *never-rarely* to *very often*) and two using a given set of answers (e.g., *With whom do you talk about sexuality? mother | father | siblings | partner | friends(male) | friends(female) | doctors | nobody | other*).

Manipulation Check

To measure how strong the participants believe that the tool is a "specialist" or "generalist," two questions (e.g., *The survey tool has high expertise in the topic*) were answered using a 5-point Likert scale.

Procedure

The study took about 40 min, starting with COVID-19 hygienic routines (warm-up phase: washing and antisepticizing hands, answering a questionnaire and wearing a mouth-nose-mask). Since the experimental supervisor left the room for the actual experiment, participants could discard their face masks during the interaction with the VA. In the warm-up phase, participants were instructed to do a short tutorial with the VA, which asked some trivial questions (e.g., *How is the weather?* or *Do you like chocolate?*). When the participants confirmed to be ready to start the experiment, they were instructed to read the introduction text about the anamnesis tool and to start the interaction with the VA (experimental phase). After finishing, the participants answered the questionnaires and were briefly interviewed about the experience with the VA by the supervisor (cool-down phase). **Figure 1** illustrates the procedure of the study and the three experimental phases.

Design, Hypothesis

Accordingly to previous studies (e.g., Gore and Madhavan, 1993; Reeves and Nass, 1996; Kareklas et al., 2015), the experiment followed a between-subjects design with two conditions ("specialist" or "generalist" VA). In line with the first research question, referring to the effects of perceived expertise on perceived trustworthiness (differentiated regarding source layers), the first hypotheses postulated that the perceived trustworthiness (across the source layers) would be higher in the specialist condition than in the generalist condition. The second research question addressed the impact of individual dispositions on perceived trustworthiness. In line with Mayer et al. (1995) and McKnight et al. (2002), the second hypotheses assumed that higher individual trust-related dispositions result in higher trustworthiness ratings. Finally, the third research question explorative asked whether the perceived trustworthiness of the multiple source layers (e.g., the assistant tool, the providers, the receiver) interact with each other (Koh and Sundar (2010). **Table 1** gives an overview of the hypothesis.

Data Analyses

Data have been prepared as proposed by the corresponding references. To facilitate the comparability of measures of trustworthiness, items answered on a 7-point Likert scale were converted to a 5-point scale. Five t-tests for independent groups (specialist condition vs. generalist condition) were conducted to test the first hypotheses. The second group of hypotheses was tested, conducting five linear regression analyses with the source layers' trustworthiness as the five criteria variables and the individual characteristics as the predictor variables. Finally, a linear regression analysis was conducted regarding the explorative research question with trustworthiness of the VA as the criteria and the other layers of trustworthiness as the predictors. The following section report means (M), standard deviations (SD) of scales as well as the test statistic parameters such as the *t*-value (t), and *p*-value (*p*).

RESULTS

Impact of Expert Condition on Source Indicators' Perceived Trustworthiness

As expected, the perceived trustworthiness of the tool was higher in the specialist condition (M = 3.206, SD = 1.037) than in the generalist condition (M = 2.634, SD = 0.756). However, the result was just not significant ($t_{(38)}$ = 2.019, p = 0.051, d = 0.638). Contrary to our expectations, the perceived trustworthiness of the platform provider, the tool provider, the data receiver, and the automatic speech recognition did not differ significantly between

TABLE 1 | Hypotheses and a short overview of corresponding results.

Class of hypotheses	Dependent variables	Results
H1: Source indicators' perceived trustworthiness is higher in the specialist condition than in the generalist condition.	H1a: tool's trustworthiness	Confirmed with adjusted analysis.
	H1b: Tool provider's trustworthiness	Confirmed with adjusted analysis.
	H1c: Platform provider's trustworthiness	Confirmed with adjusted analysis.
	H1d: Data receiver's trustworthiness	Confirmed with adjusted analysis.
	H1e: Trustworthiness of automatic speech recognition	Confirmed with adjusted analysis.
H2: Higher individual trust-related dispositions and tendencies result in higher trustworthiness ratings.	H2a: disposition to trust	Confirmed for the trustworthiness of automatic speech recognition. Confirmed for the tool's trustworthiness.
	H2b: Tendency to sexual self-disclosure	Confirmed by trend for the trustworthiness of automatic speech recognition.
RQ3: Does the perceived trustworthiness for multiple source indicators interact with each other?	–	As higher the tool provider's scores, as higher is the perceived trustworthiness of the tool itself. By trend: As higher the speech recognition scores, as higher is the tool's perceived trustworthiness. By trend: As higher the platform provider's (i.e., Amazon) scores, the tool's perceived trustworthiness is lower.

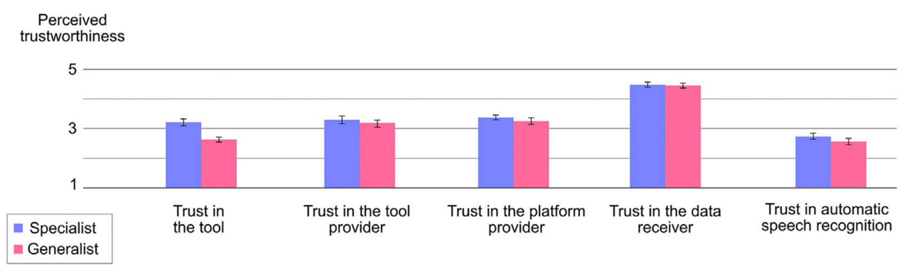


FIGURE 2 | Shows the unadjusted analyses of the different trustworthiness indicators distinguishing between the specialist condition (blue) and the generalist condition (red). The scale ranged from 1 to 5, with higher values indicating higher ratings of trustworthiness.

TABLE 2 | Results of unadjusted analyses.

Indicator of trustworthiness	Specialist		Generalist		Significance test
	M	SD	M	SD	
Tool	3.206	1.037	2.634	0.756	$t(38) = 2.019$, $p = 0.051$, $d = 0.638$
Tool provider (German company)	3.290	1.145	3.190	1.053	$t(38) = 0.287$, $p = 0.775$, $d = 0.091$
Platform provider (amazon)	3.250	0.969	3.370	0.700	$t(38) = -0.449$, $p = 0.656$, $d = 0.142$
Data receiver (gynecologists/urologist)	4.483	0.791	4.450	0.767	$t(38) = 0.135$, $p = 0.893$, $d = 0.042$
Automatic speech recognition	2.725	0.862	2.550	0.955	$t(38) = 0.608$, $p = 0.547$, $d = 0.192$

the conditions (see **Figure 2** and **Table 2** for descriptive results and t-test results of unadjusted analyses).

Impact of Manipulation Check Success on Source Indicators' Perceived Trustworthiness

Although a pre-test confirmed the manipulation of the two conditions, the manipulation check of the main study revealed a lack of effectivity: two control questions showed that the specialist tool was not rated as significantly "more special" than the generalist tool (see **Table 3** for descriptive and *t*-test results). Consequently, the assignment to the two groups did not result in significantly different levels of perceived expertise of the VA.

Consequently, we needed to adjust the statistical analyses of the group comparisons. We re-analyzed the ratings of the control questions (asking for the VA's expertise). Based on the actually perceived expertise, participants were divided into two groups with ratings below and above the averaged scales' median (MD = 3.333). Independently of the intended manipulation, 22

TABLE 3 | Results of manipulation check.

Control question: "The survey tool . . .	Specialist M (SD)	Generalist M (SD)	Significance test
. . . has high expertise in the topic."	3.300 (0.923)	3.050 (0.887)	t(df) = 0.873, p = 0.388, d = 0.276
. . . was developed for a special purpose."	3.650 (1.268)	3.350 (1.040)	t(df) = 0.818, p = 0.419, d = 0.259

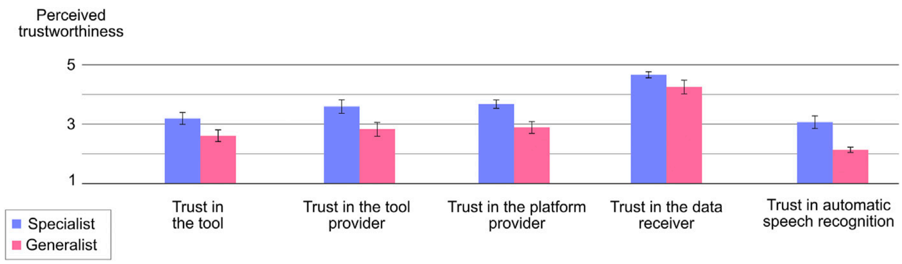


FIGURE 3 | Shows the adjusted analyses of the different trustworthiness indicators distinguishing between the specialist condition (blue) and the generalist condition (red). The scale ranged from 1 to 5, with higher values indicating higher ratings of trustworthiness.

participants revealed higher ratings of the VA's expertise (> 3.5; group 1) indicating that they rather perceived a specialist tool. In contrast, 18 participants revealed lower ratings of the expertise (< 3.0; group 2) indicating that they rather perceived a generalist tool. In sum, the re-analyses (referred to as the "adjusted analysis" below) will analyze the perceived trustworthiness of participants, who actually perceived the VA as a specialist or a generalist independently of the intended manipulation (see **Figure 3**).

In line with the hypotheses, the **perceived trustworthiness of the tool** would be significantly higher, if the tool was actually perceived as a specialist (M = 3.182, SD = 0.938) compared to the perception of a generalist (M = 2.599, SD = 0.836; $t_{(38)}$ = 2.051, p = 0.045, d = 0.652). Also in line with expectations, the **platform provider's perceived trustworthiness** would be significantly higher, if the VA was perceived as a specialist (M = 3.581, SD = 1.313) than a generalist (M = 2.822, SD = 1.170, $t_{(38)}$ = 2.318, p = 0.025, d = 0.737). The same applies to the **provider's perceived trustworthiness**: the provider of the specialist tool was rated to be significantly more trustworthy (M = 3.663, SD = 0.066) than the provider of a generalist tool (M = 2.878, SD = 0.984; $t_{(38)}$ = 3.331 p = 0.02, d = 1.503). Likewise, the **perceived trustworthiness of automatic speech recognition** was significantly higher for a perceived specialist (M = 3.057, SD = 0.994) than a generalist (M = 2.125, SD = 0.376; $t_{(38)}$ = 3.757, p = 0.01, d = 1.194). However, the data receiver's perceived trustworthiness did not differ significantly between a specialist (M = 4.652, SD = 0.488) and a generalist (M = 4.241, SD = 0.982; $t_{(38)}$ = 1.722, p = 0.093, d = 0.547).

Additional Predictors of the Perceived Trustworthiness

Individual Characteristics

Linear regression analyses were conducted with the five trustworthiness source layers as the criteria variables and the general disposition to trust and to disclose sexual health information as predictor variables. Only two regressions involved significant predictions: the perceived trustworthiness of the VA and of the speech recognition in general. First, the prediction of the tool's perceived trustworthiness was significant (β = 0.655, $t_{(37)}$ = 2.056, p = 0.047) with the tendency to disclose health information contributing significantly: the higher this tendency, the higher the perceived trustworthiness of the tool. Second, the prediction of the automatic speech recognition's trustworthiness was significant, with the participants' disposition to trust contributing significantly to the prediction (β = 0.306, $t_{(37)}$ = 2.276, p = 0.029). Moreover, the tendency to disclose sexual health information contributed substantially but only by trend (β = 0.370, $t_{(37)}$ = 1.727, p = 0.093).

Further Source Layers of Trustworthiness

The final regression analysis investigated whether the further source layers of trustworthiness indicators (providers, speech recognition, receiver) predicted the tool's trustworthiness. Results revealed that the **trustworthiness of the provider** of the tool (i.e., ignimed UG) significantly predicted the trustworthiness of the tool itself (β = 0.632, $t_{(35)}$ = 3.573, p = 0.001): the higher the provider's trustworthiness, the higher the trustworthiness of the tool. Similar, but not significantly, the trustworthiness of the automatic speech recognition predicted the trustworthiness of the tool (β = 0.371, $t_{(35)}$ = 1.768, p = 0.086): the higher the trustworthiness of the speech recognition, the higher the perceived trustworthiness of the tool. In contrast, the higher the platform provider's (i.e., Amazon) scores, the lower the trustworthiness of the tool, only by trend, however (β = −0.446, $t_{(35)}$ = −1.772, p = 0.085).

DISCUSSION

Aim of the Present Study

Voice-based (artificial intelligence) systems serving as digital assistants have evolved dramatically within the last few years.

The healthcare sector has been referred to as an emerging market for these systems, which imposes different requirements on the systems than private usage scenarios do. Data revealed here are more personal and more sensitive, resulting in increasing engineering requirements regarding data security, for example. To establish voice-based systems in a more sensitive context, the users' perspective needs to be considered. In a healthcare context, users need to trust their technological counterpart to disclose personal information. However, the trustworthiness of most of the systems in the market and the users' willingness to trust in the applications has not been analyzed yet. The present study bridged this research gap. In an empirical study, the trustworthiness of a voice-based anamnesis tool was analyzed. In two different conditions, participants either interacted with a VA, which was introduced as a "specialist" or a "generalist". Then, they rated the trustworthiness of the tool, distinguishing between different source layers of trust (provider, platform provider, automatic speech recognition in general, data receiver). To ensure external reliability, participants interacted with an anamnesis tool for sexual health, which collected health data by asking questions regarding their puberty, sexual orientation, diseases/hygiene and sexual activity. They were informed that the tool uses artificial intelligence to provide a diagnosis, which would be sent to the gynecologists/urologist.

Answering the Research Questions

The present study investigated three research questions: 1) Does the expert framing of a voice assistant increase its trustworthiness in the context of Further, 2) Does individual dispositions influence the perceived trustworthiness? Finally 3) Do different trustworthiness source indicators (e.g., the assistant tool, the providers, the receiver), interact with each other?

In line with previous studies, the present results revealed that participants, who perceive the VA tool as a specialist tool, reported higher levels of trustworthiness across all different source layers—compared to participants, who perceived the tool as a generalist. Considering, that the tool acted completely identically in both conditions and that the conditions only differed in terms of the introduction of the VA to the participant (written introduction and introduction presented by the VA itself), the present study highlights the manipulability of the users' perception of the system and the effects this perception has on the evaluation of the trustworthiness of the system. The way a diagnostic tool is introduced to the patient seems to be of considerable importance when it comes to the patient's perception of the tool and the willingness to interact with it. As the present study reveals, a few words can fundamentally change the patients' opinions of the tool, which might affect their willingness to cooperate. In the presence of the ongoing worldwide pandemic, we all learnt about the need for intelligent tools, which support physicians with remote anamnesis and diagnosis to unburden the stationary medical offices. Our study shows how important it is to not only consider engineering aspects and ensure that the system functions properly but to consider also the users' perception of the tool and the resulting trustworthiness. Thus, our results offer promising first insights for developers and designers. However, our results also refer to risks. Many health-related tools conquer the market without any quality checks. If these tools framed themselves as experts or specialists, users could be easily misled.

Following the basic assumption of the media equation approach and its significant research body confirming the idea that social rules and dynamics, which guide human-human interaction, similarly apply to human-computer interaction (Reeves and Nass, 1996), voice-based systems could be regarded as a new era of technological counterpart. Being able to recognize process and produce human language, VA adopt features that have been exclusively human until recently. Consequently, VA can verbally introduce themselves to the users resulting in a powerful manipulation of the users' perception. The presented results show how easily and effectively the impression of a VA can be manipulated. Furthermore, our results indicate an area of research, today's HCI research tends to miss too often. While its primary focus is on the effect of gestalt design on usability and user experience, our results encourage to refer to the users' perspective on the system and the perceived trustworthiness as an essential aspect of a responsible and serious design, which bears chances and risks for both high-quality and low-quality applications.

Referring to methodological challenges, the present study reveals limitations of the way we manipulated the impression of the VA. Participants read an introduction text, which referred to the VA as either a specialist or a generalist tool. Moreover, the VA introduced itself as a specialist or a generalist. Although a pre-test was conducted to ensure the manipulation, not all participants took the hints resulting in participants of the "specialist condition", who did not refer to the VA as a specialist. Similarly, not all participants took the hints resulting in participants of the "generalist condition", who did not refer to the VA as a generalist. Future studies in this area need to conduct a manipulation check to ensure their manipulation or to adopt their analysis strategy (e.g., post-hoc assignments of groups). In our study, unadjusted analyses, which strictly followed the intended manipulation, resulted in reduced effects compared to the newly composed groups following the participants' actual perception of the tool. Additionally, future research should focus on manipulations that are more effective. Following the *source-credibility* model and the *source-attractiveness model,* the perspective on perceived competence is more complex. Besides the perception of expertise, authoritativeness, competence, qualification, or a system perceived as being trained, informed, and educated could contribute to the attribution of competence (Ohanian, 1990; McKnight et al., 2002). Future studies should use the variety of possibilities to manipulate the perceived competence of a VA.

From a theoretical perspective, competence is only one dimension describing the human trustee's main characteristics. Benevolence and integrity of the trustee are also relevant indicators (e.g., Ohanian, 1990; McKnight et al., 2002). In terms of an artificial trustees, performance (analogous to human competence), clarity (analogous to human benevolence), and transparency (analogous to integrity) are further dimension, which determine the impression (Backhaus, 2017). Future studies should widen the perspective and refer to the multiple dimensions. Additionally, human information processing was introduced to follow two different routes: the systematic (central) or the heuristic (peripheral) route (Cacioppo and Petty, 1986) with personal relevant topics increasing the probability to be processed systematically. Sexual health, the topic of the anamnesis tool of the present study, was shown to be of personal relevance (Kraft Foods, 2009), indicating systematic processing of judgment-relevant cues (e.g., source's expertise) (Cacioppo and Petty, 1986). This might explain the limited effect of the manipulation on the ratings of trustworthiness. Possibly, the personally relevant topic of sexual health triggers the central route of processing resulting in the rather quick labeling of the VA as a heuristic (peripheral) cue to have a limited effect. Moreover, the interaction with the tool might have further diminished the effect of the manipulation. That might also explain why only participants, who explicitly evaluated the tool as a specialist, showed more trustworthiness. Thus, it should be further investigated whether effects of the heuristical design of health-related websites, for example (Gore and Madhavan, 1993), can be transferred to voice-based anamnesis tools assessing highly personal relevant topics.

Regarding the second research question, the present results show only minor effects of the participants' individual characteristics. From the multiple source layers of trustworthiness, the participants' general disposition to trust only impacted the perceived trustworthiness of automatic speech recognition. Possibly, our sample was too homogenous regarding the participants' disposition to trust: mean values of trust-disposition were relatively high (M = 3.567 on a 5-point scale) and rather low (SD = 1.03). Future studies could consider to incorporate predictors, which are more closely connected with the selected use case such as the tendency to disclose sexual health (Mayer et al., 1995).

The third research question explored the relationship between the different source layers of trustworthiness. When predicting the trustworthiness of the tool, only the tool provider's perceived trustworthiness was a significant predictor. The trustworthiness of the platform provider and automatic speech recognition are related by trend while the data receiver's (gynecologists/urologist) trustworthiness was of minor importance. However, as our participants knew that their data would be only saved on the university's server, the latter results might have been different if the data were transferred to the attending physician (or if participants assumed data transfer). Nevertheless, results are interpreted as a careful first confirmation of Koh and Sundar (2010), who postulate that the perception of an artificial counterpart is not only influenced by the characteristics of the tool itself but also by indicators related to the tool. Referring to today's most popular VAs for private use, Amazon Echo and Google Home, both tools might be closely associated with the perceived image or trustworthiness of the companies. If such consumer products are used in the context of healthcare, reservations regarding the companies might have an impact. Furthermore, the general view of automatic speech recognition affected the perceived trustworthiness of the tool. Thus, current public debates about digitalization or artificial intelligence should also be considered when designing and using VA for health-related applications.

Limitations and Future Work

To summarize the limitations and suggestions for future work presented above, the manipulation of competence and the additional indicators of trustworthiness need to be reconsidered. Future work might consider more fine-grained and more in-depth operationalization of different expert levels (e.g., referring to the performance of the tool), include further manipulations of the competence dimensions (e.g., referring to a trained system), or incorporate the dimensions of clarity (analogous to human benevolence) or transparency (resample integrity). The perceived trustworthiness might result from a systematic (central) information process due to potential high personal relevance. Future work should investigate whether the effects of the heuristic design of health-related websites, for example (Gore and Madhavan, 1993; Kareklas et al., 2015), can be transferred to voice-based anamnesis tools assessing highly personal relevant topics. The data receiver (gynecologists/urologist) played only a subordinate role in the present study. As participants knew that their disclosed data would be stored on university servers, they were of minor importance. Future work should increase the external validity of the experiment by incorporating the data receiver more explicitly. Finally, only the tool's expert status and not of the additional source layers of trustworthiness have been manipulated, resulting in relatively simple analyses of interaction effects between the trustworthiness indicators. Future studies might choose a more elaborate design. Finally, perceived trustworthiness is an essential topic for different application areas such as education (Troussas et al., 2021). Another important field might be the perceived trustworthiness of multilingual voice assistants applicated in multilingual societies (Mukherjee et al., 2021) A different approach to the perceived trustworthiness would be testing the impact of different dialogue architectures (e.g., Fernández et al., 2005). Strategies of dialogue design can be very different and impact on user's trustworthiness. Future studies should investigate if hardcoded intents or flexible and natural spoken interactions have a different impact.

Conclusion and Contribution

Voice assistants gain in importance in healthcare contexts. With remote anamnesis and diagnoses gaining in importance these days, voice-based systems offer promising contributions, for instance, in the area of medical diagnoses. Using voice assistants in data sensitive contexts draws attention to the concept of trust: if patients were to reveal personal, sensitive information to the voice-based systems, they would need to trust them. However, the analysis and the understanding of the psychological processes characterizing the patient-voice assistant interaction is still in their early stages. For human-human relationships, psychological research revealed the characteristics of individuals, who give trust (trustor) and those who receive trust (trustee). Moreover, research established models of the characteristics, which are processed and attributed (e.g., *source-credibility model*, *source-attractiveness model*, *HSM*, *ELM*). Researchers in the field of human-computer interaction transferred this knowledge to interactions with technological counterparts (e.g., television, web pages, web agents). However, little is known about voice-based tools, which have become increasingly popular, and which involve more complex, more humanlike features (speech processing) compared to technology so far. The present study contributes to close this research gap by presenting ideas for the design of VAs, which have been derived from literature. Furthermore, the study provides empirical data of human users interacting with a device to disclose health-related information. Results showed that participants, who perceived the VA tool as a specialist tool, reported higher trustworthiness scores than participants, who thought to interact with a generalist tool. To conclude, the users' perception significantly influences the trust users have in the VA. Furthermore, influencing this perception was shown to be rather easy: a short-written introduction and a "spoken" introduction presented by the VA itself were sufficient to affect the users' perception and their trust in the system significantly. In sum, we want to draw attention to the importance of the human user's perspective when interacting with technology. Future studies need to address the trustworthiness of technology to contribute to more responsible and serious design processes to take the chances technology offers and to avoid the risks of low-quality applications.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CW: concept of study, theoretical background, argumentation line, data analyses, supervision; CR: skill implementation, study conductance, data analyses; AC: theoretical background, argumentation line, supervision.

FUNDING

This publication was supported by the Open Access Publication Fund of the University of Wuerzburg.

REFERENCES

Backhaus, N. (2017). User's Trust and User Experience in Technical Systems: Studies on Websites and Cloud Computing 1–281. doi:10.14279/depositonce-5706

Bär, N. (2014). Human-Computer Interaction and Online Users' Trust. PhD dissertation. TU Chemnitz. Retrieved from: http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa149685

Bär, N., Hoffmann, A., and Krems, J. (2011). "Entwicklung von Testmaterial zur experimentellen Untersuchung des Einflusses von Usability auf Online-Trust," in *Reflexionen und Visionen der Mensch-Maschine-Interaktion – Aus der Vergangenheit lernen, Zukunft gestalten*. Editors S. Schmid, M. Elepfandt, J. Adenauer, and A. Lichtenstein, 627–631.

Bates, B. R., Romina, S., Ahmed, R., and Hopson, D. (2006). The Effect of Source Credibility on Consumers' Perceptions of the Quality of Health Information on the Internet. *Med. Inform. Internet Med.* 31, 45–52. doi:10.1080/14639230600552601

Baumeister, J., Sehne, V., and Wienrich, C. (2019). "A Systematic View on Speech Assistants for Service Technicians," in LWDA Jäschke Robert and Weidlich Matthias Editors. Berlin, Germany, 195–206. doi:10.1136/bmjspcare-2019-huknc.228

Beierlein, C., Kemper, C. J., Kovaleva, A., and Rammstedt, B. (2012). "Kurzskala zur Messung des zwischenmenschlichen Vertrauens: Die Kurzskala Interpersonales Vertrauen (KUSIV3)[Short scale for assessing interpersonal trust: The short scale interpersonal trust (KUSIV3)],"in GESIS Working Papers 2012|22, Kölm.

Cacioppo, J. T., and Petty, R. E. (1986). in "He Elaboration Likelihood Model of Persuasion," in *Advances in Experimental Social Psychology*. Editor L. Berkowitz (New York: Academic Press), 123–205.

Casaló, L. V., Flavián, C., and Guinalíu, M. (2007). The Role of Security, Privacy, Usability and Reputation in the Development of Online Banking. *Online Inf. Rev.* 31 (5), 583–603. doi:10.1108/14684520710832315

Chaiken, S., and Maheswaran, D. (1994). Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgment. *J. Personal. Soc. Psychol.* 66, 460–473. doi:10.1037/0022-3514.66.3.460

Chaiken, S. (1987). "The Heuristic Model of Persuasion," in Social influence: the ontario symposium M. P. Zanna, J. M. Olson, and C. P. Herman Editors. . Lawrence Erlbaum Associates, Inc, 5, 3–39.

Clark, L. V. (1987). *Sexual Self-Disclosure to Parents and Friends*. Texas: University Press.

Corritore, C. L., Kracher, B., and Wiedenbeck, S. (2003). On-line Trust: Concepts, Evolving Themes, a Model. *Int. J. Human-Computer Stud.* 58, 737–758. doi:10.1016/s1071-5819(03)00041-7

Crisci, R., and Kassinove, H. (1973). Effect of Perceived Expertise, Strength of Advice, and Environmental Setting on Parental Compliance. *J. Soc. Psychol.* 89, 245–250. doi:10.1080/00224545.1973.9922597

Eisend, M. (2006). Source Credibility Dimensions in Marketing Communication–A Generalized Solution. *J. Empir. Gen. Mark. Sci.* 10.

Endreß, M. (2010). "Vertrauen–soziologische Perspektiven," in *Vertrauen–zwischen Sozialem Kitt Senkung Von Transaktionskosten*, 91–113.

ePharmaINSIDER (2018). Das sind die Top 12 Gesundheits-Chatbots. (Accessed March 17, 2021). Available at: https://www.epharmainsider.com/die-top-12-gesundheits-chatbots/.

Fernández, F., Ferreiros, J., Sama, V., Montero, J. M., Segundo, R. S., Macías-Guarasa, J., et al. (2005). "Speech Interface for Controlling an Hi-Fi Audio System Based on a Bayesian Belief Networks Approach for Dialog Modeling," in Ninth European Conference on Speech Communication and Technology.

Giffin, K. (1967). The Contribution of Studies of Source Credibility to a Theory of Interpersonal Trust in the Communication Process. *Psychol. Bull.* 68, 104–120. doi:10.1037/h0024833

Gore, P., and Madhavan, S. S. (1993). Credibility of the Sources of Information for Non-prescription Medicines. *J. Soc. Adm. Pharm.* 10, 109–122.

Hernandez, A. (2021). The Best 7 Free and Open Source Speech Recognition Software Solutions. Goodfirms. Available at: https://www.goodfirms.co/blog/best-free-open-source-speech-recognition-software (Accessed March 17, 2021).

Hoff, K. A., and Bashir, M. (2015). Trust in Automation. *Hum. Factors* 57, 407–434. doi:10.1177/0018720814547570

Hörner, T. (2019). "Sprachassistenten im Marketing," in *Marketing mit Sprachassistenten*. Springer, Gabler: Wiesbaden, 49–113. doi:10.1007/978-3-658-25650-0_3

Hovland, C. I., Janis, I. L., and Kelley, H. H. (1953). *Communication and Persuasion*. New Haven, CT: Yale University Press

idealo (2020). E-Commerce-Trends 2020: Millennials Treiben Innovationen Voran. Idealo. Available at: https://www.idealo.de/unternehmen/pressemitteilungen/ecommerce-trends-2020/(Accessed March 17, 2021).

Kareklas, I., Muehling, D. D., and Weber, T. J. (2015). Reexamining Health Messages in the Digital Age: A Fresh Look at Source Credibility Effects. *J. Advertising* 44, 88–104. doi:10.1080/00913367.2015.1018461

Kim, K. J. (2014). Can Smartphones Be Specialists? Effects of Specialization in mobile Advertising. *Telematics Inform.* 31, 640–647. doi:10.1016/j.tele.2013.12.003

Kim, K. J. (2016). Interacting Socially with the Internet of Things (IoT): Effects of Source Attribution and Specialization in Human-IoT Interaction. *J. Comput-mediat Comm.* 21, 420–435. doi:10.1111/jcc4.12177

Koh, Y. J., and Sundar, S. S. (2010). Heuristic versus Systematic Processing of Specialist versus Generalist Sources in Online media. *Hum. Commun. Res.* 36, 103–124. doi:10.1111/j.1468-2958.2010.01370.x

Kraft Foods (2009). Tabuthemen. Statista. Available at: https://de.statista.com/statistik/daten/studie/4464/umfrage/themen-ueber-die-kaum-gesprochen-wird/(Accessed March 17, 2021).doi:10.1007/978-3-8349-8235-3

Leshner, G., Reeves, B., and Nass, C. (1998). Switching Channels: The Effects of Television Channels on the Mental Representations of Television News. *J. Broadcasting Electron. Media* 42, 21–33. doi:10.1080/08838159809364432

Liew, T. W., and Tan, S.-M. (2018). Exploring the Effects of Specialist versus Generalist Embodied Virtual Agents in a Multi-Product Category Online Store. *Telematics Inform.* 35, 122–135. doi:10.1016/j.tele.2017.10.005

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Amr* 20, 709–734. doi:10.5465/amr.1995.9508080335

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). Developing and Validating Trust Measures for E-Commerce: An Integrative Typology. *Inf. Syst. Res.* 13, 334–359. doi:10.1287/isre.13.3.334.81

McKnight, D. H., Cummings, L. L., and Chervany, N. L. (1998). Initial Trust Formation in New Organizational Relationships. *Amr* 23, 473–490. doi:10.5465/amr.1998.926622

Meticulous Market Research (2021). Healthcare Virtual Assistant Market By Product (Chatbot And Smart Speaker), Technology (Speech Recognition, Text-To-Speech, And Text Based), End User (Providers, Payers, And Other End User), And Geography - Global Forecast To 2025. Available at: http://www.meticulousresearch.com/(Accessed March 17, 2021).

Mukherjee, S., Nediyanchath, A., Singh, A., Prasan, V., Gogoi, D. V., and Parmar, S. P. S. (2021). "Intent Classification from Code Mixed Input for Virtual Assistants," in 2021 IEEE 15th International Conference on Semantic Computing (ICSC) (IEEE), 108–111.

Nass, C., and Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *J. Soc. Isssues* 56, 81–103. doi:10.1111/0022-4537.00153

Ohanian, R. (1990). Construction and Validation of a Scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness. *J. Advertising* 19, 39–52. doi:10.1080/00913367.1990.10673191

Pornpitakpan, C. (2004). The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *J. Appl. Soc. Pyschol* 34, 243–281. doi:10.1111/j.1559-1816.2004.tb02547.x

Reeves, B., and Nass, C. (1996). *The media Equation: How People Treat Computers, Television, and New media like Real People and Places*. Cambridge, UK: Cambridge University Press.

Riedl, B., Gallenkamp, J., and Hawaii, A. P. (2013). "The Moderating Role of Virtuality on Trust in Leaders and the Consequences on Performance," in 2013 46th Hawaii International Conference on System Sciences . IEEE, 373–385. Available at: https://ieeexplore.ieee.org/Xplore/home.jsp. doi:10.1109/hicss.2013.644

Söllner, M., Hoffmann, A., Hoffmann, H., and Leimeister, J. M. (2012). Vertrauensunterstützung für sozio-technische ubiquitäre Systeme. *Z. Betriebswirtsch* 82, 109–140. doi:10.1007/s11573-012-0584-x

The Medical Futurist (2020). The Top 12 Healthcare Chatbots. *Med. Futur*. Available at: https://medicalfuturist.com/top-12-health-chatbots (Accessed March 17, 2021).

Troussas, C., Krouska, A., Alepis, E., and Virvou, M. (2021). Intelligent and Adaptive Tutoring through a Social Network for Higher Education. *New Rev. Hypermedia Multimedia*, 1–30. doi:10.1080/13614568.2021.1908436

Venkatraman, V., Dimoka, A., Pavlou, P. A., Vo, K., Hampton, W., Bollinger, B., et al. (2015). Predicting Advertising success beyond Traditional Measures: New Insights from Neurophysiological Methods and Market Response Modeling. *J. Marketing Res.* 52, 436–452. doi:10.1509/jmr.13.0593

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002b). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research* 13, 334–359. doi:10.1287/isre.13.3.334.81

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

METHODS
published: 21 June 2021
doi: 10.3389/fcomp.2021.686050

Check for updates

Crowdsourcing Ecologically-Valid Dialogue Data for German

Yannick Frommherz[1]* and Alessandra Zarcone[2]

[1]Audio and Media Technologies, Semantic Audio Processing, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, [2]Audio and Media Technologies, HumAIn, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

Despite their increasing success, user interactions with smart speech assistants (SAs) are still very limited compared to human-human dialogue. One way to make SA interactions more natural is to train the underlying natural language processing modules on data which reflects how humans would talk to a SA if it was capable of understanding and producing natural dialogue given a specific task. Such data can be collected applying a Wizard-of-Oz approach (WOz), where user and system side are played by humans. WOz allows researchers to simulate human-machine interaction while benefitting from the fact that all participants are human and thus dialogue-competent. More recent approaches have leveraged simple templates specifying a dialogue scenario for crowdsourcing large-scale datasets. Template-based collection efforts, however, come at the cost of data diversity and naturalness. We present a method to crowdsource dialogue data for the SA domain in the WOz framework, which aims at limiting researcher-induced bias in the data while still allowing for a low-resource, scalable data collection. Our method can also be applied to languages other than English (in our case German), for which fewer crowd-workers may be available. We collected data asynchronously, relying only on existing functionalities of Amazon Mechanical Turk, by formulating the task as a dialogue continuation task. Coherence in dialogues is ensured, as crowd-workers always read the dialogue history, and as a unifying scenario is provided for each dialogue. In order to limit bias in the data, rather than using template-based scenarios, we handcrafted situated scenarios which aimed at not pre-*script*-ing the task into every single detail and not priming the participants' lexical choices. Our scenarios cued people's knowledge of common situations and entities relevant for our task, without directly mentioning them, but relying on vague language and circumlocutions. We compare our data (which we publish as the CROWDSS corpus; n = 113 dialogues) with data from MultiWOZ, showing that our scenario approach led to considerably less scripting and priming and thus more ecologically-valid dialogue data. This suggests that small investments in the collection setup can go a long way in improving data quality, even in a low-resource setup.

Keywords: dialogue data, voice assistants, crowdsourcing, Wizard-of-Oz, German, ecological validity, situated knowledge

OPEN ACCESS

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

Reviewed by:
Marius Hamacher, University of Duisburg-Essen, Germany
Christos Troussas, University of West Attica, Greece
Sebastian Kahl, Bielefeld University, Germany

****Correspondence:***
Yannick Frommherz
yannick.frommherz@iis.fraunhofer.de

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

Received: *26 March 2021*
Accepted: *26 May 2021*
Published: *21 June 2021*

Citation:
Frommherz Y and Zarcone A (2021) Crowdsourcing Ecologically-Valid Dialogue Data for German. Front. Comput. Sci. 3:686050. doi: 10.3389/fcomp.2021.686050

INTRODUCTION

Recently, smart speech assistants (SAs) have found their way into the lives of more and more people (Byrne et al., 2019; Yuan et al., 2020). Despite their increasing success, their main applications are currently limited to command-and-control via short commands ("Play the next song"), to simple question-answering or to performing tasks via *slot filling*, that is a conversation pattern which allows the SA to request pieces of information from the user in a structured manner (Asri et al., 2017). Complex tasks such as *booking a table at a restaurant* are typically simplified as instances of slot filling (i.e., the interaction is shaped by the SA requesting slots like the *time* or *number of people* for the booking, and the user informing the SA about them). However, SAs could potentially help humans solve complex tasks in a more sophisticated way, for example, by building common ground, negotiating information or supporting deviations from the "happy path". For this, single-turn or strictly-designed interfaces are not adequate anymore, but a flexible and efficient way of interacting over multiple turns is required, not unlikely the way humans interact in a dialogue while they collaborate on a task.

Dialogue, however, comes with a new set of challenges for machines, including, but not limited to, context-sensitivity, anaphora, ellipsis and dynamic error management (Williams and Young, 2007; Grosz, 2018; Serban et al., 2018; Byrne et al., 2019; de Vries et al., 2020). SAs need to be able to handle these dialogue-specific phenomena, not only to assist in complex tasks, but also to make SA interactions more natural in general. In fact, studies suggest that humans generally prefer dialogue as a mode of interacting with SAs (de Vries et al., 2020).

One way to model dialogue in SAs and make SA interactions overall more natural, is to train the underlying natural language processing (NLP) modules on linguistic data that is representative of natural dialogue (Rieser and Lemon, 2011). To collect large-scale datasets, recent approaches have leveraged scenarios generated from simple templates where entity placeholders are replaced by possible entity surface forms (e.g., "Find a [CUISINE] restaurant" > "Find a Japanese restaurant"; e.g., Budzianowski et al., 2018; Wang et al., 2012). However, such template-based collection efforts may provide too fixed of a script for the dialogue and prime the participants into using specific words (de Vries et al., 2020).

Therefore, in this paper we present a framework to crowdsource dialogue data for the SA domain which is specifically aimed at limiting the kind of bias induced by the template-based approach while still allowing for collecting dialogue data in a low-resource and scalable way. Crowdsourcing, that is, relying on a large, remotely-located pool of workers who perform small tasks on a dedicated website like Amazon Mechanical Turk (AMT), is now a well-established and reliable method for collecting large-scale datasets in a time- and cost-effective way (Schnoebelen and Kuperman, 2010; Buhrmester et al., 2011; Garcia et al., 2020). More concretely, our approach pursues three goals: 1) The collected data should be of good quality, that is, the dialogues should be coherent, diverse and natural. 2) The data should be collected in a low-resource fashion, that is, with as little overhead as possible and relying to the greatest extent possible on existing technologies. 3) Our approach should allow for collecting data in languages other than English (in our case German), where fewer crowd-workers may be available.

Our contribution with this paper is threefold. 1) We describe a novel approach for collecting dialogue data which strikes a balance between limiting researcher-induced bias and a low-resource, scalable data collection setup. Rather than using template-based scenarios for eliciting the dialogues, we used *situated scenarios* formulated in a way that was aimed at reducing bias (scripting and priming). Our scenarios were designed to tap into the participants' situated knowledge, in order to afford them the opportunity to go about solving their task more freely, while at the same time avoiding explicit reference to relevant entities. Additionally, as method sections are typically short in dataset publications (e.g., in Budzianowski et al., 2018; Wen et al., 2017) and design choices are not always argued for, our in-depth description can help researchers collect dialogue data in a simple way. 2) In our data analysis, we present novel operationalizations of quality metrics and show that small efforts in the data collection setup can lead to less bias in the dialogue data, thus making it more ecologically valid. Ecological validity is a notion introduced by de Vries et al. (2020) into the NLP community, which specifies "the degree to which [data] generalize[s] to naturally occurring scenarios" (de Vries et al., 2020). A dataset is ecologically valid, and thus allows for such generalizations, if it consists of (simulations of) human-machine interactions, that is, data which "reflect[s] the intents and linguistic phenomena found in real-world applications" (de Vries et al., 2020). 3) We release ***Crow****dsourced* ***W****izard of Oz* ***D****ialogue dataset based on* ***S****ituated* ***S****cenarios* (*CROWDSS*), a dataset labeled with dialogue acts (DAs) which, to the best of our knowledge, is the first German task-oriented dialogue dataset. It can be used for a variety of NLP tasks like DA classification or response selection.

The paper is structured as follows. First, we describe what makes dialogue an efficient and flexible form of interaction and why dialogue is needed for SAs to be able to help users accomplish complex tasks in a natural way. Next, we review previous approaches to dialogue data collection, and we describe our own method, arguing for how our design choices fit in with our three goals (good-quality data, low-resource approach, feasibility in languages with limited crowd-worker availability). We briefly explain our annotation efforts, and go on to analyze our data regarding data quality. For this purpose, we compare our data to a sample of MultiWOZ (Budzianowski et al., 2018). De Vries et al. (2020) used that very corpus to demonstrate the presence of scripting and priming in datasets for NLP. Crucially though, MultiWOZ makes for an ideal comparison as it was collected in the same way as CROWDSS, with one key difference which is the stimuli used to elicit the dialogues (template-based vs. situated scenarios). Assessing pre-*script*-edness regarding task-relevant entities (scripting), lexical overlap between scenario and dialogue (priming) as well as diversity between dialogues elicited from the same scenario (scalability), we can show that small investments in the collection setup can greatly improve the dialogue quality.

Dialogue as the Mode for Solving Complex Tasks

Current SAs including popular services like Apple's Siri or Amazon's Alexa do a good job at single-turn commands and simple multi-turn interactions. However, SAs could also support humans in complex tasks like *finding a restaurant* and *booking a table* there, *comparing shopping items*, or *searching large databases* (Asri et al., 2017; de Vries et al., 2020). To the extent to which SAs are already capable of assisting humans in such complex tasks, they are typically implemented in a simplified way: a *restaurant booking task* would, for example, be implemented as *slot filling*, which limits the exchange to the assistant *requesting information from the user* and the user *informing the assistant* (e.g., in Rasa—Bocklisch et al., 2017). In human-human interaction, on the other hand, such tasks would be collaboratively solved in a sophisticated way using *dialogue* (Stalnaker, 1978; Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Clark, 1996; Pickering and Garrod, 2004; Xu and Reitter, 2018).

Human dialogue is a flexible mode of interaction without a turn limit, enabling information to be exchanged dynamically, depending on the dialogue flow, a sudden change of mind, new incoming information, etc. The linguistic features that constitute dialogue make it a very efficient way of negotiating information while building common ground (Stalnaker, 1978; Clark and Wilkes-Gibbs, 1986; Clark, 1996; Xu and Reitter, 2018): already-introduced entities can be referred to with shorter expressions (anaphora, Poesio and Rieser, 2011); understanding from context makes it possible to omit superfluous elements (context sensitivity, ellipsis, Levelt and Kelter, 1982); mutual understanding is continuously displayed and monitored (back-channeling, Schegloff, 1982) and if need be enforced (error management), and there is an elaborate system for floor management (turn-taking, Oreström, 1983). Its flexibility and efficiency may explain why dialogue is so easy to process for humans (Fox and Jean, 1999; Branigan et al., 2011) and why it presumably is also their preferred format of SA interaction (de Vries et al., 2020).

In sum, the flexibility and efficiency of dialogue as a mode of interaction as well as a (presumable) general human preference for this type of communication indicate that SAs need to be able to handle natural dialogue if they are to assist humans in more complex everyday tasks. Dialogue as an interaction mode is, of course, not limited to task-oriented settings where SAs help users get a task done, but also social SA interactions can (and should) adopt a dialogical style. However, the following discussion of previous work solely focuses on data collection for task-oriented systems.

Previous Work

Dialogue datasets have been collected or generated in three different ways, distinguished by who is interacting with whom in the data collection/generation process.

Dialogue Data Collection Settings

In a **human-human** setting, two or more human users talk or write to each other, thereby generating a dialogical interaction. Since all interlocutors are human and thus dialogue-competent, the aforementioned dialogue-specific phenomena are ideally present in and can be learnt from data collected in this way. Yet, humans talk differently to machines than with their fellow human beings (de Vries et al., 2020, but see discussion below). Thus, simply employing existing human-human dialogue corpora (e.g., recorded dialogues from customer service interactions) may not be representative enough of human-machine interaction to improve current SAs. The human-human setting was applied, among others, for the MultiWOZ corpus (Budzianowski et al., 2018), as well as by Wen et al. (2017), Eric et al. (2017), Asri et al. (2017), Byrne et al. (2019) and for training Google Duplex (Leviathan and Matias 2018; Chen and Metz 2019), a SA which can interact with businesses on behalf of customers in the restaurant booking domain.

Apart from that, data has also been collected in a **human-machine** setting where a human participant interacts with a machine, for example in the setup of the second and third DSTC challenge (Henderson et al., 2013). While this type of setup may serve to improve an already-existing system, the improvement can only happen within the capabilities of that system (Wen et al., 2017; Budzianowski et al., 2018). Crucially though, natural dialogue-specific phenomena will only be present to a limited extent in data collected in this way. They will not be present in the machine turns at all. The users, then, may speak to the machine in a dialogical way, but this typically results in the machine's failure to react "dialogically", and, in the long run, in the users' downscaling their linguistic behavior to a level adequate to their machine interlocutor (see below).

Furthermore, to overcome data scarcity, data has also been generated by having **two machines** interact with each other, for example by Shah et al. (2018) and by Rastogi et al. (2020). The machine-produced interactions are typically an exchange of intents or DAs coupled with entities which are subsequently translated into natural language, for example by crowd-workers. While researchers are saved from the time-consuming and error-prone task of annotating the dialogues, data collected like that can in no way serve to improve a SA's dialogue competency as neither the "user" nor the system are dialogue-competent. Nonetheless, even datasets collected in this way are called "dialogue datasets" and are used to train dialogue systems.

Following this, if the goal is to enable machines to handle dialogue in a natural way, only data collected from human-human interactions seems adequate (Budzianowski et al., 2018; Byrne et al., 2019). Dialogue data from human participants has either been collected in an overt setup where both participants know they are interacting with another human being (Asri et al., 2017) or in the style of Wizard-of-Oz (Kelley, 1983).

Wizard-of-Oz Framework

The Wizard-of-Oz (WOz) framework aims at striking a balance between the naturalness of human-human interaction while still accounting for the observation that humans talk differently to machines than with their fellow human beings. This is generally attributed to the fact that humans always adapt their language towards the recipient and what they perceive as the recipient's capabilities (Shatz and Gelman 1973; de Vries et al., 2020, but see discussion below).

Therefore, in a WOz collection, the participant playing the role of the user (here: *she*) is led to believe she is interacting with a machine. The machine, on the other hand, is enacted by a different participant, who plays the role of the assistant (Kelley, 1983). Naturally, the assistant (here: *he*) exhibits better natural language understanding (NLU) than machines and he produces a greater variety of answers to user requests than existing natural language generation (NLG) modules do (Byrne et al., 2019). In doing so, he generates system turns that are more representative of how human interlocutors would behave in similar dialogical scenarios (Merdivan et al., 2020). At the same time, he formulates his replies knowing that he is mimicking a machine, respecting crucial boundaries: he only "operates" within the task that he is "designed for", producing mostly task-oriented utterances and engaging less in social chat. On the other side, "[g]iven the human-level natural language understanding, [the participant playing the user] quickly realize[s she] can comfortably and naturally express [her] intent rather than having to modify behaviors as is normally the case with a fully automated assistant" (Byrne et al., 2019). Doing that, she generates user turns that are also more in line with natural dialogue as it would unfold between human interlocutors, employing, for example, context-dependent anaphora or elliptic structures. Yet, she also operates within the boundaries of an interaction with a machine which was designed for a specific task, and therefore avoids conversing all too freely with it (Byrne et al., 2019) as well as typical idiosyncrasies of human dialogue like back-channeling. It may, of course, be the case that an individual user participant becomes suspicious of the illusion of interacting with a machine, but this is not necessarily detrimental for the data, as this participant will still conform to the context of addressing a machine for accomplishing their task.

For human-machine interactions, there is a discussion regarding whether, in addition to the recipient-oriented talk, humans conceptualize machines as different, non-human entities, and (also) for that reason talk differently to them [de Visser et al. (2016) call this the "unique-agent hypothesis"]. Crucially, this would mean that human-machine interactions will always be of a different kind than human-human interactions. In contrast to this, the Media Equation theory, which originated from the *Computers as social actors* (CASA) framework (Nass et al., 1994; Reeves and Nass, 1996; Nass and Lee, 2001), argues that computers, too, are social actors, and that humans apply the same social rules and norms when interacting with them as they do when interacting with fellow human beings (de Visser et al., 2016). This theory can explain why people currently behave differently towards machines (i.e., due to their limited capabilities), but contrary to the "unique-agent hypothesis", it predicts that these differences will disappear with machines becoming more human-like. If machines take on different "personalities", possibly even user-adapted ones, users would still adapt their speech towards their machine interlocutors, but more so to the specific machine they are interacting with, to its "personal" style of behavior, rather than in a generic machine-directed way. For a discussion on how humanlike-ness is conceptualized and how it can potentially be reached by machines, refer to de Visser et al. (2016).

Since machines are a long way from being able to converse in a human-like way, whether or not human-machine interaction will always be fundamentally different from human-human interaction is beyond the scope of this paper. For now, we must reconcile the fact that current human-machine interaction lags behind human-human interaction (the latter being the model to strive for and learn from due to its advantages: flexibility, efficiency, suitability for complex tasks) and the knowledge that humans (at least for now) generally talk to machines differently. WOz provides a suitable solution for both these issues (de Vries et al., 2020) and has accordingly been applied in some of the major recent data collection efforts (e.g., Eric et al., 2017; Budzianowski et al., 2018; Byrne et al., 2019).

Additionally, WOz approaches try to simulate a context analogous to the eventual deployment context, which is "of utmost importance" for machine learning (de Vries et al., 2020). The Spoken Dialogue Systems community has always taken this simulation of the deployment context very seriously, actually striving for ecological validity and focusing on smaller, good-quality datasets (Rieser 2008; Rieser and Lemon 2011; Schlangen 2019). Lately, however, the focus has shifted to large-scale collections, both due to demands of huge data-driven models and thanks to the availability of online data collection platforms (e.g., Wang et al., 2012; Eric et al., 2017; Wen et al., 2017; Budzianowski et al., 2018; Wei et al., 2020). Such collections typically use templates to elicit data, which has been argued to induce researcher bias into the collected data, lowering its ecological validity (see below; de Vries et al., 2020). An interesting open question is then how we can optimally set up a WOz data collection on the web for a tradeoff between good-quality data and a low-resource collection setup.

(A)synchronous Interaction

Dialogue data from human interlocutors has either been collected from live interactions (Garcia et al., 2020) or in an asynchronous fashion (Eric et al., 2017; Wen et al., 2017; Budzianowski et al., 2018). Live interactions can be done with hired workers or with crowd-workers (e.g., over Facebook's ParlAI). In any case, a large pool of participants is required, because two participants have to be available at the same time to be paired up (Garcia et al., 2020), and in the case of crowd-workers there is a considerable risk of dropouts.

In asynchronous data collection, dialogues are collected on a turn-by-turn basis. This means that a participant's task only consists in continuing an ongoing dialogue with *one* turn. The dialogue including the new turn is then handed to the next available crowd-worker (in the opposite role) which can happen at any later point in time. This simplifies the setup and eliminates the need that two crowd-workers be available at the same time. Despite the involvement of multiple rather than two interlocutors, turn-based data collection has been shown to generate coherent data (Wen et al., 2017; Budzianowski et al., 2018). All-important coherence in dialogues is ensured by having participants read the dialogue history before they respond to it, as well as through a unifying scenario for the user participants and in some cases also a unifying user persona (Jonell et al., 2019).

Modality

Dialogue data collections further differ regarding whether the data is collected in spoken or written modality. As language is not only recipient-dependent, but also modality-dependent (Serban et al., 2018), data from spoken interactions is, naturally, most representative for the smart *speech* assistant domain. Collecting spoken rather than written data requires a more complex data collection setup in which participants record their utterances with microphones. These utterances, then, need to be transcribed by an automatic speech recognition (ASR) module, which can be error-prone (Asri et al., 2017).

While the spoken modality makes sense for live interactions, for asynchronous data collection, it poses the non-trivial problem of how the dialogue history (and even the scenario) should be presented to new crowd-workers who are continuing the dialogue. One option would be to present the dialogue history either completely in written form or with the last turn being synthesized. Yet, this would result in a mix of modalities where some or all of the dialogue history is in text form, but the crowd-workers are asked to phrase the dialogue continuation using their voice. A second option would be to synthesize the whole dialogue history using two different voices, but that would be rather strange for the crowd-workers as one of the two voices is supposed to be their own. Furthermore, only listening to the dialogue may make it more difficult for crowd-workers to familiarize themselves with the dialogue that they should continue, since with text one can more easily revisit what has already been discussed. Thus, for asynchronous data collection, using the spoken modality not only means more technical overhead, but also a modality mix which likely does not result in higher data representativity for the SA domain than completely replacing the spoken modality with the written one. Therefore, asynchronous datasets have been collected in the written modality (Eric et al., 2017; Wen et al., 2017; Byrne et al., 2019 [note that they also collect spoken data for the same dataset]; Budzianowski et al., 2018), often without even discussing modality as a factor.

Furthermore, with online chats having become a popular means of interaction, the line between the written and spoken modality seems to have become blurrier (Nishimaki 2014). Chats exhibit some of the features that previously had been proprietary to spoken language (Dürscheid and Brommer, 2009; Nishimaki, 2014), including, but not limited to shorter, spontaneously produced units, ellipsis, colloquial style, (near) synchronicity, and dynamic error management. At the same time, chat platforms like WhatsApp or Threema (where the modality focus traditionally was on written chats) also provide the possibility to send voice messages, and users often mix the written and spoken modality. Interestingly, when switching to the spoken modality, the messages are decidedly non-dialogical not permitting any form of interaction or feedback until the message is recorded, sent and listened to. In any case, collecting data in an asynchronous way (i.e., typing a message into a response field after viewing the dialogue history which is typically presented in differently-colored text boxes, aligned on either side of the window), could remind crowd-workers of chat interactions and stimulate a more oral-style language.

Summary

The main distinction regarding dialogue data collection approaches comes down to who is interacting with whom in the collection process, and, along with it, which purpose the dataset should serve: train better, dialogue-competent NLP models (human-human setting), improve an existing system (human-machine setting), or overcome data scarcity (machine-machine setting). Most datasets collected from human-human interactions apply the WOz framework in order to obtain data that is not only representative of natural dialogue, but also of a human-machine interaction context. WOz data collections can be distinguished regarding whether the participants interact live or asynchronously, and along with it, whether they do so in the written (both) or spoken modality (only live interaction).

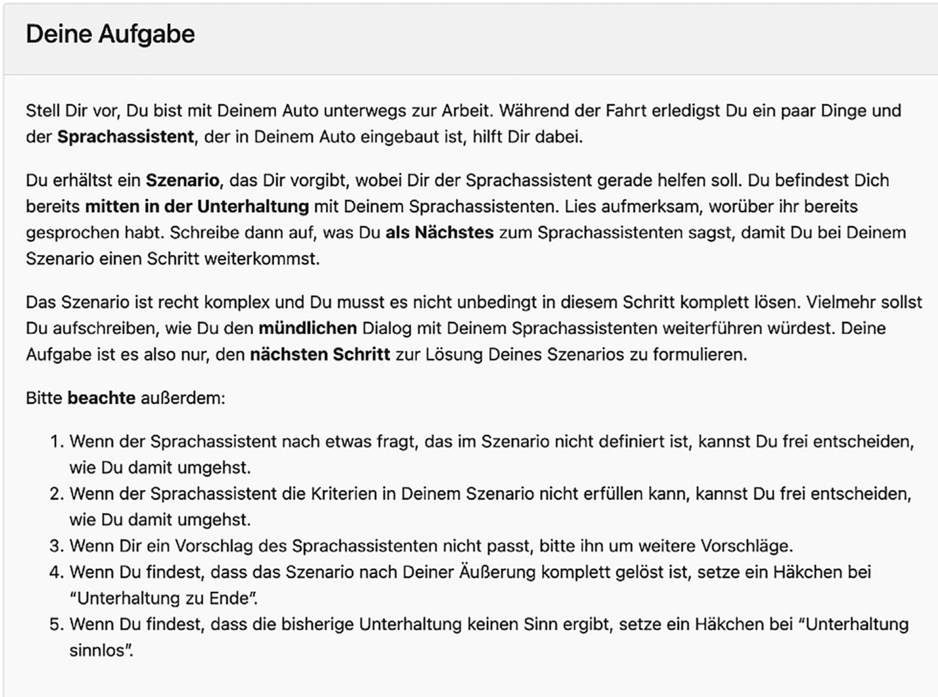
Our Approach

Based on the preceding overview, we collect written, task-oriented dialogues from human-human interactions in the WOz framework and an asynchronous fashion, aiming at a tradeoff between good-quality data and a low-resource collection setup. The dialogues are collected with the help of crowd-workers on AMT. We propose to collect data asynchronously as it is in accordance with our low-resource goal. Crucially, this design choice is also suitable for collecting data in languages that are less represented on AMT than English, as pairing up two simultaneously-available crowd-workers becomes obsolete. The asynchronous approach entails that we collect written data, which may lower the data's representativity. However, we argue that we can mitigate this issue by designing materials which stimulate an oral context (see below). Furthermore, we collect data in batches (i.e., all first turns of all dialogues in the first batch, then, all second turns of all dialogues in the second batch, etc.; see below) which further simplifies the collection setup.

We follow a similar approach to the one used to collect MultiWOZ (Budzianowski et al., 2018), as we crowdsource written dialogues from human interlocutors in an asynchronous WOz setup. However, de Vries et al. (2020) used MultiWOZ to exemplify the presence of scripting and priming in NLP datasets, arguing that this lowers its ecological validity. Scripting and priming in the data can only be induced by the materials used to elicit the dialogues. Therefore, rather than using template-based scenarios like in MultiWOZ, we propose to use *situated scenarios* (see below), aimed at collecting good-quality data despite the low-resource setup.

MATERIALS AND EQUIPMENT

In order to collect task-based dialogues, there needs to be a task that the crowd-workers playing the user need assistance with, and



Deine Aufgabe

Stell Dir vor, Du bist mit Deinem Auto unterwegs zur Arbeit. Während der Fahrt erledigst Du ein paar Dinge und der **Sprachassistent**, der in Deinem Auto eingebaut ist, hilft Dir dabei.

Du erhältst ein **Szenario**, das Dir vorgibt, wobei Dir der Sprachassistent gerade helfen soll. Du befindest Dich bereits **mitten in der Unterhaltung** mit Deinem Sprachassistenten. Lies aufmerksam, worüber ihr bereits gesprochen habt. Schreibe dann auf, was Du **als Nächstes** zum Sprachassistenten sagst, damit Du bei Deinem Szenario einen Schritt weiterkommst.

Das Szenario ist recht komplex und Du musst es nicht unbedingt in diesem Schritt komplett lösen. Vielmehr sollst Du aufschreiben, wie Du den **mündlichen** Dialog mit Deinem Sprachassistenten weiterführen würdest. Deine Aufgabe ist es also nur, den **nächsten Schritt** zur Lösung Deines Szenarios zu formulieren.

Bitte **beachte** außerdem:

1. Wenn der Sprachassistent nach etwas fragt, das im Szenario nicht definiert ist, kannst Du frei entscheiden, wie Du damit umgehst.
2. Wenn der Sprachassistent die Kriterien in Deinem Szenario nicht erfüllen kann, kannst Du frei entscheiden, wie Du damit umgehst.
3. Wenn Dir ein Vorschlag des Sprachassistenten nicht passt, bitte ihn um weitere Vorschläge.
4. Wenn Du findest, dass das Szenario nach Deiner Äußerung komplett gelöst ist, setze ein Häkchen bei "Unterhaltung zu Ende".
5. Wenn Du findest, dass die bisherige Unterhaltung keinen Sinn ergibt, setze ein Häkchen bei "Unterhaltung sinnlos".

FIGURE 1 | Instructions familiarizing the user participants with their task.

that they should solve collaboratively with the crowd-workers playing the assistant. The user participants (UPs) got instructions for their role and a scenario which specified the task for a given dialogue. The assistant participants (APs) also got instructions for their role and a database which they could query to help the UPs accomplish their task.

Instructions for User Participants

The instructions for UPs introduced a simulated situation: they were driving with their car to their workplace (see **Figure 1**)[1] and while doing so they should solve a specific task with the help of an intelligent in-car SA. That task was detailed in the scenario (see below). It was stressed multiple times that the UPs should carry out *only one step* in order to come closer to solving the task (see bold words in **Figure 1**). This should prevent UPs from taking too lengthy turns, which would not be representative of spoken dialogue, where pieces of information are typically negotiated step by step. Also, the instructions put emphasis on the fact that it was a simulation of an *oral interaction* taking place in a hands-free environment. The UPs were asked to read what they had already *talked* about with their SA and to write what they would *say* next to the SA if it were an *oral* dialogue.

The instructions further detailed a few other points that the UPs should consider, including that they could freely choose how to react if the SA could not meet their request or if they were presented with a question that they could not answer by relying on the scenario. They were furthermore encouraged to evaluate offers from the SA according to their own liking. These points were included to give the UPs more freedom, which should lead to more diverse dialogues. Furthermore, the UPs were instructed to tick a box, 1) if they deemed that the dialogue was completed, and/or 2) if they thought that the preceding dialogue history was incoherent. This built-in, crowd-sourced dialogue validation mechanism could make scaling easier. Lastly, they had the option to leave us a comment.

Situated Scenarios for User Participants

The task that the UPs should pursue in their dialogue (*finding a restaurant* and, in some scenarios, also *booking a table*, see below) was specified in a scenario. Asynchronous data collection has been shown to generate coherent dialogues (Wen et al., 2017; Budzianowski et al., 2018) as long as all UPs working on the same dialogue are presented with the same scenario and the dialogue history up to their turn. Template-based scenarios can help achieve coherence, but they arguably offer the participants too fixed of a script to solve the task and prime their lexical choices (de Vries et al., 2020). As our first goal was to collect good-quality data that is not only coherent, but also diverse and natural, our scenarios had to specify the right amount of information to, on the one hand, achieve a unified intent (in the interest of coherence), but, on the other hand, give enough freedom to the UPs, making sure that the dialogues are not completely pre-*script*-ed (in the interest of diversity and naturalness).

[1]Find translations of all German-language figures, tables and examples in the **Supplementary Material**.

In order to achieve all of that, we designed *situated scenarios* to tap into the participant's situated knowledge of *restaurant booking*. Situations such as *finding a restaurant* or *booking a table* are represented in our brain as complex simulations of perceived situations (Barsalou, 2009), which include information regarding relevant people and objects, typical actions, background settings as well as introspections, intentions and emotions. We can tap into this situated knowledge without having to spelling out every single detail of a situation (as do template-based scenarios), but using minimal lexical or visual cues (McRae et al., 2018). For example, it can be assumed to be common knowledge that some restaurants do not allow pets and hence "your dachshund Benno should also be allowed to tag along" (see Example 1 below) should be enough of a cue to let UPs know they should consider this criterion in finding a restaurant. Situated knowledge also includes what typically motivates a situation: people do not just book a restaurant matching a set of criteria for the sake of booking a restaurant, but they engage in such a situation with a specific motivation in mind. Therefore, to make the scenarios more realistic, we also included some background information about the person that the UPs were simulating as well as their motivation (e.g., in Example 1, the holidays in Brittany and the there-acquired love for French cuisine as background; taking out the girlfriend as motivation).

Example 1 shows an example scenario including a booking task:

"Example 1: Die letzten paar Urlaube hast Du mit Deiner Freundin in der Bretagne verbracht. Da hast Du die Küche dieses Landes lieben gelernt: Baguette, Croissants, Käse und Rotwein... Heute Abend möchtest Du mit ihr essen gehen. Euer Dackel Benno sollte auch mitkommen dürfen. Der Abend sollte nicht zu teuer werden, es muss aber auch nicht das billigste Restaurant sein. Finde ein passendes Restaurant und buche einen Tisch für Euch. Bringe außerdem die Adresse des Restaurants in Erfahrung."

English translation: "*You've spent your last couple of holidays with your girlfriend in Brittany. You've got to love the cuisine of that country: baguette, croissants, cheese and red wine... Tonight you want to take her out to dinner. Your dachshund Benno should be allowed to tag along. The evening shouldn't be too expensive, but it need not be the cheapest place either. Find a matching restaurant and book a table for you guys. Also, inquire about the restaurant's address.*"

Further, the information that we chose to provide for finding a restaurant should not jeopardize the naturalness of the dialogues on the level of wording. The scenario in Example 1 nicely demonstrates our efforts to avoid lexical priming in the dialogues. It specifies three criteria for finding a suitable restaurant (*French cuisine, dog-friendly, medium price range*) and two for booking a table there (*tonight, two people*). Crucially, none of these criteria are *explicitly* given *in full detail*: *French cuisine* is hinted at by the vacations in Brittany as well as typical French dishes, *pet-friendly* is paraphrased by means of "your dachshund Benno should also be allowed to tag along" (note that, in German, the word *Hund* [dog] is not used), *medium price range* can logically be deduced from "not too expensive, but not too cheap either", *tonight* is explicitly specified, the exact time, however, is not, and, lastly, *two people* must be derived from "you and your girlfriend". We used vague language and circumlocutions for all the criteria in the scenarios, in order to afford crowd-workers the opportunity to phrase their utterances in their own words, rather than priming them into using the exact same words as in the scenarios.

In total, we used ten different handcrafted scenarios as seeds for the dialogues (see Batch-wise data collection). Five of them only concerned *finding a restaurant*, and the other five also required *booking a table* there. Each scenario contained three search criteria, which mirrored the search criteria in the database that the APs had at their disposal (see below). Thus, the scenarios were so detailed that the UPs were likely to go about solving them in multiple steps, that is, engaging in a dialogue.

In sum, our situated scenarios describe the task in a way that should reduce priming while still cueing the target situation, and at the same time permit some degree of freedom for the UPs, and we expect them to yield coherent, yet diverse, and natural dialogue data.

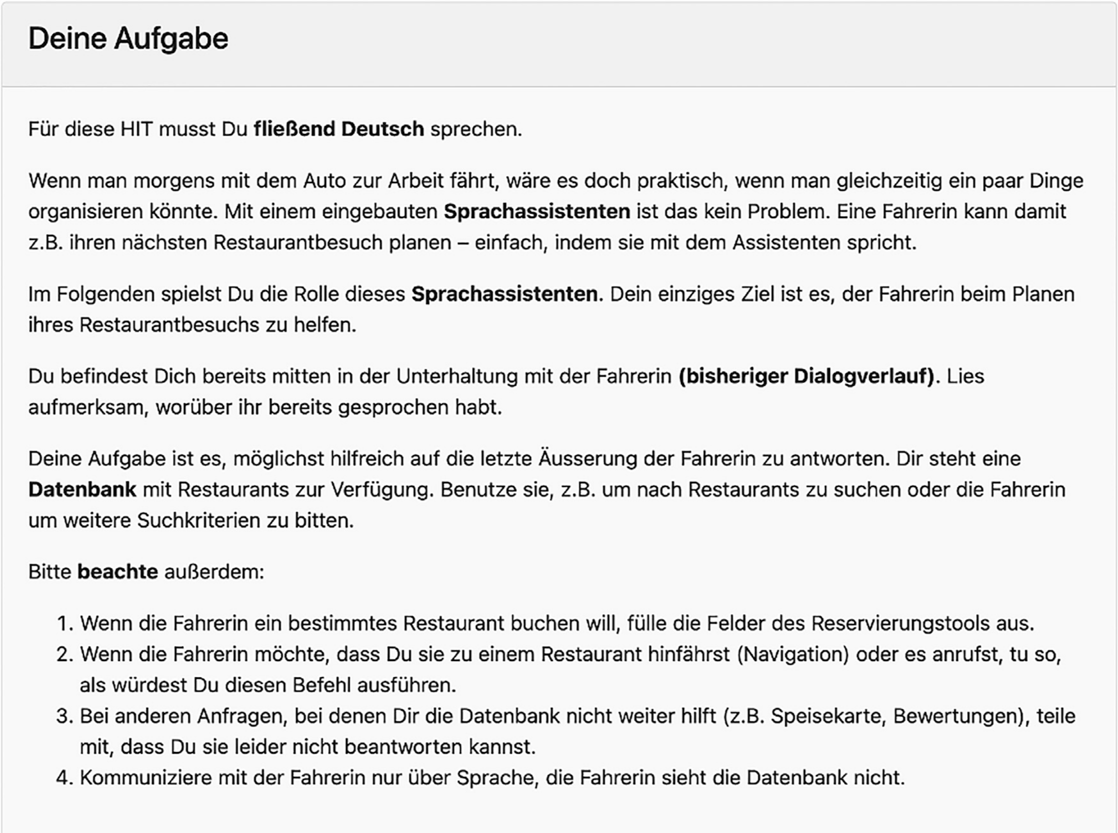
Instructions for Assistant Participants

For the APs, the instructions explained that their task was to play the role of an in-car SA and help a human driver find and book a restaurant (and *only* do that; see **Figure 2**). They were instructed to carefully read the dialogue history, extract relevant information from it and use it to query a simple database which we provided (see below), to find a restaurant matching the user's request. The APs also had the option to mark dialogues as completed or as incoherent, as well as leave us a comment.

APs were further informed about their capabilities as SAs: They could simulate making bookings, calling venues as well as navigating there. To simulate bookings, APs continuing the dialogues past the fourth turn (see Batch-wise data collection) saw three additional text fields in their graphical user interface (GUI; see below), where they were asked to enter any booking information (*name of restaurant*, *day* and *time of booking*, *number of guests*, respectively) as soon as the users had settled on one of them. Bookings were always successful, as they only consisted in filling in these fields. The database was the APs' only source of information. Thus, if a user asked for reviews about a restaurant, this request should be considered as "out of scope", as reviews were not part of the database (see below). Again, we put emphasis on the fact that the APs were engaging in what should be an oral interaction.

Database for Assistant Participants

The database (see **Figure 3**) that the APs had at their disposal was implemented as a simple combination of HTML forms and client-side JavaScript which was integrated into the HTML code for the GUI (see below). It consisted of 200 different restaurants, where each had a unique name and was defined regarding the search criteria cuisine (*American, Chinese, French, German, Italian, Mexican, Turkish, vegan*), location (*downtown, North, South, East, West, countryside*), price range (*cheap, moderate, expensive*) as well as an additional feature (*live music, wheelchair accessible, dog-friendly, featuring garden/terrace, accepts credit cards*; note that, for simplicity, these features were mutually exclusive). Furthermore, each



Deine Aufgabe

Für diese HIT musst Du **fließend Deutsch** sprechen.

Wenn man morgens mit dem Auto zur Arbeit fährt, wäre es doch praktisch, wenn man gleichzeitig ein paar Dinge organisieren könnte. Mit einem eingebauten **Sprachassistenten** ist das kein Problem. Eine Fahrerin kann damit z.B. ihren nächsten Restaurantbesuch planen – einfach, indem sie mit dem Assistenten spricht.

Im Folgenden spielst Du die Rolle dieses **Sprachassistenten**. Dein einziges Ziel ist es, der Fahrerin beim Planen ihres Restaurantbesuchs zu helfen.

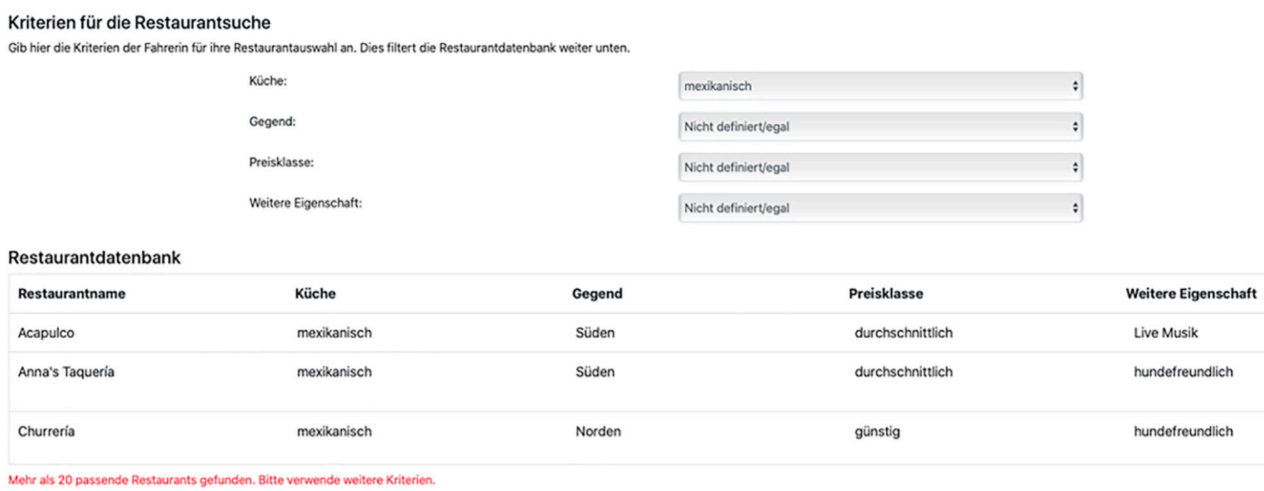
Du befindest Dich bereits mitten in der Unterhaltung mit der Fahrerin **(bisheriger Dialogverlauf)**. Lies aufmerksam, worüber ihr bereits gesprochen habt.

Deine Aufgabe ist es, möglichst hilfreich auf die letzte Äusserung der Fahrerin zu antworten. Dir steht eine **Datenbank** mit Restaurants zur Verfügung. Benutze sie, z.B. um nach Restaurants zu suchen oder die Fahrerin um weitere Suchkriterien zu bitten.

Bitte **beachte** außerdem:

1. Wenn die Fahrerin ein bestimmtes Restaurant buchen will, fülle die Felder des Reservierungstools aus.
2. Wenn die Fahrerin möchte, dass Du sie zu einem Restaurant hinfährst (Navigation) oder es anrufst, tu so, als würdest Du diesen Befehl ausführen.
3. Bei anderen Anfragen, bei denen Dir die Datenbank nicht weiter hilft (z.B. Speisekarte, Bewertungen), teile mit, dass Du sie leider nicht beantworten kannst.
4. Kommuniziere mit der Fahrerin nur über Sprache, die Fahrerin sieht die Datenbank nicht.

FIGURE 2 | Instructions familiarizing assistant participants with their task.



Kriterien für die Restaurantsuche

Gib hier die Kriterien der Fahrerin für ihre Restaurantauswahl an. Dies filtert die Restaurantdatenbank weiter unten.

Küche: mexikanisch

Gegend: Nicht definiert/egal

Preisklasse: Nicht definiert/egal

Weitere Eigenschaft: Nicht definiert/egal

Restaurantdatenbank

Restaurantname	Küche	Gegend	Preisklasse	Weitere Eigenschaft
Acapulco	mexikanisch	Süden	durchschnittlich	Live Musik
Anna's Taquería	mexikanisch	Süden	durchschnittlich	hundefreundlich
Churrería	mexikanisch	Norden	günstig	hundefreundlich

Mehr als 20 passende Restaurants gefunden. Bitte verwende weitere Kriterien.

FIGURE 3 | Restaurant database which the assistant participants could query to find venues matching the user's request.

restaurant featured an address including an indication of the distance from the users' current location which the UPs could request from the APs (see Example 1). The database could be filtered using any combination of the search criteria.

Presupposing that UPs would extract the intended search criteria from the scenario, request them in their utterance, and that the APs would use them to filter the database, four out of the ten scenarios were designed to lead to multiple entries in the

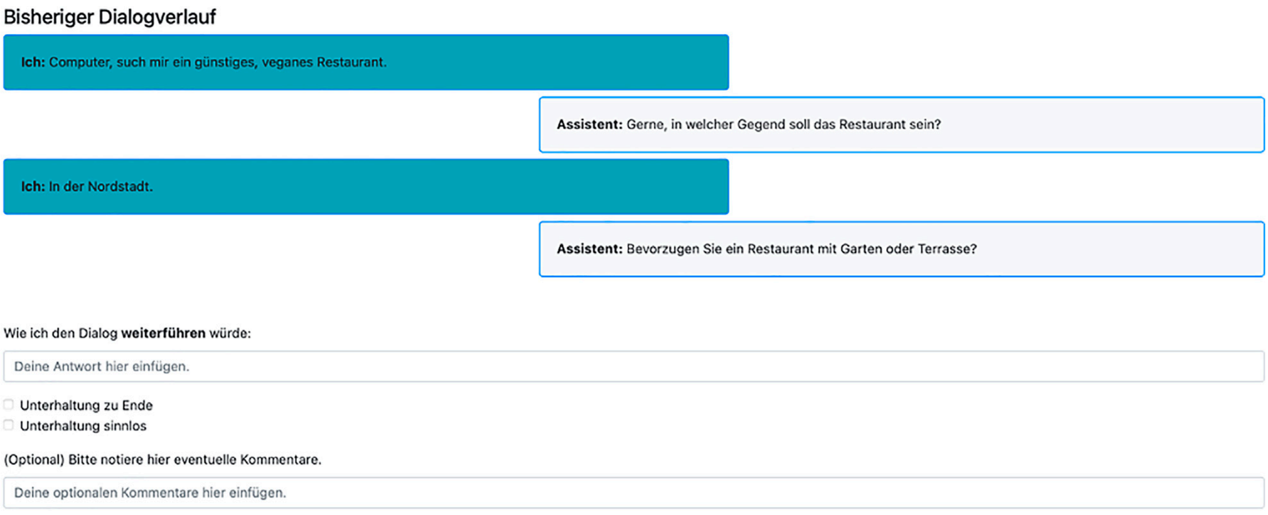


FIGURE 4 | Graphical User Interface for user participants with dialogue history, a response and optional comment field as well as end/spam check boxes.

FIGURE 5 | Map of fictional city "Bornberg" which was embedded into the Graphical User Interface of both participant groups, establishing common ground between them.

database, another four would instead lead to one entry each, and the last two to zero entries. This should again produce more diverse dialogues.

As the dialogues were gradually continued (see Batch-wise data collection), APs always saw the database output with the most recently-applied filters from the previous AP turn, but they could always modify them to yield different search results.

The architecture of AMT is centered around the publication of batches of *human intelligence tasks* (HITs). One batch contains one or more HITs, each of which can be assigned to one or multiple crowd-workers. The platform offers templates for common crowdsourcing setups (e.g., sentiment analysis) for task completion on-site. Alternatively, it can be used to only recruit participants for task completion on an external website. Further, it features an HTML editor where researchers can design their custom GUI for task completion on-site. In pursuit of our low-resource goal, we opted for this possibility and designed two complete GUIs inside AMT's HTML editor:

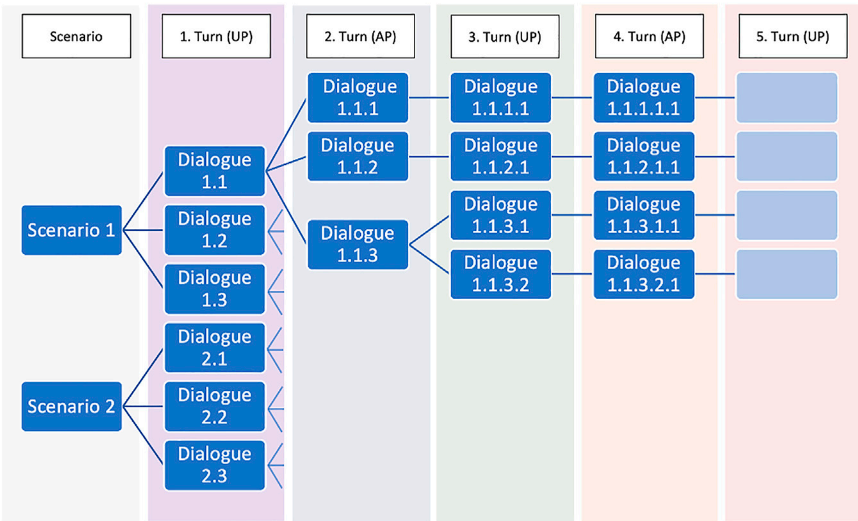


FIGURE 6 | One-to-many ratio showing how one dialogue seed was used to elicit multiple dialogue continuations.

one for UPs (see **Figure 4**), and one for APs including the searchable database.

For common grounding, participants of both roles further saw a map of the fictional city "Bornberg" in their GUI (see **Figure 5**). The map featured a pin indicating the user's current location. This should enhance the situatedness of the task of *finding a restaurant* in a specific geographic environment, by providing an additional cue in the visual modality.

METHODS

Having prepared instructions, scenarios, the database and the GUIs, we launched our data collection. We collected the data in batches, which is fully in line with AMT's architecture.

Batch-wise Data Collection

In batch 1, we presented the instructions and a scenario to UPs, who (according to their role) were assigned a scenario to pursue, and who had to initiate the dialogue. Applying a one-to-three ratio, we assigned each of the ten scenarios to three unique workers, to collect first utterances for each scenario, thus generating 30 dialogue beginnings in batch 1 (see **Figure 6**)[2]. Having collected them, we ran a post-processing script on the data, which spell-checked the utterances (relying on spaCy and pyspellchecker; Honnibal et al., 2020; Barrus 2019), corrected common misspellings (like wrong lower-casing of the polite form *Sie*) and flagged duplicate, empty or overly long answers (twelve words or more) for manual review. Our asynchronous batch-wise collection setup allowed us to do this offline and under no time pressure.

Next, in batch 2, we used the 30 dialogue beginnings as seeds to collect dialogue continuations (i.e., the second turn) from APs. The APs were presented with instructions, the database and a dialogue history which, at this point, only consisted of one previous turn. We again applied a one-to-three ratio and assigned each beginning to three unique APs, thus collecting 90 two-turn dialogues by the end of batch 2. We excluded 7 of them, as they had been marked as incoherent, and ran our post-processing script on the remaining 83 dialogues, manually reviewing flagged ones.

In batch 3 we then collected dialogue continuations from UPs, who were now instructed to familiarize themselves with the dialogue history as well and to continue the dialogues. Further, if they were working on a scenario which required *booking a table*, they were only now presented with this additional task as we wanted to prevent information overload in batch 1. For batch 3, we did not follow the one-to-three ratio between dialogue seeds and new continuations, but we assigned about a third of the 83 dialogues (30) to two UPs each (one-to-two ratio) and the rest (53) to only one UP each (one-to-one ratio). This was an arbitrary choice to limit the overall size of the dataset, as we primarily wanted to gain experience in collecting dialogue data in an ecologically-valid way, pathing the way for future domain-specific collections. At the end of batch 3 we obtained 113 unique dialogues. For batch 4 and onward, we continued assigning these 113 dialogues to participants following a one-to-one ratio, until they were complete. A dialogue was complete when one of the participants marked it as completed. Such dialogues were reviewed by us and, if complete, excluded from subsequent batches. Post-processing was performed after each batch.

Participants

In total, 57 unique participants contributed to the dialogues, of which 34 as UPs and 23 as APs. All crowd-workers had to be

[2]Note that the dialogue ids do not correspond to the dialogue ids in the dataset, as numbering in the dataset started with the first turn rather than with the scenario.

TABLE 1 | Comparison of CROWDSS with similar datasets (partially from Budzianowski et al., 2018; numbers for MultiWOZ and the dataset collected for Eric et al., 2017 are for the training split, for FRAMES the division is not specified, for CROWDSS the numbers refer to the whole dataset).

Metric	MultiWOZ Budzianowski et al. (2018)	FRAMES Asri et al. (2017)	Dataset collected for Eric et al. (2017)	CROWDSS
# Dialogues	8,438	1,369	2,425	113
Total # turns	113,556	19,986	Not provided	897
Total # tokens	1,490,615	251,867	Not provided	9,487
Avg. turns per dialogue	13.46	14.60	5.25	7.94
Avg. tokens per turn	13.13	12.60	9	10.1
Total unique tokens	23,689	12,043	1,601	906
# Slots	24	61	15	12
# Values	4,510	3,871	284	26

located in Germany. Additionally, they were asked to only participate if they were fluent in German. Some participants had to be excluded from further participation because of their non-fluent level in German, which became apparent when our post-processing script flagged their utterances due to, for example, misspellings. Participants should have an approval rate greater than 95% on AMT and we paid them above German minimum wage.

To uphold the WOz illusion, crowd-workers were consistently excluded from participating in the opposite role using AMT's built-in worker management, but not from participating in subsequent batches of the same role. However, in order to allow for more worker diversity in the data, we split batches into (up to eight) sub-batches and used "intra-role" worker exclusion, meaning that, for example, a UP from sub-batch 3.2 could not participate in sub-batch 3.3.

Annotation

We annotated our dialogues with DAs relying on the "Hierarchical Schema of Linked Dialog Acts" proposed by Pareti and Lando (2019). DA annotation schemes are typically developed either for human-human interaction (e.g., Bunt et al., 2010), or to meet domain-specific engineering requirements (e.g., the scheme used for MultiWOZ). Pareti and Lando's scheme bridges this gap as it was specifically developed to account for human-machine interaction, but in a broadly applicable, domain-agnostic way. It features "useful categories that can help […] understand the human dialog input as well as generate a suitable machine reaction" (Pareti and Lando 2019). The schema is hierarchical with three levels of granularity (but can be extended, also with lower-level domain-specific tags if needed; see **Table 1**), where a full tag is composed of three sub-tags, one for each level, for example *request.instruct.task*. Further, the schema is expectation-based, meaning that the tags should be "informative of the [most salient] conversational expectations at each given point in the dialog" (Pareti and Lando 2019). Hence, for example, offers made by the assistant to the user fall into the high-level category of *requests*, as they create an expectation on the user to *accept* (or *reject*) it. In the original proposal, DAs are further linked with each other (beyond simple order), but for reasons of simplicity, we did not use this feature.

Conducting an iterative annotation pilot involving three trained linguists, we generally deemed it intuitive to find a tag from the schema for a given sequence in our data, and would typically also agree on the choice of the tag. However, as the schema does not provide descriptions of the individual tags, we created our own annotation guidelines. We share them (see **Supplementary Material**), as this may help other researchers annotate dialogues with the same scheme. We did not modify the scheme, but disposed of some tags that were not needed for our data (e.g., the original scheme contains further *assert* tags for *opinions* and *elaborations*). Annotation of all dialogues was then performed by two of the three annotators using Doccano (Nakayama et al., 2018; see **Figure 7**). The dialogues were pre-segmented into utterances, but we did not restrict the number of consecutive tags that the annotators could assign to an utterance. This resulted in some cases (19 out of 897 utterances) in a different number of tags assigned by the annotators to the same utterance. For example, one annotator would use one tag (e.g., *request.instruct.task*), whereas the other would further segment the utterance and use two (e.g., *social.greetings.opening, request.instruct.task*). In order to compute inter-annotator agreement (Cohen's kappa), we considered the smallest annotated unit (by either annotator) and, when needed, we doubled the tags assigned by the other annotator to ensure an equal number of tags assigned to each utterance. The resulting inter-annotator agreement (excluding the dialogues annotated during the pilot) was very high at 0.91. This is encouraging, as it suggests that crowd-workers could perform this task in large-scale collections, increasing the potential of scalability. Finally, the few inconsistencies were reconciled.

For the data quality analyses, we also annotated both dialogues and scenarios with entities, labelling all entities that are needed for *finding a restaurant* (`cuisine, location, price range`), *booking a table* there (`number of people, day of booking, time of booking`) or that could be requested by the user and retrieved from the database (`address`). Additional features (`live music, wheelchair accessible, dog-friendly, featuring garden/terrace, accepts credit card`) were annotated as Boolean entities. Annotation was performed in Doccano by one trained linguist. For each entity, we also saved the corresponding surface form in the scenario or dialogue and additionally normalized that surface form into entity categories, leaving us with, for example, the surface form

FIGURE 7 | An example dialogue annotated with dialogue acts in Doccano.

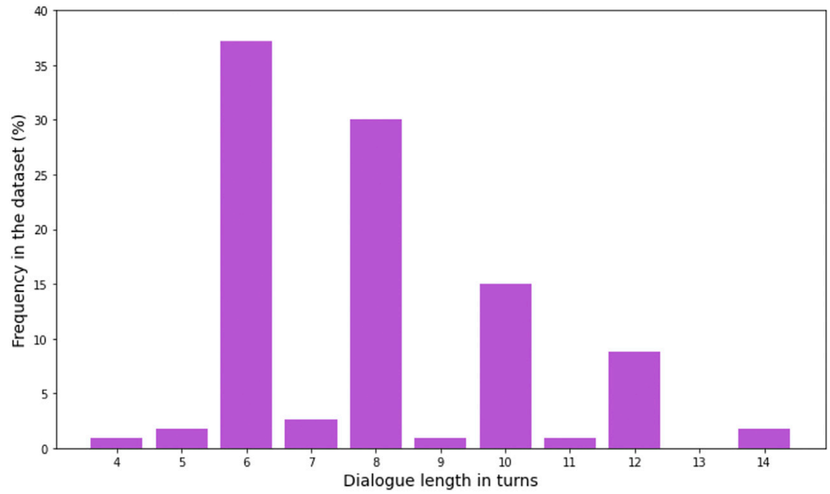


FIGURE 8 | Distribution of dialogue length in number of turns in CROWDSS.

"Küche des Reichs der Mitte" for the entity category *Chinese* for the entity `cuisine`.

Since our data quality analyses are partly comparative, we performed the same entity annotations for a sample of dialogues including the corresponding scenarios from MultiWOZ (see below).

RESULTS

In total, we collected 113 dialogues with a mean length of 7.94 turns per dialogue (*SD* = 2.13). The shortest dialogue consisted of four turns, the longest of 14 turns. Mean turn length was 10.1 tokens (*SD* = 3.98). The type-token-ratio (TTR) is 0.09 for the whole dataset and 0.71 on average for single dialogues. The two numbers differ substantially, as TTR is very sensitive to text length. In contrast, the Measure of Textual Lexical Diversity (McCarthy and Jarvis, 2010), a measure which avoids correlation with text length, is 67.72 for the whole dataset and on average 70.58 for single dialogues. **Table 1** presents a comparison between CROWDSS and similar datasets. **Figure 8** shows the distribution of dialogue length in number of turns in CROWDSS. The high number of dialogues ending with an assistant turn (even-number dialogue length) rather than with a user turn is due to the fact that APs typically marked dialogues as completed as soon as they had filled in all booking

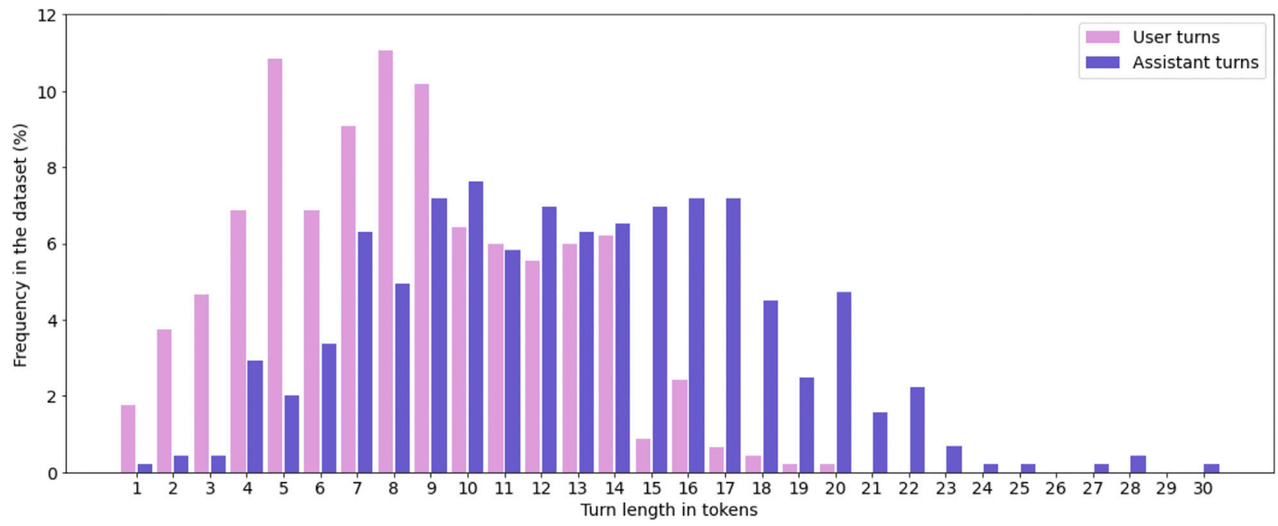


FIGURE 9 | Distribution of turn length in number of tokens for user and assistant turns in CROWDSS.

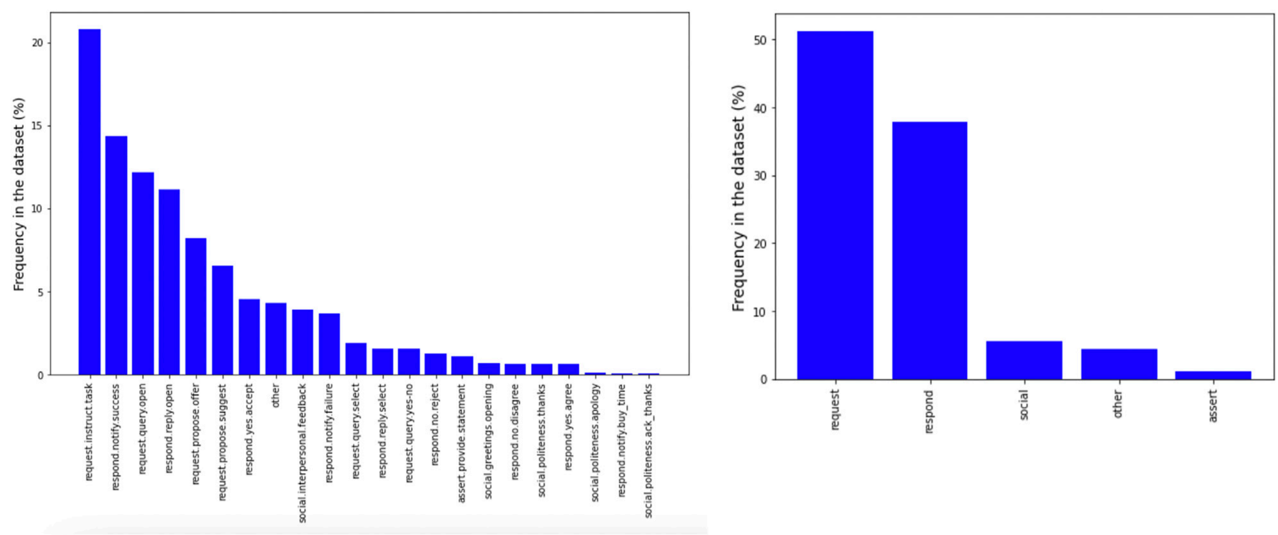


FIGURE 10 | Distribution of dialogue acts and high-level dialogue acts in CROWDSS.

fields and we manually confirmed that this was the case. **Figure 9** shows the distribution of turn length in tokens for user and assistant turns, respectively. **Figure 10** shows the distribution of all dialogue acts and of high-level dialogue acts, respectively (see Annotation). Finally, **Figure 11** shows the distribution of number of dialogue acts per turn.

Data Quality Analysis

Our first goal was to collect good-quality data that is both coherent, diverse and natural. In terms of coherence, we observe that our built-in spam detector was rarely used by the participants, suggesting that they had no problems making sense of the ongoing dialogue and continuing it with a meaningful next turn. Also, the optional comment field was not used to indicate any trouble regarding that. Furthermore, having annotated CROWDSS, it is our impression that the dialogues are coherent.

Diversity and naturalness, then, can be subsumed under the concept of ecological validity (see Introduction). A dataset can deviate from an ideal ecologically-valid methodology in five ways: 1) if it relies on synthetic language, or, given that the dataset contains authentic language, 2) if it was collected using an artificial task without real-world correspondence, 3) if it does not come

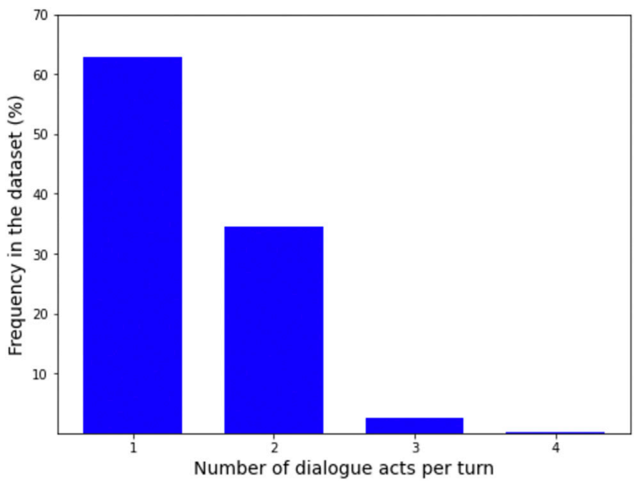


FIGURE 11 | Distribution of number of dialogue acts per turn in CROWDSS.

from users who may eventually benefit from a SA capable of a given, meaningful task, 4) if it contains single-turn interactions lacking the "conversational aspect", or 5) if template-based scenarios were used for data collection, which can lead to scripting and priming in the dataset (de Vries et al., 2020). WOz setups have typically aimed at sidestepping most of these issues. The use of template-based scenarios, however, can potentially be critical and affect the ecological validity by pre-*script*-ing the dialogues too much and by priming the participants' lexical choices.

We thus focus our data quality analysis on investigating the presence (or, ideally, the absence) of scripting and priming in CROWDSS. For that, we compare our dataset to the English-language MultiWOZ dataset (Budzianowski et al., 2018). We find that despite being in a different language, MultiWOZ makes for an ideal comparison. First, because de Vries et al. (2020) used that very dataset to show the presence of scripting and priming. Second, MultiWOZ was collected in the same way as CROWDSS (even the domain is the same) but with the potentially critical difference regarding the stimuli: template-based scenarios in MultiWOZ, handcrafted situated scenarios in CROWDSS. As for the language difference, we cannot exclude the possibility that the vocabularies for the task at hand are not equally rich in the two languages (e.g., in an extreme case, the English vocabulary could be so restricted that participants would make a given lexical choice, irrespective of a stimulus priming it or not). However, we argue that it is reasonable to assume similar vocabularies, given the closeness of English and German as well as the everyday nature of the task. For the analyses of lexical overlap (see Priming), we lemmatized all tokens so that different German case endings could not lead to an excessively low rate of overlap for German.

We first extracted a sample from MultiWOZ (n = 10,438 dialogues spanning multiple domains) to match the size of our dataset (n = 113). This was done by computing a random same-size sample from all single-domain dialogues in the restaurant domain. We observed that the restaurant dialogues in MultiWOZ

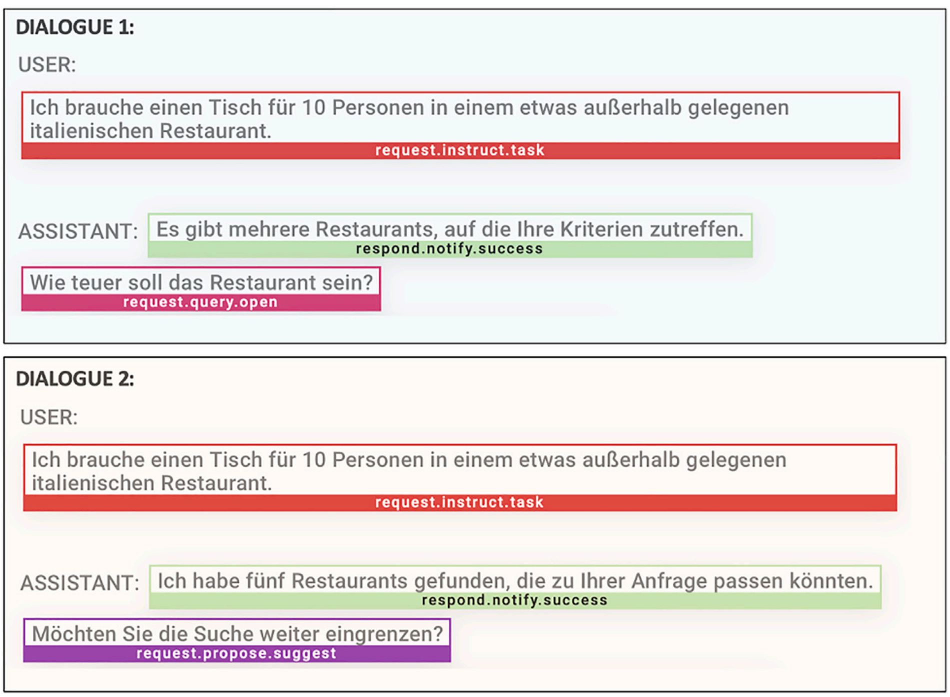


FIGURE 12 | Dialogue Act Sequence Similarity between turns elicited by same dialogue beginning: the two second turns share the first dialogue act and differ regarding the second one resulting in a sequence similarity of 0.5.

TABLE 2 | Entities in scenario and dialogue of MultiWOZ example dialogue.

Location	Order	Entity	Category	Surface form
Scenario	1	`location`	*south*	"south"
Scenario	2	`cuisine`	*international*	"international"
Scenario	3	`price_range`	*expensive*	"expensive"
Scenario	4	`cuisine`	*indian*	"indian"
Scenario	5	`number_of_people`	*4*	"4 people"
Scenario	6	`time_of_booking`	*T16:15*	"16:15"
Scenario	7	`day_of_booking`	*wednesday*	"wednesday"
Scenario	8	`time_of_booking`	*T15:15*	"15:15"
Scenario	9	`reference_number_req`	*reference_number_req*	"reference number"
Dialogue	1	`location`	*south*	"south"
Dialogue	2	`cuisine`	*international*	"international"
Dialogue	3	`cuisine`	*indian*	"Indian"
Dialogue	4	`number_of_people`	*4*	"4 people"
Dialogue	5	`time_of_booking`	*T16:15*	"16:15"
Dialogue	6	`day_of_booking`	*wednesday*	"wednesday"
Dialogue	7	`time_of_booking`	*T15:15*	"15:15"

(M = 9.16 turns/dialogue, SD = 2.69) generally are longer than the dialogues in CROWDSS (M = 7.94 turns/dialogue, SD = 2.14). For better comparability, we required the random sample[3] to consist of similar-length dialogues. The random sample had a mean dialogue length of 8.25 (SD = 2.16), which, as an unpaired t-test revealed, did not differ significantly from the mean dialogue length in CROWDSS (t = 1.08, p = 0.28). Next, we pre-processed all dialogue turns in both datasets, removing HTML-tags, punctuation and custom stop words (articles, pronouns, prepositions; ≈140 English stop words and ≈320 German stop words with all different case endings, where applicable), as well as tokenizing and lemmatizing the dialogues (relying on spaCy; Honnibal et al., 2020).

Scripting

Our first analysis was aimed at assessing how much freedom the participants had in terms of *what* they should accomplish in the dialogue. We thus compared the entities mentioned in a dialogue's user turns with the entities mentioned in the corresponding scenario (we excluded assistant turns because only users could potentially be influenced by the entities in the scenario). Entities are critical for the tasks of *finding* and *booking a restaurant* as they define key criteria like *cuisine* or *time of booking*. More specifically, we computed the overlap of entity categories (e.g., *Italian, American* for the entity `cuisine`) between dialogue and scenario as a proxy for how pre-*script*-ed a dialogue is: if an entity category in a user turn also appears in the scenario, this counts as overlap. For the MultiWOZ sample, on average 95% of the entity categories mentioned in the user turns in a dialogue also appear in the corresponding scenario for that dialogue (M = 0.95, SD = 0.10). In comparison, in our data, only 75% of entity categories in the user turns also appear in the corresponding scenario (M = 0.75, SD = 0.19). An unpaired t-test revealed that this difference is significant (t = 10.39, p < 0.001). The example dialogue from MultiWOZ in **Table 2** mirrors this result: all entity categories brought up in the user turns also appear in the corresponding scenario. The example dialogue from CROWDSS (**Table 3**), then, shows higher entity category overlap than average, but it points to another interesting aspect: the user asks for a restaurant "[i]n der Nordstadt" (*in the north of town*), even though location is not specified in the corresponding scenario. The map of "Bornberg" (see **Figure 5**) indicates the user's car location with a pin (at the border of the north and east part of town), and this may have led the user to pick *Nordstadt* as a location. The map, besides providing common ground for user and assistant, provides an additional (visual) source of information which is grounded in the task. The map can be used by the users to make an independent choice (picking *Nordstadt* or another part of town) and thus contributes to reducing scripting. Textual elements in the map (the label *Nordstadt*) may have primed the user's lexical choice, but this can easily be avoided, for example, by using a compass instead.

The differing levels of entity category overlap indicate that the user-assistant interactions in the MultiWOZ sample are almost entirely "limited by the complexity of the script" (de Vries et al., 2020; *script* being synonymous with *scenario*), and that the template-based *script* kept the users from making any independent choices which could have led to more diverse dialogues. In CROWDSS, on the other hand, a quarter of all entity categories in the user turns is not *scripted* by the scenario and, thus, they represent independent choices by the users. It seems that our efforts — intentionally not specifying all entities in the scenario and explicitly granting the users some freedom to make their own choices (see Materials and Equipment) — led to relatively diverse dialogues on the level of entity categories.

Second, we had a closer look at dialogues where the same entity appears twice, but in different categories. This would typically happen when the first choice of the user was not available, and an alternative was requested. Such cases should appear in both datasets: in CROWDSS, because we intentionally created scenarios that would lead to zero entries in the assistants' database, making a change of entity category necessary for the dialogue to continue (see Materials and Equipment); and in the

[3]Find the dialogue ids of the MultiWOZ sample in **Supplementary Material**.

TABLE 3 | Entities in scenario and dialogue of CROWDSS example dialogue.

Location	Order	Entity	Category	Surface form
Scenario	1	`number_of_people`	*2*	"Zum Muttertag möchtest du deine Mama zum Essen einladen"
Scenario	2	`cuisine`	*vegan*	"Ihr seid beide sehr naturverbunden und esst keine tierischen Lebensmittel"
Scenario	3	`garden/terrace`	*TRUE*	"Da das Wetter schön sein soll, möchtet ihr gerne draußen sitzen können"
Scenario	4	`price_range`	*cheap*	"Du befindest dich gerade auf einer finanziellen Durststrecke und hast nur ein begrenztes Budget"
Scenario	5	`day_of_booking`	*tomorrow*	"morgen"
Scenario	6	`time_of_booking`	*noon*	"Mittagessen"
Scenario	7	`address_req`	*address*	"Adresse"
Dialogue	1	`price_range`	*cheap*	"günstiges"
Dialogue	2	`cuisine`	*vegan*	"veganes"
Dialogue	3	`location`	*north*	"In der Nordstadt"
Dialogue	4	`garden/terrace`	*TRUE*	"mit Garten"
Dialogue	5	`number_of_people`	*2*	"2"
Dialogue	6	`day_of_booking`	*tomorrow*	"morgen"
Dialogue	7	`time_of_booking`	*noon*	"Mittag"

MultiWOZ sample, because there are scenarios following an if-then-logic (e.g., from the scenario in **Table 2**: "The restaurant [. . .] should serve international food. If there is no such restaurant, how about one that serves indian [sic!] food [. . .] you want to book a table [. . .] at 16:15 on wednesday [sic!]. If the booking fails how about 15:15"). Across all dialogues of the MultiWOZ sample, there are 43 entities which appear twice, but in different categories, and for all of them not only the category of the first mention, but also the category of the second mention appears in the scenario. Thus, not only the *first choice* of the user has been scripted, but also the *alternative* that they request when the first choice is unavailable. In comparison, in CROWDSS, out of a total of 31 entities appearing twice in a dialogue, but with different categories, only in four cases both the category of the first mention and second mention appear in the corresponding scenario. These are cases where location is mentioned twice in both the scenario (e.g., as "in deiner Nähe im Osten der Stadt" [*close to your location in the east of town*]) and the dialogue, where the user would first ask for a restaurant *in the east* and then one *nearby* (or vice versa). Crucially, though, in 27 cases the users mention the second entity in a category that is not specified in the scenario. While it is not a surprise that in the MultiWOZ sample all second entity mentions are s*cripted* with regard to the entity category (as the scenarios explicitly predefine the behavior in case a request fails), it is encouraging that the users in CROWDSS continued the dialogues making their own choices in the face of a failed first request, which again increases the dialogue diversity.

Third and last, we investigated whether the dialogues are *scripted* with regard to the entity order. That is, if the order in which entities are mentioned in the scenario provides a sort of a script which is reproduced in the order of entity mentions in the dialogue. For that, we simply compared the order of entities as they are brought up in the dialogue with the order in which they are mentioned in the corresponding scenario (ignoring entities which are specified in the scenario, but not mentioned in the dialogue). In the MultiWOZ sample, 46 of 113 dialogues exhibit the exact same order of entities between dialogue and scenario. In CROWDSS, this is only true of five out of 113 dialogues. Again, the examples in **Table 2** and **Table 3** show that the entity order is identical between dialogue and scenario in the MultiWOZ example, whereas the entities appear in a very different order in the example from CROWDSS. This result, then, also suggests that our situated scenarios (where some entities had to be deduced from multiple pieces of information spread across the scenario, e.g., in **Table 3** the scenario ended with "buche einen Tisch für Euch", where *Euch* had to be derived from "Zum Muttertag möchtest Du Deine Mama zum Essen einladen", at the very beginning of the scenario) led to more diverse dialogues than the scenarios used in MultiWOZ. An interesting open question is whether people have a natural preference for the order in which entities are mentioned. However, our data may be too small to draw any conclusions about this.

Priming

Our next analysis is aimed at assessing how much freedom the participants had in terms of *how* they lexically accomplished their task in the dialogue. For that, we looked at the lexical overlap between a given scenario and the corresponding user turns (we again excluded assistant turns because only the users could potentially be primed by the scenario). Let s be the scenario, u the user turns in a dialogue, and L_s and L_u the set of content word types in s and u, we computed the *lexical overlap* between s and u as

$$LexOverlap_{su} = \frac{|L_s \cup L_u|}{|L_s \cup L_u|}$$

that is, the proportion of content word types in a scenario which are re-used in the user turns. For the MultiWOZ sample, on average more than 50% of the scenario's content word types also appear in the user turns of the corresponding dialogue ($M = 0.51$, $SD = 0.11$). In comparison, in CROWDSS on average only 15% of content word types in the scenario also appear in the user turns ($M = 0.15$, $SD = 0.06$). An unpaired t-test revealed that this difference is significant ($t = 30.29$, $p < 0.001$).

Second, as most of our efforts were focused on how to phrase the *entities* in the scenario, in addition to the lexical priming analysis on the whole content vocabulary, we also compared the lexical overlap for entities only. Compared to our previous analysis of entities, we are now not looking at the normalized categories but at the overlap between entity surface forms in the user turns with entity surface forms in the corresponding scenario. If an entity surface form in a user turn also appears in the exact same form in the scenario, this counts as overlap. For the MultiWOZ sample, on average more than 84% of the user turn entity surface forms also appear in the corresponding scenario (M = 0.85, SD = 0.16). In comparison, in CROWDSS on average only 15% of user turn entity surface forms also appear in the scenario (M = 0.15, SD = 0.15). An unpaired t-test revealed that this difference is significant (t = 33.27, p < 0.001). The examples in **Table 2** and **Table 3** nicely illustrate that: The surface forms between user turns and corresponding scenario in the MultiWOZ example overlap in their entirety. In the example from CROWDSS, on the other hand, only "morgen" (*tomorrow*) for the day of booking is mentioned in the exact same form in the corresponding scenario. All the other entities take on different surface forms, mainly thanks to the circumlocutions used in the scenario.

Both looking at the whole content vocabulary and looking at entities only, the results suggest that our scenarios led to considerably less priming in the dialogues, when compared to the MultiWOZ sample. Budzianowski et al. (2018) argue that their template-based approach provides "easy-to-follow goals for the users [which] resulted in a bigger diversity and semantical richness of the collected data". However, contrary to this claim, our analysis on the sample appears to confirm the observation by de Vries et al. (2020) that in the worst case the user participants in MultiWOZ seem to have copy-pasted parts of the scenario into their utterances, and in the best case were still "heavily influenced" by the scenario, leading to "unnatural" requests. In CROWDSS, on the other hand, it appears that our relatively small efforts to create "priming-reduced" scenarios encouraged the user to formulate their utterances using their own vocabulary, generating more diverse and natural data.

Scalability

The preceding analyses suggest that our scenarios led to good-quality data. However, handcrafting scenarios like we did is more time-consuming than adopting a template-based approach, even if it arguably leads to more ecologically-valid data. As we were not only aiming at good-quality data, but were also interested in a low-resource approach, we only designed a small set of ten scenarios, which we used as seeds to elicit dialogues by assigning them to participants in a one-to-many ratio. In order to evaluate this tradeoff between good-quality data and low-resource approach, we investigated whether using the same seed still led to diverse dialogue continuations. If this is the case, it would speak for the scalability of our approach, as one could invest time in handcrafting a small set of situated scenarios and elicit diverse dialogues from the same scenario.

In our case, we not only assigned the scenario seeds to different participants, but also assigned dialogue beginnings after the first and in some cases after the second turn to more than one participant as seeds (see Batch-wise data collection). Specifically, we applied a one-to-three participant assignment ratio for collecting batch 1 and 2, and a mixed one-to-one (53 dialogues)/one-to-two ratio (resulting in 30 pairs with the same second turn) for batch 3. In our analysis we look at how diverse dialogues sharing an identical beginning turn out to be from the point where they continue on their own branch (as an example, in **Figure 6**, the dialogues 1.1.1, 1.1.2 and 1.1.3 all share the same scenario and first turn and continue on their own branch from the second turn on). We identified 83 dialogues stemming from 30 dialogue beginnings (unique combinations of scenario + first turn) in order to compare the continuations at two points: 1) at turn/batch 2, and 2) at the end of the dialogues. 53 of these 83 dialogues continued on their own branch at turn 2 (see above). The remaining 30 were selected by randomly choosing one from each of the 30 dialogue pairs which continued on their own branch only at turn 3.

We analyzed each set of dialogues sharing an identical beginning both on the level of DA sequences and vocabulary. A low degree of DA sequence overlap would mean that dialogues with an identical beginning differed substantially regarding *what* participants wanted to accomplish with their utterances. Similarly, a low degree of vocabulary overlap within one set would mean that *how* (i.e., with what words) the participants went about accomplishing their task varied, as they used different words to continue these dialogues, despite their identical beginnings.

For the DA comparison, we relied on Python's built-in SequenceMatcher which looks for the longest contiguous sequences of elements and computes a measure of similarity ranging from 0 to 1. Consider the two second turns which were elicited using the same dialogue beginnings in **Figure 12**. Here, the sequence similarity at the level of the second turn would be 0.5.

We computed the similarity of DA sequences for all possible pairs of dialogues within one set and averaged these similarities over the number of pairs within that set. For the vocabulary comparison, we computed the lexical overlap (see above) between any dialogue pair in the set as the ratio of overlapping word types over all word types in the pair and then averaged over all possible pairs within that set.

Looking only at the second turn, there is a mean similarity of DA sequences of 41% (M = 0.41, SD = 0.28) and a mean lexical overlap of 32% (M = 0.32, SD = 0.15). Thus, both on the level of DAs and vocabulary, these second turns *turned out* to be rather diverse, despite their identical dialogue beginnings. Looking at all the turns, there is a mean similarity of DA sequences of 45% (M = 0.45, SD = 0.14) and a mean lexical overlap of 40% (M = 0.40, SD = 0.07). Thus, the diversity found for the second turns is maintained until the end of the dialogues. There is a slight decrease in diversity (i.e., an increase in similarity/overlap), but that is to be expected given that there is a finite set of DAs and relevant word types in the given context. The decrease in vocabulary diversity can further be explained by the fact that things which are not addressed in the second turn in one dialogue of a given set may be addressed in a later turn in a different dialogue of the same set. In that case, however, things are done in a different order, which increases the diversity among dialogues on a structural level.

DISCUSSION

Below, we discuss our three goals *good-quality data*, *low-resource approach*, *feasibility in languages with limited crowd-worker availability*.

Good-Quality Data

Looking at scripting and priming, CROWDSS seems to be of higher quality than the MultiWOZ sample. Yet, MultiWOZ is a much larger dataset and we cannot exclude the possibility that the random sample we used is not representative of the dataset as a whole. Our analyses are, however, in line with de Vries et al.'s (2020) observations about MultiWOZ. While CROWDSS does not exhibit any of the other deviations from an ideal, ecologically-valid data collection methodology as listed in de Vries et al. (2020; see above), and despite the encouraging results for scripting and priming, our dataset is in no way perfectly representative of human-machine interaction either. Especially the fact that we collected written data for the SA domain is disadvantageous. As explained, it is, however, not meaningful to collect spoken data in an asynchronous fashion. To collect spoken data, a live interaction on a dedicated platform would have been necessary, but that did not fit in with our second and third goal (*low-resource, feasibility in languages with limited crowd-worker availability*). Therefore, we restricted ourselves to putting emphasis on the fact that the dialogues are a simulation of spoken SA interactions (see Materials and Equipment). The mean turn length in tokens is relatively short in CROWDSS (M = 8.4, SD = 1.7, compared to M = 11.46, SD = 2.37 in the MultiWOZ sample), which is encouraging considering the fact that spoken dialogue typically consists of shorter utterances compared with written interaction. Hauptmann and Rudnicky (1988) report an average command turn length of 6.1 tokens for a WOz setup in the spoken modality (speaking to a computer/wizard; see also Fraser and Gilbert 1991). In sum, our investments into reducing scripting and priming seem to have paid off.

Low-Resource Approach

Our second goal was to collect dialogue data with as little resources as possible. For that reason, we collected data in an asynchronous way, which essentially obviated the need for a technically more complex setup that would enable live interactions. The asynchronous setup entailed that we collected written data, which was again low-resource because participants only needed a screen and a keyboard. Microphones and an ASR module on our side were not necessary. Relying only on AMT's HTML editor naturally constrained our design possibilities, but it is important to note that we were able to design fully functional and aesthetically appealing interfaces. This made it possible to collect all data on-site, rather than having to host the data collection on an external website, which would have required more resources.

Thus, it turned out to be feasible to reduce the technical overhead and only rely on an existing platform for data collection. At the same time, there was considerable manual overhead, mostly in between the batches for running a post-processing script on the most recent batch, reviewing flagged dialogues, excluding poorly performing participants, accepting the HITs so that the participants would receive payment, and preparing the following batch. Most of these steps can be automated using scripts, and the manual review in between batches can be skipped or downscaled depending on where the compromise between quality and quantity should be made.

An encouraging finding regarding the manual overhead is that our resource-friendly approach to only handcraft ten scenarios and use them to elicit more than the tenfold of dialogues still appears to have led to diverse dialogues. Therefore, we argue that the investments into good-quality data that we propose, namely the situated scenarios, are also implementable in large-scale dialogue data collections, as it does not seem necessary to have one unique scenario per dialogue. Thus, it appears that collecting good-quality data can go hand in hand with a low-resource setup, which is a step towards reconciling quality with quantity.

Feasibility in Languages With Limited Crowd-Worker Availability

Our third goal was to make sure that our data collection approach worked for German, where fewer crowd-workers are available compared to English. This proved to work well for our data collection. We see two main reasons for that. First, thanks to the batch-wise setup, one participant could contribute to multiple dialogues, though always in the same role. Using sub-batches we could still ensure diversity in participants which, if the data is used for training an algorithm, could allow the model to generalize better (Geva et al., 2019). Second, the asynchronous setup obviated the pressure to pair up two simultaneously available crowd-workers for a live interaction. Instead, we could launch a batch and simply wait until all dialogues were continued which would typically take a couple of hours.

Since we only collected a small dataset, we did not test the boundaries of German-speaking crowd-worker availability and we can therefore not say how large the German worker pool is. However, as NLP datasets are needed for languages other than English where they often are collected in ways that are not feasible in smaller languages (e.g., live interactions) we argue that our approach is a step in the right direction, enabling dialogue data collections for different languages than English.

Future Research

The main limitation of this work is the small size of the dataset. Thus, a larger dataset collected with our situated scenarios may be needed to strengthen the generalizability of the analyses reported above. Another limitation is the written modality which we had to rely on for resource and crowd-worker availability reasons. A next step could be to compare our dataset to a corpus of spoken SA interactions (e.g., Siegert, 2020) in order to evaluate potential modality-induced differences. Speaking of modalities, it would also be interesting to include more visual elements like the map in the scenarios, which could enhance the situatedness of the task and further reduce scripting and priming. Lastly, it could be worth combining a low-resource template-based approach with our suggestions concerning vague language and circumlocutions. One could use *whole* sentences formulated along these lines to build scenarios, rather than just inserting one-word entities in designated placeholders in otherwise ready-made scenarios, as is the case with the traditional template-based approach.

CONCLUDING REMARKS

In this paper, we have presented a novel approach for dialogue data collection. We argue that our framework makes it possible to crowdsource good-quality dialogues in a fairly low-resource fashion which is furthermore feasible in other languages than English. Additionally, we show that putting little effort in creating instructions and scenarios for the participants can lead to better-quality data than a template-based approach like in MultiWOZ. Our comparison with a sample from MultiWOZ suggests that our endeavors led to significantly less scripting and priming and thus more ecologically-valid dialogue data.

With the shift from rule-based to machine learning-based NLP systems, recently, datasets have focused heavily on quantity, under the assumption that the mass of data still comes in at least "appropriate" quality (de Vries et al., 2020). While quantity will always come at the cost of quality, we argue that small investments in the collection setup such as the ones we propose, and which are also possible for large-scale collections, can go a long way in improving data quality. WOz is a very helpful tool to "bootstrap out of [the] chicken and egg problem" (de Vries et al., 2020), that is, the problem that we do not know how humans would talk to human-like machines if they existed, however to make them come into existence we need data from this type of interaction. We argue that if one puts time and effort into a WOz collection, it is worth to also invest in reducing researcher bias and instead allowing the participants to interact as naturally as possible in the given context. Taking it to the extreme, you could even wonder what WOz data can give you beyond a machine-generated "dialogue" if you do not afford participants the opportunity to phrase their utterances in a natural way.

DATA AVAILABILITY STATEMENT

The dataset presented in this study can be found in an online repository. The name of the repository and accession number can be found here: https://dx.doi.org/10.24406/fordatis/124.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YF and AZ contributed to conception and design of the data collection. YF managed the data collection process and performed data quality analyses. YF wrote the first draft of the manuscript. YF and AZ wrote sections of the manuscript. YF and AZ contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to acknowledge Anna Sauer, Valmir Kosumi and Zahra Kolagar for their contribution to the data collection and analysis as well as Junbo Huang for his feedback on the article. We acknowledge the support of the Federal Ministry for Economic Affairs and Energy (BMWi) project SPEAKER (FKZ 01MK20011A).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2021.686050/full#supplementary-material

REFERENCES

Asri, L. E. l., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., et al. (2017). "Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems," in Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 2017, 207–219. Available at: https://www.aclweb.org/anthology/W17-5526

Barrus, T. (2019). *Pure Python Spell Checker Based on Work by Peter Norvig*. Available at: pypi.org/project/pyspellchecker/.

Barsalou, L. W. (2009). Simulation, Situated Conceptualization, and Prediction. *Phil. Trans. R. Soc. B* 364 (1521), 1281–1289. doi:10.1098/rstb.2008.0319

Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017). Rasa: Open Source Language Understanding and Dialogue Management. *ArXiv*, 1–9. http://arxiv.org/abs/1712.05181.

Branigan, H. P., Catchpole, C. M., and Pickering, M. J. (2011). What Makes Dialogues Easy to Understand? *Lang. Cogn. Process.* 26 (10), 1667–1686. doi:10.1080/01690965.2010.524765

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). "MultiWoz - A Large-Scale Multi-Domain Wizard-Of-Oz Dataset for Task-Oriented Dialogue Modelling," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, Brussels, Belgium, October–November 2018, 5016–5026. doi:10.18653/v1/d18-1547

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspect. Psychol. Sci.* 6 (1), 3–5. doi:10.1177/1745691610393980

Bunt, H., Jan, A., Jean, C., Choe, J. W., Alex, C. F., Hasida, K., et al. (2010). "Towards an ISO Standard for Dialogue Act Annotation," in Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC, Valletta, Malta, May 2010, 2548–2555.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., et al. (2019). Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. Proceedings of the Conference, 4516–4525. doi:10.18653/v1/d19-1459

Chen, B. X., and Metz., C. (2019). "Google's Duplex Uses A.I. To Mimic Humans (Sometimes). *New York Times*. May 22, 2019. Available at: https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html.

Clark, H. H. (1996). *Using Language*. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/CBO9780511620539

Clark, H. H., and Wilkes-Gibbs., D. (1986). Referring as a Collaborative Process. *Cognition* 22 (1), 1–39. doi:10.1016/0010-0277(86)90010-7

de Vries, H., Bahdanau, D., and Manning, C. (2020). Towards Ecologically Valid Research on Language User Interfaces. ArXiv http://arxiv.org/abs/2007.14435.

de Visser, E. J., Monfort, S. S., Ryan, M. K., Melissa, A., Smith, B., McKnight, P. E., et al. (2016). Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *J. Exp. Psychol. Appl.* 22 (3), 331–349. doi:10.1037/xap0000092

Dürscheid, C., and Brommer, S. (2009). Getippte Dialoge in Neuen Medien. Sprachkritische Aspekte Und Linguistische Analysen. *Linguistik Online* 1 (37), 3–20. Available at: https://www.zora.uzh.ch/id/eprint/24486/.

Eric, M., Krishnan, L., Charette, F., and Manning, C. D. M. (2017). "Key-Value Retrieval Networks for Task-Oriented Dialogue," in Proceedings of the Conference, Saarbrücken, Germany, August 2017, 37–49. doi:10.18653/v1/w17-5506

Fox, Tree., and Jean, E. (1999). Listening in on Monologues and Dialogues. *Discourse Process.* 27 (1), 35–53. doi:10.1080/01638539909545049

Fraser, N. M., and Gilbert, G. N. (1991). Simulating Speech Systems. *Comput. Speech Lang.* 5 (1), 81–99. doi:10.1016/0885-2308(91)90019-m

Garcia, Chiyah., Francisco, J., Lopes, José., Liu, X., and Hastie, H. (2020). "CRWIZ: A Framework for Crowdsourcing Real-Time Wizard-Of-Oz Dialogues," in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, May 2020, 288–297. Available at: https://www.aclweb.org/anthology/2020.lrec-1.36

Garrod, S., and Anderson, A. (1987). Saying what You Mean in Dialogue: A Study in Conceptual and Semantic Co-ordination. *Cognition* 27 (2), 181–218. doi:10.1016/0010-0277(87)90018-7

Geva, M., Goldberg, Y., and Berant, J. (2019). "Are We Modeling the Task or the Annotator? an Investigation of Annotator Bias in Natural Language Understanding Datasets," in EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, Hong Kong, China, November 2019 (Association for Computational Linguistics). 1161–1166. doi:10.18653/v1/d19-1107

Grosz, B. J. (2018). Smart Enough to Talk with Us? Foundations and Challenges for Dialogue Capable AI Systems. *Comput. Linguistics* 44 (1), 1–15. doi:10.1162/COLI_a_00313

Hauptmann, A. G., and Rudnicky, A. I. (1988). Talking to Computers: An Empirical Investigation. *Int. J. Man-Machine Stud.* 28 (6), 583–604. doi:10.1016/S0020-7373(88)80062-2

Henderson, M., Thomson, B., and Williams, J. (2013). Dialog State Tracking Challenge 2 & 3." Dialogues with Social Robots, No. September: 1–22. Available at: http://camdial.org/~mh521/dstc/downloads/handbook.pdf.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). SpaCy: Industrial-Strength Natural Language Processing in Python. Zenodo. doi:10.5281/zenodo.1212303

Jonell, P., Fallgren, P., Doğan, F. I., Lopes, J., Irmak Doğan, F., Wennberg, U., et al. (2019). Crowdsourcing a Self-Evolving Dialog Graph. CUI'19: Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, Ireland, August 2019. doi:10.1145/3342775.3342790

Kelley, J. F. (1983). "An Empirical Methodology for Writing User-Friendly Natural Language Computer Applications," in CHI '83: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, December 1983, 193–196.

Levelt, W. J. M., and Kelter., S. (1982). Surface Form and Memory in Question Answering. *Cogn. Psychol.* 14 (1). doi:10.1016/0010-0285(82)90005-6

Leviathan, Yaniv., and Matias, Yossi. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks over the Phone. Available at: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

McCarthy, P. M., and Jarvis., S. (2010). MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behav. Res. Methods* 42 (2), 381–392. doi:10.3758/BRM.42.2.381

McRae, K., Nedjadrasul, D., Pau, R., Bethany, P. H. L., Lo, B. P.-H., and King, L. (2018). Abstract Concepts and Pictures of Real-World Situations Activate One Another. *Top. Cogn. Sci.* 10 (3), 518–532. doi:10.1111/tops.12328

Merdivan, E., Singh, D., Hanke, S., Kropf, J., Holzinger, A., and Geist, M. (2020). Human Annotated Dialogues Dataset for Natural Conversational Agents. *Appl. Sci.* 10 (762), 762. doi:10.3390/app10030762

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). Doccano: Text Annotation Tool for Human. https://github.com/doccano/doccano.

Nass, C., and Lee, K. M. (2001). Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction. *J. Exp. Psychol. Appl.* 7 (3), 171–181. doi:10.1037//1076-898X.7.3.17110.1037/1076-898x.7.3.171Lee

Nass, C., Steuer, J., and Tauber, E. R. T. (1994). "Computers Are Social Actors," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'94) (New York, NY: Association for Computing Machinery), 72–78. doi:10.1145/191666.191703

Nishimaki, K. (2014). Characteristics of Spoken and Written Communication in the Opening and Closing Sections of Instant Messaging. Dissertations and Theses. Paper 1548. Available at: https://pdxscholar.library.pdx.edu/open_access_etds.

Oreström, B. (1983). *Turn-Taking in English Conversation*. Lund: Krieger Publishing Company.

Pareti, Silvia., and Lando, Tatiana. (2019). "Dialog Intent Structure: A Hierarchical Schema of Linked Dialog Acts," in LREC 2018 - 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, May 2018, 2907–2914. Available at: https://www.aclweb.org/anthology/L18-1460/

Pickering, M. J., and Garrod., S. (2004). Toward a Mechanistic Psychology of Dialogue. *Behav. Brain Sci.* 27 (2). doi:10.1017/s0140525x04000056

Poesio, M., and Rieser, H. (2011). An Incremental Model of Anaphora and Reference Resolution Based on Resource Situations. *dad* 2 (1), 235–277. doi:10.5087/dad.2011.110

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). "Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset," in Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, February 2020. 34, 8696–96. doi:10.1609/aaai.v34i05.6394

Reeves, B., and Nass, C. (1996). *Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Stanford, CA: The Center for the Study of Language and Information Publications. doi:10.1300/j105v24n03_14

Rieser, V. (2008). *Bootstrapping Reinforcement Learning-Based Dialogue Strategies from Wizard-Of-Oz Data*." Saarbrücken, Germany: German Research Center for Artificial Intelligence.

Rieser, V., and Lemon, O. (2011). "Reinforcement Learning for Adaptive Dialogue Systems. A Data-Driven Methodology for Dialogue Management and Natural Language Generation," in *Theory and Applications of Natural Language Processing*. Editors G Hirst, E Hovy, and M Johnson (Heidelberg, Dordrecht, London, New York: Springer). doi:10.1007/978-3-642-24942-6http://www.springer.com/series/8899

Schegloff, E. A. (1982). *"Discourse as an Interactional Achievement: Some Uses of 'uh Huh' and Other Things that Come between Sentences*. Washington, DC: Roundtable on Languages and Linguistics.

Schlangen, D. (2019). Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research. ArXiv https://arxiv.org/abs/1908.10747. doi:10.18653/v1/w19-0424

Schnoebelen, T., and Kuperman, V. (2010). Using Amazon Mechanical Turk for Linguistic Research. *Psihologija* 43 (4), 441–464. doi:10.2298/PSI1004441S

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2018). A Survey of Available Corpora for Building Data-Driven Dialogue Systems: The Journal Version. *dad* 9 (1), 1–49. doi:10.5087/dad.2018.101

Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., and Heck, L. (2018). "Building a Conversational Agent Overnight with Dialogue Self-Play," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, June 2018 (Association for Computational Linguistics), Volume 3, 41–51. (Industry Papers) https://www.aclweb.org/anthology/N18-3006.pdf.

Shatz, M., and Gelman, R. (1973). The Development of Communication Skills: Modifications in the Speech of Young Children as a Function of Listener. *Monogr. Soc. Res. Child Develop.* 38 (5), 1. doi:10.2307/1165783

Siegert, I. (2020). in Alexa in the Wild' - Collecting Unconstrained Conversations with a Modern Voice Assistant in a Public Environment." LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, Marseille, France, May 2020, 615–619. https://www.aclweb.org/anthology/2020.lrec-1.77/.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics* 9, 315–332.

Wang, W. Y., Bohus, D., Kamar, E., and Horvitz, E. (2012). "Crowdsourcing the Acquisition of Natural Language Corpora: Methods and Observations. "Crowdsourcing the Acquisition of Natural Language Corpora: Methods and

Observations.," in 2012 IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings, Miami, FL, December 2012. doi:10.1109/SLT.2012.6424200

Wei, W., Le, Q., Dai, A., and Li, J. (2018). "Airdialogue: An Environment for Goal-Oriented Dialogue Research," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, Brussels, Belgium, October–November 2018. doi:10.18653/v1/d18-1419LeLi2018

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašíc, M., Rojas Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). "A Network-Based End-To-End Trainable Task-Oriented Dialogue System," in 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, Valencia, Spain, April 2017, 1, 438–449. doi:10.18653/v1/e17-1042

Williams, J. D., and Young, S. (2007). Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Comput. Speech Lang.* 21 (2), 393–422. doi:10.1016/j.csl.2006.06.008

Xu, Y., and Reitter, D. (2018). Information Density Converges in Dialogue: Towards an Information-Theoretic Model. *Cognition* 170, 147–163. doi:10.1016/j.cognition.2017.09.018

Yuan, S., Brüggemeier, B., Hillmann, S., and Michael, T. (2020). "User Preference and Categories for Error Responses in Conversational User Interfaces," in CUI'20: Proceedings of the 2nd Conference on Conversational User Interfaces, Bilbao, Spain, July 2020. doi:10.1145/3405755.3406126

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

frontiers in Psychology

OPINION
published: 01 July 2021
doi: 10.3389/fpsyg.2021.684012

Potential and Pitfalls of Digital Voice Assistants in Older Adults With and Without Intellectual Disabilities: Relevance of Participatory Design Elements and Ecologically Valid Field Studies

Anna Schlomann[1,2], Hans-Werner Wahl[1,3], Peter Zentel[4], Vera Heyl[5], Leonore Knapp[2], Christiane Opfermann[1], Torsten Krämer[2] and Christian Rietz[2]*

[1] Network Aging Research, Heidelberg University, Heidelberg, Germany, [2] Institute for Educational Sciences, Heidelberg University of Education, Heidelberg, Germany, [3] Institute of Psychology, Heidelberg University, Heidelberg, Germany, [4] Institute of Prevention, Integration and Rehabilitation Research, Ludwig-Maximilians-Universität Munich, Munich, Germany, [5] Institute for Special Needs Education, Heidelberg University of Education, Heidelberg, Germany

OPEN ACCESS

Edited by:
Benjamin Weiss,
Hanyang University, South Korea

Reviewed by:
Maria Klara Wolters,
University of Edinburgh,
United Kingdom

Correspondence:
Anna Schlomann
schlomann@nar.uni-heidelberg.de

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

Received: *22 March 2021*
Accepted: *01 June 2021*
Published: *01 July 2021*

Citation:
Schlomann A, Wahl H-W, Zentel P, Heyl V, Knapp L, Opfermann C, Krämer T and Rietz C (2021) Potential and Pitfalls of Digital Voice Assistants in Older Adults With and Without Intellectual Disabilities: Relevance of Participatory Design Elements and Ecologically Valid Field Studies. Front. Psychol. 12:684012. doi: 10.3389/fpsyg.2021.684012

Keywords: speech assistant, smart speaker, aging, old age, digitization, ecological validity, digital diary

INTRODUCTION

While digital voice assistants (VAs) are increasingly becoming part of everyday life (Auxier, 2019), specific user groups that might benefit from such technology are often not considered in research. This applies to older adults (Stigall et al., 2019), a population with pronounced heterogeneity in terms of cognitive, sensory and functional competencies. Given the rapid expansion of digital tools, it is important to consider the use of VAs as a promising tool able to maintain and enhance social participation, autonomy, and leisure activities for a wide range of older adults. There is still a digital divide between younger and older generations (Anderson, 2019) and older adults might benefit to a smaller extent from the advantages of new technologies such as VAs. Older people with intellectual disabilities are at particular risk of digital exclusion (Ehlers et al., 2020) and digital technologies such as VAs have hardly been evaluated for this target group (e.g., Smith et al., 2020). In this opinion paper, we synthesize current research in the context of VAs for older adults. Building on this, we propose specific research designs to provide better insights into the adoption and use of VAs in advanced age.

BENEFITS OF VOICE ASSISTANTS

An overview of literature provides six clusters of insights. First, VAs offer technology access for individuals who do not use conventional computing devices. *Usability* problems caused by small font or buttons are eliminated (Ziman and Walsh, 2018; Kowalski et al., 2019; Corbett et al., 2021), which is especially useful for individuals with limited motor, sensory, or cognitive functions (e.g., Yaghoubzadeh et al., 2013; Wulf et al., 2014). Second, in the *social domain*, VAs enable contact and communication with others, especially for older people with disabilities like limited vision or impaired hand movement (Kowalski et al., 2019; Scherr et al., 2020; Trajkova and Martin-Hammond, 2020). Additionally, the VA itself can be a social companion to some extent (Scherr et al., 2020; Smith et al., 2020; Corbett et al., 2021). Third, a benefit has been identified in the *health domain*. VAs can assist with daily well-being activities such as health tracking, medication management, or dietary planning (Tsiourti et al., 2014; Nallam et al., 2020; Trajkova and Martin-Hammond, 2020). Fourth, the use of VAs might contribute to enjoyable *leisure time* experiences, via entertainment features like music, videos, and jokes (Scherr et al., 2020; Corbett et al., 2021). Fifth, VAs can provide *support for independent living* including aspects of time-structuring (e.g., setting a timer or reminders) and instrumental activities like access to online information (Nallam et al., 2020; Pradhan et al., 2020; Scherr et al., 2020; Corbett et al., 2021). Sixth, taken together, VAs may have a positive effect on a person's *agency* because VAs can support older adults with intellectual disabilities in better managing different aspects of their everyday lives (Smith et al., 2020).

CHALLENGES OF VOICE ASSISTANTS AND LIMITATIONS IN CURRENT RESEARCH ON VOICE ASSISTANTS IN OLDER ADULTS

Although it boasts good usability in general, *problems interacting* with VAs are frequently observed, because the user is required to follow a pre-structured form of dialogue, thus limiting the conversational abilities of VAs (Scherr et al., 2020). Older adults often have problems recalling the specific commands necessary to operate the devices (Wulf et al., 2014; Pradhan et al., 2020). Another limiting factor can be the *lack of added value*, which may result in a preference for devices already used (Trajkova and Martin-Hammond, 2020). VAs are perceived as time consuming and a *lack of compatibility* is criticized (Kowalski et al., 2019). Furthermore, a barrier for using VAs can be seen in reported fear of *losing one's own competences and autonomy*, because the VA may take care of a number of tasks without considering the competencies of the user (Kowalski et al., 2019; Trajkova and Martin-Hammond, 2020). In this way, the benefit of support for independent living and a reduction of dependency on personal assistance may also result in a *higher level of dependency* on VA assistance. Finally, concerns about *privacy and data security* are potential barriers to using VAs (Nallam et al., 2020; Trajkova and Martin-Hammond, 2020).

On a more general level, two major limitations in research on VAs for older adults exist. There is limited knowledge on VA use in *specific groups of older adults*. We identified two studies that focus on benefits and challenges for older adults with cognitive impairment (e.g., dementia; Wargnier et al., 2015; Wolters et al., 2016) and two other studies that included older people with intellectual disabilities (Braun et al., 2020; Smith et al., 2020). We also see *limitations at the methodological and design level*. In particular, we identified six studies that apply field data collection—collecting data in everyday life settings to analyze the use of VAs among older adults (Tsiourti et al., 2014; Kopp et al., 2018; Oh et al., 2020; Pradhan et al., 2020; Scherr et al., 2020; Corbett et al., 2021). Similarly, user-centered research approaches have been infrequently applied thus far.

RECOMMENDATIONS FOR FUTURE RESEARCH

Based on these insights, we derive recommendations for research on VAs (see **Table 1**). We put emphasis on subgroups of older adults who have been largely neglected, i.e., older adults with cognitive impairment like dementia-related disorders or intellectual disabilities.

Need for Participatory Design Elements in Research With Older Adults

We recommend a high level of user involvement in research about commercial VAs for older people. A participatory design strategy should identify the best ways to introduce VA use, benefits as well as limits and risks of VAs to different groups of older adults. In particular, this allows to address the identified challenges (see above) and may ensure that older people can use VAs according to their preferences. We argue that the need for instruction is higher among older adults compared to younger people because they did not grow up with digital technologies. As an example, we propose a participatory conception and implementation of user trainings and manuals[1]: To realize a participatory design in this domain, older adults should be involved in different co-designing activities such as co-design workshops to develop and discuss own ideas (e.g., Davidson and Jensen, 2013). Thereby, older users will be able to directly and actively influence the design of different kinds of material:

- *User trainings:* Educational programs and training courses can lower barriers in VA interaction (Czaja et al., 2019). The contents and format should be developed together with the target group and tailored to varying prior technology experiences, disability, and cognitive competencies. Relevant aspects may include the requirements for a successful interaction with VAs (e.g., specific voice commands), the possibilities of using VAs for different purposes (e.g., social domain, health, leisure time), and possible concerns (e.g., privacy issues, losing competencies due to using VAs). We recommend including older users with different skill levels in the training conception to make VAs widely accessible. Due to the different competencies (e.g., in reading and writing, attentional control, executive functions), the presentation and complexity must be adapted individually in each case.
- *User manuals:* Another aspect of VA learning and adoption are user manuals, an aspect that has so far been insufficiently addressed in research. Well-designed guidelines could help older adults to explore the possibilities of VAs according to their needs and in their own pace. Different versions of user manuals should be discussed with older adults to achieve the best design possible. These group-specific manuals are especially helpful for older adults with cognitive impairments

[1]We put emphasis on commercial VAs that are available "off the shelf" but offer lower customization and co-designing options. However, in the case of *VA design and development*, participatory research and co-designing is equally important to better understand users' needs and their preferences. A co-design of VAs will guarantee that style, content, wording and tasks of the VAs will fit the mental models of older adults (Wolters et al., 2016; Fischer et al., 2020).

TABLE 1 | Recommendations for future research on Voice Assistants (VAs) in heterogeneous groups of older adults.

	Participatory design elements	Field studies
Aims	• Conception and implementation of user training and manuals for using VAs together with the target group(s)	• Collect ecologically valid information on user experience • Gain insights into real-life use (and non-use) of VAs
Methods and design	• Co-designing, e.g., design workshops, prototype evaluation • Semi-structured interviews • Focus group discussions	• Diary studies, including cultural probes • Automatically collected data on usage (i.e., back-end data, audio- or video recordings) • Emotion analysis based on video and audio data
Expected outcomes and synergies	• Introduce VA use, benefits as well as limits and risks of VAs to different groups of older adults • Address common challenges of VAs for users with different levels of competences	• Get insights into user experiences beyond verbal and subjective feedback • Assess potentials of VAs for different groups • Improve user training and manuals

or intellectual disabilities who may have specific needs, like instructions in easy-to-read language (Maaß, 2020) or a visualization of the instructions (Spriggs et al., 2017).

Need for Field Studies in the Everyday Lives of Older Adults

Field studies can contribute to a high level of *ecological validity* of the research–implying the empirical validity of findings in everyday life settings. Depending on the research questions, researchers should thoroughly consider the appropriate observation period. Existing field studies about VA use report durations from only a few days (Lopatovska et al., 2019), several months (Scherr et al., 2020), up to 1 year (Trajkova and Martin-Hammond, 2020). Our recommendation would be an observation period of at least 4 weeks, which would allow analysis of more routinized interactions with VAs after an initial phase of learning and curiosity.

Despite existing field studies, we claim that the potentials are not yet fully exploited, and should be extended focusing on innovative approaches and heterogeneous user groups:

- *Diary studies*: In open-ended and closed questions, the users can report their experiences, likes, and dislikes about using a VA. A digital diary format allows to provide additional assistance if necessary. Participants can be reminded of the diary with prompts, and questions can be repeated and adapted to the individual. Participants can be actively encouraged to provide comprehensive feedback on enjoyable, useful, and negative experiences of VA interactions. In addition to this, cultural probes such as self-taken photos, cards with reflection tasks about VA use, and other activities (e.g., creation of relationship or neighborhood maps) can be applied to gain further insights into the everyday lives of the participants and to capture older adults' experiences with the devices in a comprehensive manner (Jarke and Gerhard, 2017).
- *Analysis of automatically collected data*: Beyond users' self-reports, we see high potential in collecting additional data associated with usage behaviors like back-end data combined with external recordings of audio (Porcheron et al., 2018) or video (Lahoual and Frejus, 2019) data of VA use. These data provide information about which VA functions are used by individuals with different levels of expertise and competences, about used voice commands, and changes of use patterns over time. In this context, researchers should in any case consider ethical concerns of automatically collected data such as a threat of permanent observation or the fear of providing too intimate information. Attention should be paid to the design of the informed consent and continuous support of study participants concerning these aspects should be provided.
- *Emotion analysis:* State-of-the-art software solutions allow to automatically analyze emotional experiences based on speech and facial expressions (Garcia-Garcia et al., 2017; Dupré et al., 2018). The emotional experience of VA use is another aspect that is important to our understanding of their benefits and challenges for older adults. An analysis of these data allows researchers to study user experiences of VAs in situations when verbal feedback is scarce due to possible cognitive impairments or intellectual disabilities. Still, due to the novelty of the approach, automatic emotion analysis is not always reliable. Regarding older adults with intellectual disabilities, their emotional expression may differ from older adults without disabilities due to more frequent motor impairments (e.g., spasticity), differences in physical appearance (especially regarding genetic syndromes), cognitive deficits (e.g., in perception or appraisal processes), stereotypical behaviors, or earlier aging processes (von Gontard, 2013). Facial expression may be altered so that automatic face recognition cannot detect known patterns. The validity of automatic emotion analysis in this group has not yet been proven (Adams and Oliver, 2011; Martínez-González and Veas, 2019) and should be the focus of future research.

In particular, the triangulation of the different data sources can provide an overarching picture of user experiences of VAs, e.g., by analyzing in which situations the VA is used, how the older person evaluates this interaction, and how it is experienced emotionally.

CONCLUSIONS

In this opinion paper, we offer a set of recommendations that may guide future research on VAs for and with older adults. In a nutshell, we posit that future research designs should strongly rely on analyzing VA interactions in everyday ecologies and strictly apply participatory design elements where possible. Data protocols should include a balanced mixture of automatized data including emotional aspects as well as structured and open assessments. From our perspective, considering these recommendations can significantly help to create evidence-based findings able to inform interventions with VAs in heterogeneous groups of older adults. This also contributes to the goal of getting the best out of VA systems to improve quality of life and avoid possible risks.

AUTHOR CONTRIBUTIONS

All authors provided substantial contributions to this article from conception to final approval.

FUNDING

This publication was part of the *AI-Aging* project (AI-based voice assistance for older adults with and without intellectual disabilities), which was funded by the Baden-Württemberg Stiftung within the funding line Responsible Artificial Intelligence. The article processing charge was funded by the Baden-Württemberg Ministry of Science, Research and Culture and the Heidelberg University of Education in the funding programme Open Access Publishing.

REFERENCES

Adams, D., and Oliver, C. (2011). The expression and assessment of emotions and internal states in individuals with severe or profound intellectual disabilities. *Clin. Psychol. Rev.* 31, 293–306. doi: 10.1016/j.cpr.2011.01.003

Anderson, M. (2019). *Mobile Technology and Home Broadband 2019*. Available online at: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/06/PI_2019.06.13_Mobile-Technology-and-Home-Broadband_FINAL2.pdf (accessed May 17, 2021).

Auxier, B. (2019). *5 Things to Know About Americans and Their Smart Speakers*. Available online at: https://www.pewresearch.org/fact-tank/2019/11/21/5-things-to-know-about-americans-and-their-smart-speakers/ (accessed May 20, 2021).

Braun, M., Wölfel, M., Renner, G., and Menschik, C. (2020). "Accessibility of different natural user interfaces for people with intellectual disabilities," in *2020 International Conference on Cyberworlds (CW)* (Caen: IEEE), 211–218.

Corbett, C. F., Combs, E. M., Wright, P. J., Owens, O. L., Stringfellow, I., Nguyen, T., et al. (2021). Virtual home assistant use and perceptions of usefulness by older adults and support person dyads. *Int. J. Environ. Res. Public Health* 18, 1–13. doi: 10.3390/ijerph18031113

Czaja, S. J., Boot, W., Charness, N., and Rogers, W. A. (2019). *Designing for Older Adults: Principles and Creative Human Factors Approaches, 3rd Edn*. London: Chapman and Hall/CRC.

Davidson, J. L., and Jensen, C. (2013). "Participatory design with older adults," in *Proceedings of the 9th ACM Conference on Creativity and Cognition*, eds E. Y.-L. Do, S. Dow, J. Ox, S. Smith, K. Nishimoto, C. T. Tan, Y. Nagai, S. Ferguson, and T. Hewett (Sydney, NSW: ACM), 114–123.

Dupré, D., Andelic, N., Morrison, G., and McKeown, G. (2018). "Accuracy of three commercial automatic emotion recognition systems across different individuals and their facial expressions," in *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (Athens: IEEE), 627–632.

Ehlers, A., Heß, M., Frewer-Graumann, S., Olbermann, E., and Stiemke, P. (2020). "Digitale Teilhabe und (digitale) Exklusion im Alter: Expertise zum Achten Altersbericht der Bundesregierung," in *Expertisen zum Achten Altersbericht der Bundesregierung*, eds C. Hagen, C. Endter, and F. Berner (Berlin: Deutsches Zentrum für Altersfragen).

Fischer, B., Peine, A., and Östlund, B. (2020). The importance of user involvement: a systematic review of involving older users in technology design. *Gerontologist* 60, e513–e523. doi: 10.1093/geront/gnz163

Garcia-Garcia, J. M., Penichet, V. M. R., and Lozano, M. D. (2017). "Emotion detection: technology review," in *Proceedings of the XVIII International Conference on Human Computer Interaction–Interacción '17*, ed J. M. González-Calleros (Cancun: ACM Press), 1–8. doi: 10.1145/3123818.3123852

Jarke, J., and Gerhard, U. (2017). "Using cultural probes for co-creating a digital neighbourhood guide with and for older adults," in *Mensch und Computer–Workshopband*, eds M. Burghardt, R. Wimmer, C. Wolff, and C. Womser-Hacker (Regensburg), 79–85.

Kopp, S., Brandt, M., Buschmeier, H., Cyra, K., Freigang, F., Krämer, N., et al. (2018). "Conversational assistants for elderly users–the importance of socially cooperative dialogue," in *Proceedings of the AAMAS 2018 Workshop on Intelligent Conversational Agents in Home and Geriatric Care Applications* (Stockholm).

Kowalski, J., Jaskulska, A., Skorupska, K., Abramczuk, K., Biele, C., Kopeć, W., et al. (2019). "Older adults and voice interaction: a pilot study with google home," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, eds S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos (Glasgow: ACM), 1–6.

Lahoual, D., and Frejus, M. (2019). "When users assist the voice assistants: from supervision to failure resolution," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, eds S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos (Glasgow: ACM), 1–8.

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

Maaß, C. (2020). *Easy Language—Plain Language—Easy Language Plus: Balancing Comprehensibility and Acceptability*. Berlin: Frank and Timme.

Martínez-González, A. E., and Veas, A. (2019). Identification of emotions and physiological response in individuals with moderate intellectual disability. *Int. J. Dev. Disabil.* 1–6. doi: 10.1080/20473869.2019.1651142

Nallam, P., Bhandari, S., Sanders, J., and Martin-Hammond, A. (2020). A question of access: exploring the perceived benefits and barriers of intelligent voice assistants for improving access to consumer health resources among low-income older adults. *Gerontol. Geriatr. Med.* 6:2333721420985975. doi: 10.1177/2333721420985975

Oh, Y. H., Chung, K., and Ju, D. Y. (2020). Differences in interactions with a conversational agent. *Int. J. Environ. Res. Public Health* 17:3189. doi: 10.3390/ijerph17093189

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). "Voice interfaces in everyday life," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems–CHI '18*, eds R. Mandryk, M. Hancock, M. Perry, and A. Cox (Montreal, QC: ACM Press), 1–12.

Pradhan, A., Lazar, A., and Findlater, L. (2020). Use of intelligent voice assistants by older adults with low technology use. *ACM Trans. Comput. Hum. Interact.* 27, 1–27. doi: 10.1145/3373759

Scherr, S. A., Meier, A., and Cihan, S. (2020). "Alexa, tell me more – about new best friends, the advantage of hands-free operation and life-long learning," in *Mensch und Computer 2020–Workshopband*, eds C. Hansen, A. Nürnberger, and B. Preim (Bonn: Gesellschaft für Informatik).

Smith, E., Sumner, P., Hedge, C., and Powell, G. (2020). Smart-speaker technology and intellectual disabilities: agency and wellbeing. *Disabil. Rehabil. Assist. Technol.* 1–11. doi: 10.1080/17483107.2020.1864670

Spriggs, A. D., Mims, P. J., van Dijk, W., and Knight, V. F. (2017). Examination of the evidence base for using visual activity schedules with students with intellectual disability. *J. Spec. Educ.* 51, 14–26. doi: 10.1177/0022466916658483

Stigall, B., Waycott, J., Baker, S., and Caine, K. (2019). "Older adults' perception and use of voice user interfaces: a preliminary review of the computing literature," in *OZCHI'19*, eds S. Berkovsky, Y. Hijikata, J. Rekimoto, M. Burnett, M. Billinghurst, and A. Quigley (Fremantle, WA: ACM Press), 55–66.

Trajkova, M., and Martin-Hammond, A. (2020). ""Alexa is a Toy": exploring older adults' reasons for using, limiting, and abandoning echo," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, eds R. Bernhaupt, F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, and R. Kocielnik (Honolulu, HI: ACM), 1–13.

Tsiourti, C., Joly, E., Wings, C., Ben Moussa, M., and Wac, K. (2014). "Virtual assistive companions for older adults: qualitative field study and design implications," in *PervasiveHealth '14: proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, Oldenburg, Germany, May 20-23, 2014*, eds A. Hein and S. Boll (Oldenburg: ICST, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 57–64.

von Gontard, A. (2013). "Genetische und biologische Grundlagen." in *Geistige Behinderung: Grundlagen, Erscheinungsformen und klinische Probleme, Behandlung, Rehabilitation und rechtliche Aspekte*, eds G. Neuhäuser, H.-C. Steinhausen, F. Häßler, K. Sarimski, A. von Gontard, J. Buchmann, O. Reis, T. Klauß, S. Wachsmuth, K. Fischer, G. Grampp, S. Elstner, M. Buchka, C. Schanze, T. Frühauf, S. Wendt, U. Hellmann, and K. Lachwitz (Stuttgart: Kohlhammer Verlag), 30–43.

Wargnier, P., Malaise, A., Jacquemot, J., Benveniste, S., Jouvelot, P., Pino, M., et al. (2015). "Towards attention monitoring of older adults with cognitive impairment during interaction with an embodied conversational agent," in *3rd IEEE VR International Workshop on Virtual and Augmented Assistive Technology (VAAT)* (Arles: IEEE), 23–28.

Wolters, M. K., Kelly, F., and Kilgour, J. (2016). Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. *Health Inform. J.* 22, 854–866. doi: 10.1177/1460458215593329

Wulf, L., Garschall, M., Himmelsbach, J., and Tscheligi, M. (2014). "Hands free–care free: elderly people taking advantage of speech-only interaction," in *Proceedings of the 8th Nordic Conference on Human-Computer Interaction Fun, Fast, Foundational-NordiCHI'14*, eds V. Roto, T. Olsson, K. Väänänen-Vainio-Mattila, J. Häkkilä, E. Hvannberg, and O. Juhlin (Helsinki: ACM Press), 203–206.

Yaghoubzadeh, R., Kramer, M., Pitsch, K., and Kopp, S. (2013). "Virtual agents as daily assistants for elderly or cognitively impaired people: Studies on acceptance and interaction feasibility," in *Intelligent Virtual Agents. IVA 2013. Lecture Notes in Computer Science*, eds R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira (Berlin, Heidelberg: Springer). doi: 10.1007/978-3-642-40415-3_7

Ziman, R., and Walsh, G. (2018). "Factors affecting seniors' perceptions of voice-enabled user interfaces." in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, eds R. Mandryk, M. Hancock, M. Perry, and A. Cox (Montreal, QC: ACM), 1–6.

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

frontiers in Communication

ORIGINAL RESEARCH
published: 29 July 2021
doi: 10.3389/fcomm.2021.675704

Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments

Michelle Cohn* **and Georgia Zellou**

Phonetics Lab, University of California, Davis, CA, United States

The current study tests whether individuals (n = 53) produce distinct speech adaptations during pre-scripted spoken interactions with a voice-AI assistant (Amazon's Alexa) relative to those with a human interlocutor. Interactions crossed intelligibility pressures (staged word misrecognitions) and emotionality (hyper-expressive interjections) as conversation-internal factors that might influence participants' intelligibility adjustments in Alexa- and human-directed speech (DS). Overall, we find speech style differences: Alexa-DS has a decreased speech rate, higher mean f0, and greater f0 variation than human-DS. In speech produced toward both interlocutors, adjustments in response to misrecognition were similar: participants produced more distinct vowel backing (enhancing the contrast between the target word and misrecognition) in target words and louder, slower, higher mean f0, and higher f0 variation at the sentence-level. No differences were observed in human- and Alexa-DS following displays of emotional expressiveness by the interlocutors. Expressiveness, furthermore, did not mediate intelligibility adjustments in response to a misrecognition. Taken together, these findings support proposals that speakers presume voice-AI has a "communicative barrier" (relative to human interlocutors), but that speakers adapt to conversational-internal factors of intelligibility similarly in human- and Alexa-DS. This work contributes to our understanding of human-computer interaction, as well as theories of speech style adaptation.

Keywords: voice-activated artificially intelligent (voice-AI) assistant, speech register, human-computer interaction, computer personification, speech intelligibility

OPEN ACCESS

Edited by:
Ingo Siegert,
Otto von Guericke University Magdeburg, Germany

Reviewed by:
Oliver Niebuhr,
University of Southern Denmark, Denmark
Hendrik Buschmeier,
Bielefeld University, Germany
Volker Dellwo,
University of Zurich, Switzerland

Correspondence:
Michelle Cohn
mdcohn@ucdavis.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *03 March 2021*
Accepted: *21 June 2021*
Published: *29 July 2021*

Citation:
Cohn M and Zellou G (2021) Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments. Front. Commun. 6:675704. doi: 10.3389/fcomm.2021.675704

INTRODUCTION

People dynamically adapt their speech according to the communicative context and (apparent) barriers present. In the presence of background noise, for example, speakers produce speech that is louder, slower, and higher pitched ("Lombard speech") (for a review, see Brumm and Zollinger, 2011), argued by some to be an automatic, non-socially mediated response (Junqua, 1993; Junqua, 1996). Other work has shown that people adapt their speech to the type of *listener* they are engaging with. One stance is that speakers presume certain types of interlocutors to have greater communicative barriers (Clark and Murphy, 1982; Clark, 1996; Oviatt et al., 1998b; Branigan et al., 2011). Supporting this account, prior work has shown that people use different speech styles when talking to non-native speakers (Scarborough et al., 2007; Uther et al., 2007; Hazan et al., 2015), hearing impaired adults (Picheny et al., 1985; Scarborough and Zellou, 2013; Knoll et al., 2015), and computers (Oviatt et al., 1998a; Oviatt et al., 1998b; Bell and Gustafson, 1999; Bell et al., 2003; Lunsford et al., 2006; Stent et al., 2008; Burnham et al., 2010; Mayo et al., 2012; Siegert et al., 2019).

For example, computer-directed speech (DS) has been shown to be louder (Lunsford et al., 2006), with durational lengthening (Burnham et al., 2010; Mayo et al., 2012), greater vowel space expansion (Burnham et al., 2010), and smaller pitch range (Mayo et al., 2012) than speech directed to a (normal hearing adult) human.

This paper explores whether speakers use a specific speech style (or "register") when talking to a voice-activated artificially intelligent (voice-AI) assistant. Voice-AI assistants (e.g., Amazon's Alexa, Apple's Siri, Google Assistant) are now a common interlocutor for millions of individuals completing everyday tasks (e.g., "set a timer for 5 min", "turn on the lights", etc.) (Bentley et al., 2018; Ammari et al., 2019). A growing body of research has begun to investigate the social, cognitive, and linguistic effects of humans interacting with voice-AI (Purington et al., 2017; Arnold et al., 2019; Cohn et al., 2019b; Burbach et al., 2019). For example, recent work has shown that listeners attribute human-like characteristics to the text-to-speech (TTS) output used for modern voice-AI, including personality traits (Lopatovska, 2020), apparent age (Cohn et al., 2020a; Zellou et al., 2021), and gender (Habler et al., 2019; Loideain and Adams, 2020). While the spread of voice-AI assistants is undeniable—particularly in the United States—there are many open scientific questions as to the nature of people's interactions with voice-AI.

There is some evidence for a different speech style used in interactions with voice-AI assistants: several studies have used classifiers to successfully identify "device-" and "non-device-directed" speech from users' interactions with Amazon Alexa (Mallidi et al., 2018; Huang et al., 2019). Yet, in these cases, the linguistic content, physical distance from the device, and other factors were not controlled and might have contributed to differences that are not speech-style adaptations per se. Critically, holding the interaction constant across a voice-AI and human interlocutor can reveal if individuals have a distinct voice-AI speech style. Some groups have aimed to compare human and voice-AI speech styles in more similar contexts. For instance, the Voice Assistant Conversation Corpus (VACC) had participants complete the same type of communicative task (setting an appointment on a calendar and doing a quiz) with an Alexa Echo and a real human confederate (Siegert et al., 2018). Several studies measuring the acoustic-phonetic features of human- and Alexa-DS in the corpus found productions toward Alexa were louder (Raveh et al., 2019; Siegert and Krüger, 2021), higher in fundamental frequency (f0, perceived pitch) (Raveh et al., 2019), and contained different vowel formant characteristics[1] (Siegert and Krüger, 2021). Yet, similar to studies of individuals using Alexa in their homes (e.g., Huang et al., 2019), differences observed in the VACC might also be driven by physical distance from the device and conversational variations. The current study holds context and physical distance from the microphone constant for the two interlocutors to address these limitations in prior work.

Making a direct human- and Alexa-DS comparison in a scripted task can speak to competing predictions across different computer personification accounts: if speech styles differ because speakers have a "routinized" way of talking to computers (in line with *routinized interaction accounts*) or if speech styles are the same (in line with *technology equivalence accounts*). *Routinized interaction accounts* propose that people have a "routinized" way of interacting with technological systems (Gambino et al., 2020), borne out of real experience with the systems, as well as a priori expectations. As mentioned, there is ample evidence for a computer-DS register (e.g., Bell and Gustafson, 1999; Bell et al., 2003; Burnham et al., 2010). Specifically, some propose that the computer faces additional communicative barriers, relative to humans (Oviatt et al., 1998b; Branigan et al., 2011). These attitudes appear to be a priori, developed before any evidence of communicative barriers in an interaction. For example, people rate TTS voices as less "communicatively competent" (Cowan et al., 2015). Therefore, one prediction for the current study is that speakers might have overall different speech styles in human- and Alexa-DS, reflecting this presumed communicative barrier and a "routinized" way of talking to voice-AI.

Technology equivalence accounts, on the other hand, propose that people automatically and subconsciously apply social behaviors from human-human interaction to their interactions with computer systems (e.g., Lee, 2008). For example, "Computers are Social Actors" (CASA) (Nass et al., 1994; Nass et al., 1997) specifies that this transfer of behaviors from human-human interaction is triggered when people detect a "cue" of humanity in the system, such as engaging with a system using language. For example, people appear to apply politeness norms from human-human interaction to computers: giving more favorable ratings when a computer directly asks about its own performance, relative to when a different computer elicits this information (Nass et al., 1994; Hoffmann et al., 2009). In line with *technology equivalence accounts*, there is some evidence for applied social behaviors to voice-AI in the way people adjust their speech, such as gender-mediated vocal alignment (Cohn et al., 2019b; Zellou et al., 2021). In the present study, one prediction from *technology equivalence accounts* is that people will adjust their speech patterns when talking to voice-AI and humans in similar ways if the communicative context is controlled.

Different Strategies to Improve Intelligibility Following a Misrecognition?

To probe *routinized interaction* and *technology equivalence* accounts, the present study further investigates if speakers adapt their speech differently after a human or a voice-AI assistant "mishears" them. There is evidence that speakers monitor communicative pressures during an interaction, varying their acoustic-phonetic output to improve intelligibility when there is evidence listeners might mishear them (Smiljanić and Bradlow, 2009; Hazan and Baker, 2011). Lindblom's (1990) Hyper- and Hypo-articulation (H&H) model proposes a real-time trade-off between speakers' needs (i.e., to preserve articulatory effort) and listeners' needs (i.e., to be more

[1]They do not report a directionality of difference.

intelligible). While the majority of prior work examining speakers' adaptations following a computer misrecognition has lacked a direct human comparison, many of the adjustments parallel those observed in human-human interaction; for example, speakers produce louder and slower speech after a dialogue system conveys that it "heard" the wrong word (Oviatt et al., 1998a; Bell and Gustafson, 1999; Swerts et al., 2000). Additionally, some studies report vowel adaptations in response to a misunderstanding that are consistent with enhancements to improve intelligibility, including vowel space expansion (Bell and Gustafson, 1999; Maniwa et al., 2009) and increase in formant frequencies (Vertanen, 2006). There is also evidence of targeted adjustments: speakers produce more vowel-specific expansion (e.g., high vowels produced higher) in response to misrecognitions by a dialogue system (Stent et al., 2008). Will speakers use different strategies to improve intelligibility following a staged word misrecognition based on who their listener is? One possibility is that speakers might have a "routinized" way of improving their intelligibility following a misrecognition made by a voice-AI assistant, which would support *routinized interaction accounts*. At the same time, Burnham et al. (2010) found no difference between speech adjustments post-misrecognition for an (apparent) human and digital avatar, but only more global differences for the computer interlocutor (i.e., speech with longer segmental durations and with greater vowel space expansion). Therefore, it is possible that speakers will produce similar intelligibility adjustments in response to a staged misrecognition made by either a voice-AI or human listener, supporting *technology equivalence accounts*.

Additionally, the current study adds a novel manipulation in addition to intelligibility pressures: emotional expressiveness. When an interlocutor "mishears", they might be disappointed and express it (e.g., "Darn! I think I misunderstood."); when they get it correct, they might be enthusiastic and convey that in their turn (e.g., "Awesome! I think I heard boot."). Emotional expressiveness is a common component of naturalistic human conversations, providing a window into how the listener is feeling (Goffman, 1981; Ameka, 1992). This "socio-communicative enhancement" might increase the pressure for speakers to adapt their speech for the listener. On the one hand, this enhanced emotional expressiveness might result in even more similar adjustments for voice-AI and human interlocutors, since adding expressiveness might increase the perception of human-likeness for the device, which could strengthen *technology equivalence*. Indeed, there is some work to suggest that emotional expressiveness in a computer system is perceived favorably by users. For instance, Brave and colleagues (2005) found when computer systems expressed empathetic emotion, they were rated more positively. For voice-AI, there is a growing body of work testing how individuals perceive emotion in TTS voices (Cohn et al., 2019a; Cohn et al., 2020a). For example, an Amazon Alexa Prize socialbot was rated more positively if it used emotional interjections (Cohn et al., 2019a). Alternatively, the presence of emotionality might lead to distinct clear speech strategies for the human and voice-AI interlocutors. For example, a study of phonetic alignment (using the same corpus in the current study) found that vowel duration alignment differed both by the social category of interlocutor (human vs. voice-AI) and based on emotionality (Zellou and Cohn, 2020): participants aligned more in response to a misrecognition, consistent with H&H theory (Lindblom, 1990), which increased even more when the voice-AI talker was emotionally expressive when conveying their misunderstanding (e.g., "Bummer! I'm not sure I understood. I think I heard sock or sack."). Still, that study examined just one acoustic difference in speech behavior (vowel duration alignment). The present study investigates whether emotionality similarly mediates targeted speech adjustments to voice-AI, an underexplored research question.

TABLE 1 | Target words and their (minimal pairs) used in the experiment dialogue.

Bat (boat)	Boot (beat)	Cheek (choke)	Coat (kate)
Cot (cat)	Deed (dude)	Dune (dean)	Hoop (heap)
Moat (meet)	Pod (pad)	Soap (seep)	Sock (sack)
Tap (top)	Toot (teat)	Tot (tat)	Weave (wove)

Current Study

The present study examines a corpus of speech directed at a human and voice-AI interlocutor which crossed intelligibility factors (staged misrecognitions) and emotionality of the interlocutor's responses in identical pre-scripted tasks (Zellou and Cohn, 2020). This is the first study, to our knowledge, to test both intelligibility and emotional expressiveness factors in speech style adaptations for a voice-AI assistant and human. Here, the Amazon Alexa voice (US-English, female) was selected for its ability to generate emotionally expressive phrases recorded by the voice actor, common in Alexa Skills Kit apps ("Speechcons"). To determine overall differences between Alexa- and human-DS, as well as more local intelligibility adjustments in response to a staged misrecognition, we measure several acoustic features associated with computer-DS and/or "clear" speech: intensity, speech rate, mean f0, f0 variation, and vowel formant characteristics (F1, F2).

METHODS

Participants

Data were taken from a corpus (Zellou and Cohn, 2020) containing 53 native English speaking participants (27 female, 26 male; mean age of 20.28 years old, sd = 2.42 years; range: 18–34) talking to a voice-AI and human interlocutor in an identical interactive task. None reported having any hearing impairment. Nearly all participants (n = 49) reported using a voice-AI system: Alexa (n = 35), Siri (n = 13), Google Assistant (n = 1). Participants were recruited from the UC Davis psychology subjects pool and completed informed consent, in pursuance with the UC Davis Institutional Review Board (IRB).

Target Words

Sixteen target words, presented in **Table 1**, were selected from Babel (2012) who had chosen the items for being low frequency in American English; higher frequency items have been shown to be more phonetically reduced in production (e.g., Pluymaekers et al., 2005). Target words were all CVC words containing either /i, æ, u, ow, a/ and a word-final obstruent

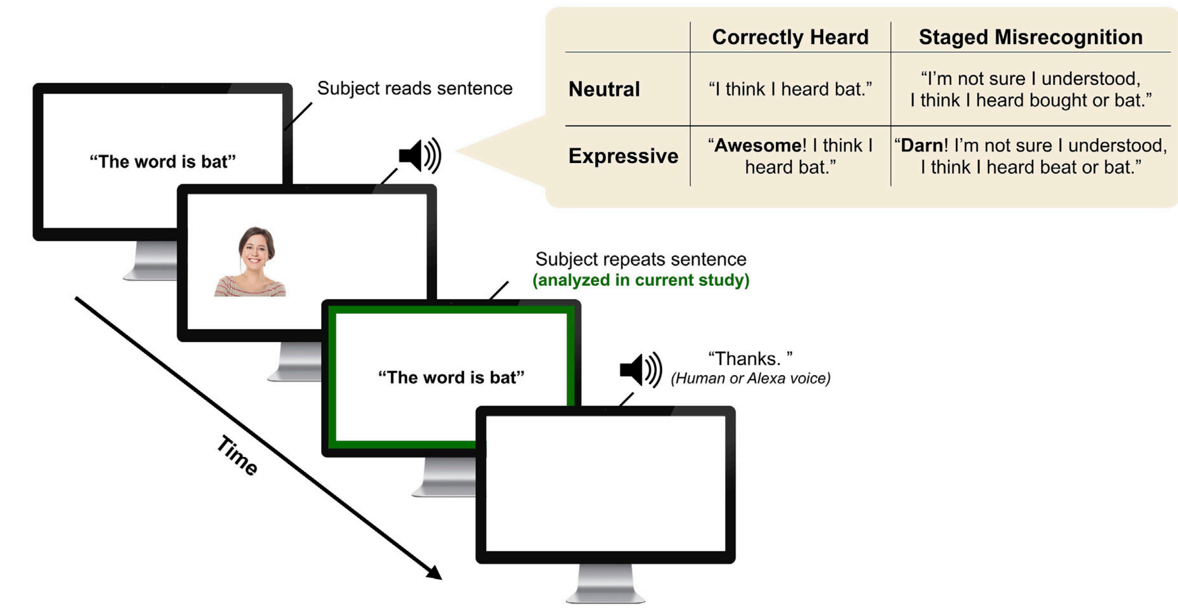


FIGURE 1 | Interaction trial schematic. After participants read a sentence, the interlocutor (human or Alexa) responds in one of the Staged Misunderstanding Conditions (correctly heard, misrecognition) and Emotionality Conditions (neutral, emotionally expressive). Then, the subject responds (the production we analyze). Finally, the interlocutor provides a follow-up response.

(e.g., /z/, /p/) (a subset of the words used in Babel, 2012). In addition, we selected a real-word vowel minimal pair, differing in vowel backness, to be used in the interlocutor responses in the misrecognition condition.

Interlocutor Recordings

The human and voice-AI interlocutor responses were pre-recorded. For the human, a female native California English speaker recorded responses in a sound attenuated booth, with a head-mounted microphone (Shure WH20 XLR). The Alexa productions were generated with the default female Alexa voice (US-English) with the Alexa Skills Kit. Both interlocutors generated introductions ("Hi! I'm Melissa. I'm a research assistant in the Phonetics Lab."/"Hi! I'm Alexa. I'm a digital device through Amazon.") and voice-over instructions for the task. We recorded each interlocutor producing two responses for each target word: a "correctly understood" response ("I think I heard bat") and an "misrecognition" response ("I'm not sure I understood. I think I heard bought or bat."). **Figure 1** provides an example of the different interlocutor responses. Order of target word and misheard word was counterbalanced across sentences, such that the "correct" word did not always occur in the same position in these response types.

Both interlocutors generated 16 emotionally expressive interjections as well: eight positive interjections (*bam, bingo, kapow, wahoo, zing, awesome, dynamite, yipee*) and eight negative interjections (*argh, baa, blarg, oof, darn, boo, oy, ouch*) selected from the Speechcons website[2] at the time of the study. We generated these interjections for the Alexa text-to-speech (TTS) output using synthesis markup language (SSML) tags. The human produced these interjections in an expressive manner (independently, not imitating the Alexa productions). We randomly assigned each interjection to the interlocutor responses, matching in whether the response was correctly understood (positive interjection) or misunderstood (negative interjection). The full set of interjections was used twice in each block (e.g., eight positive interjections randomly concatenated to 16 correct productions). The full set of interlocutor productions are available on Open Science Framework[3].

Procedure

Participants completed the experiment while seated in a sound-attenuated booth, wearing a head-mounted microphone (Shure WH20 XLR) and headphones (Sennheiser Pro), and facing a computer screen. First, we collected citation forms of the target words produced in sentences. Participants read the word in a sentence ("The word is bat.") presented on the screen. Target words were presented randomly.

Following the Citation block, participants completed identical experimental blocks with both a human talker and an Alexa talker (block order counterbalanced across subjects). First, the interlocutor introduced themselves and then went through voice-over instructions with the participant. Participants saw an image corresponding to the interlocutor category: stock images of "adult female" (used in prior work; Zellou et al., 2021) and "Amazon Alexa" (2nd Generation Black Echo).

[2]https://developer.amazon.com/en-US/docs/alexa/custom-skills/speechcon-reference-interjections-english-us.html

[3]doi: 10.17605/OSF.IO/3Y59M

Each trial consisted of four turns. Participants first read a sentence aloud containing the target word sentence-finally (e.g., "The word is bat."). Then, the interlocutor responded in one of four possible Staged Misunderstanding (correctly heard/ misrecognition) and Emotionality (neutral/expressive) Conditions (see **Figure 1**). Next, the participant responded to the interlocutor by repeating the sentence (e.g., "The word is bat."). This is the response that we acoustically analyze. Finally, the interlocutor provides a confirmation, randomized ("Thanks", "Perfect", "Okay", "Uh huh", "Got it", etc.).

In 50% of trials, the interlocutor (human, Alexa) "misunderstood" the speakers, while in the other 50% they heard correctly. Additionally, in 50% of trials, the interlocutor responded with an expressive production (distributed equally across correctly heard and misrecognition trials). Order of target words was randomized, as well as trial correspondence to the Misunderstanding and Emotionality Conditions. In each block, participants produced all target sentences once for all conditions for a total of 128 trials for each interlocutor (16 words x two misunderstanding conditions x two emotionality conditions). Participants completed the task with both interlocutors (256 total target sentences). After the speech production experiment ended (and while still in the soundbooth), participants used a sliding scale (0–100) to rate how human-like each interlocutor sounded (order of interlocutor was randomized) ("How much like a real person did [Alexa/Human] sound?" (0 = not like a real person, 100 = extremely realistic)". The overall experiment took roughly 45 min.

Acoustic Analysis

Four acoustic measurements were taken over each target sentence in both the Citation and Interaction blocks using Praat scripts (DiCanio, 2007; De Jong et al., 2017): intensity (dB), speech rate (syllables/second), mean fundamental frequency (f0) (semitones, ST, relative to 100 Hz), and f0 variation (ST). We centered the measurements from the Interaction blocks within-speaker, subtracting their Citation speech mean value (within-speaker, within-word). This measurement indicates changes from the speakers' citation form for that feature.

To extract vowel-level features, recordings were force-aligned (using the Forced Alignment and Vowel Extraction (FAVE) suite) (Rosenfelder et al., 2014). Next, vowel boundaries were hand-corrected by trained research assistants: vowel onsets and offsets were defined by the presence of both higher format structure and periodicity. Following hand-correction, we measured vowel duration and vowel formant frequency values (F1, F2) at vowel midpoint with FAVE-extract (Rosenfelder et al., 2014) for the subset of 13 words containing corner vowels: /i/ (cheek, weave, deed), /u/ (boot, hoop, toot, dune), /a/ (pod, cot, sock, tot), and /æ/ (bat, tap). We additionally scaled the formant frequency values (from Hertz) using a log base-10 transformation and centering each value to the subject's citation production values for that word (Nearey, 1978).

In order to assess whether speech changes made by participants were not simply alignment toward the interlocutors, the same sentence-level (rate, mean f0, f0 variation) and target vowel measurements (duration, F1, F2) were also taken over each interlocutor's production in Turn 2 (e.g., "I think I heard weave."). In order to compare across the interlocutors, formant frequency values (F1, F2) were centered relative to each interlocutor's mean value for that word (log mean normalization: Nearey, 1978).

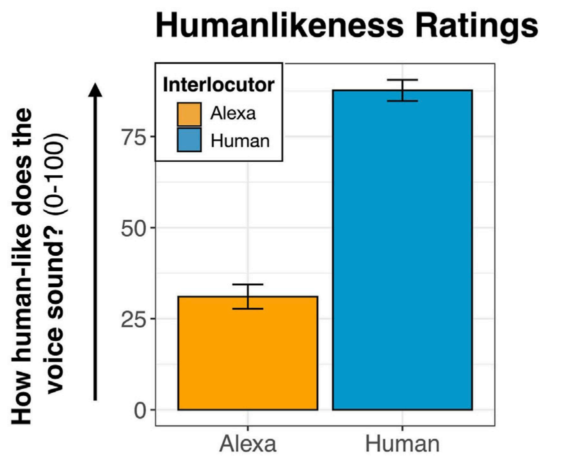


FIGURE 2 | Mean "human-like" ratings of each interlocutor. Error bars depict the standard error.

Statistical Analysis

Participants' sentence-level values for each acoustic feature (centered to speaker citation form values) were modeled in separate linear mixed effects models with the *lme4* R package (Bates et al., 2015), with identical model structure: fixed effects of Interlocutor (voice-AI, human), Staged Misunderstanding Condition (correctly heard, misrecognition), Expressiveness (neutral, expressive), and all possible interactions, with by-Sentence and by-Speaker random intercepts.

Participants' vowel-level features (F1, F2) were also modeled in separate linear mixed effects models with a similar structure as in the sentence-level models: Interlocutor, Staged Misunderstanding Condition, Expressiveness Condition, with by-Word and by-Speaker random intercepts. In both the F1 and F2 model, we included an additional predictor of Vowel Category (For the F1 (height) model, this factor included two height levels: high vs. low vowels; for the F2 (backness) model, this factor included two levels: front vs. back vowels) and all possible interactions with the other predictors (Vowel Category*Interlocutor*Misunderstanding*Emotion). The formant models (F1, F2) additionally included a fixed effect of Vowel Duration (centered within speaker).

RESULTS

Human-likeness Rating

Figure 2 provides the mean values for participants' human-like ratings of the voices. A *t*-test on participants' ratings of the voices confirmed that the Alexa voice was perceived as less human-like

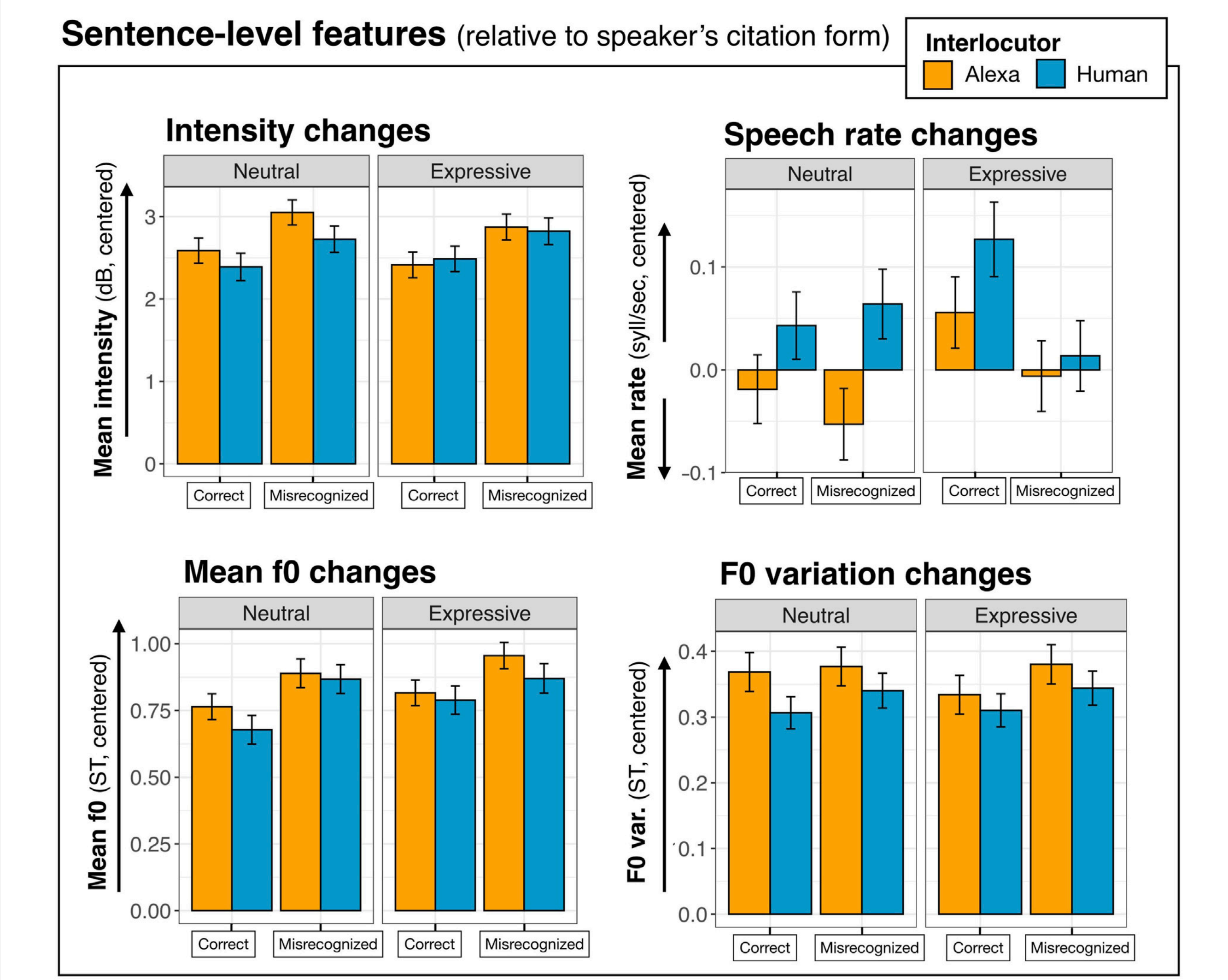


FIGURE 3 | Mean acoustic changes from speakers' citation form productions to the interaction with the Interlocutors (Alexa vs. human) for sentence intensity (in decibels, dB), speech rate (syllables per second), f0 (semitones, ST, rel. to 100 Hz), and f0 variation (ST). The *x*-axis shows Staged Misunderstanding Condition (correctly heard vs. misrecognized), while Expressiveness Condition is faceted. Values higher than 0.0 indicate an increase (relative to speakers' citation form), while values lower than 0.0 indicate a relative decrease. Error bars depict the standard error.

($\bar{x}$ = 31.06) than the human ($\bar{x}$ = 87.67) [t (104.87) = −12.84, p < 0.001].

Interlocutor Stimuli Acoustics

T-tests of the interlocutors' productions found no overall difference between the Alexa and Human speaking rate (Human $\bar{x}$ = 2.53 syll/s; Alexa $\bar{x}$ = 2.68 syll/s) [t (124.27) = −1.87, p = 0.06], but there was a significant difference in mean f0: the human had a higher mean f0 ($\bar{x}$ = 14.42 ST) than Alexa ($\bar{x}$ = 13.16 ST) [t (106.25) = 9.21, p < 0.001]. Additionally, the human produced greater f0 variation ($\bar{x}$ = 3.27 ST) than Alexa ($\bar{x}$ = 2.86 ST) [t (132.97) = 7.06, p < 0.001]. T-tests comparing formant frequency characteristics revealed no difference in vowel height (F1) for the interlocutors for high vowels (Human $\bar{x}$ = −0.37 log Hz; Alexa $\bar{x}$ = −0.41 log Hz) [t (35.28) = −1.38, p = 0.18] or low vowels (Human $\bar{x}$ = 0.43 log Hz; Alexa $\bar{x}$ = 0.47 log Hz) [t (45.42) = 1.75, p = 0.09]. Additionally, there was no difference in vowel fronting (F2) for the interlocutors for front vowels (Human $\bar{x}$ = 0.30; Alexa $\bar{x}$ = 0.35)[t (34.66) = 0.67, p = 0.51] or back vowels (Human $\bar{x}$ = -0.18; Alexa $\bar{x}$ = −0.22)[t (47.73) = 0.71, p = 0.48].

T-tests comparing the Expressiveness Conditions (neutral vs. emotionally expressive) confirmed differences: expressive productions were produced with a slower speaking rate (Expressive $\bar{x}$ = 2.45 syll/s; Neutral $\bar{x}$ = 2.76 syll/s) [t (153.88) = −4.25, p < 0.001] and with a lower mean f0 (Expressive $\bar{x}$ = 13.55 ST; Neutral $\bar{x}$ = 14.03 S T) [t (145.44) = −2.89, p < 0.01]. However, there was no difference for f0 variation (Expressive $\bar{x}$ = 3.04 ST; Neutral $\bar{x}$ = 3.09 ST) [t (157.44) = −0.60, p = 0.55].

T-tests comparing the Misunderstanding Conditions (correctly heard vs. misrecognition) showed no significant difference in speaking rate (Correct $\bar{x}$ = 2.64 syll/s; misunderstood $\bar{x}$ = 2.57 syll/s) [t (139.38) = 0.89, p = 0.37] or mean f0 (Correct $\bar{x}$ = 13.92 ST; misunderstood $\bar{x}$ = 13.65 ST) [t (114.62) = 1.60, p = 0.11]. However, they did vary in terms of f0 variation: larger for correctly understood ($\bar{x}$ = 3.15 ST) than misrecognized ($\bar{x}$ = 2.98 ST) [t (141.59) = 2.58, $p < 0.05$].

Participants' Sentence-Level Measurements

Figure 3 displays the mean acoustic values for participants' sentence-level measurements (centered to speakers' Citation form values). Model output tables are provided in **Supplementary Data Sheet 3, Appendices A1–A4**.

The Intensity model showed a significant intercept: participants increased their intensity in the interaction (relative to their citation form) [*Coef* = 2.64, *SE* = 0.45, t = 5.86, $p < 0.001$]. There was also a main effect of Misunderstanding Condition: as seen in **Figure 3**, participants' productions of sentences that the system did not understand correctly were louder than repetitions of utterances that the system understood correctly [*Coef* = 0.20, *SE* = 0.04, t = 5.13, $p < 0.001$]. No other effects or interactions were significant in the Intensity model.

The Speech Rate model showed no difference from 0 for the intercept: overall, speakers did not speed up or slow down their speech in interlocutor interactions, relative to their citation form productions. The model also revealed a main effect of Interlocutor, producing a slower speech rate (indicated by fewer syllables per second) in Alexa-DS [*Coef* = −0.03, *SE* = 0.01, t = −2.87, $p < 0.01$]. There was also a main effect of Misunderstanding Condition wherein speakers decreased their speech rate in response to a misrecognition [*Coef* = -0.02, *SE* = 0.01, t = −1.96, $p < 0.05$]. These effects can be seen in **Figure 3**. No other effects or interactions were significant in the model.

The Mean F0 model had a significant intercept, indicating that speakers increased their mean f0 in the interactions relative to the citation form productions [*Coef* = 0.83, *SE* = 0.15, t = 5.65, $p < 0.001$]. The model also showed an effect of Interlocutor: speakers produced a higher mean f0 toward the Alexa interlocutor [*Coef* = 0.03, *SE* = 0.01, t = 2.40, $p < 0.05$]. Additionally, there was an effect of Misunderstanding wherein responses to misunderstood utterances were produced with a higher f0 [*Coef* = 0.06 *SE* = 0.01, t = 5.04, $p < 0.001$], as seen in **Figure 3**. Furthermore, there was a main effect of Expressiveness Condition wherein speakers produced a higher mean f0 in response to emotionally expressive utterances [*Coef* = 0.03, *SE* = 0.01, t = 2.49, $p < 0.05$]. No other effects or interactions were observed in the Mean f0 model.

The F0 Variation model also had a significant intercept: relative to their citation form productions, speakers increased their f0 variation in the interaction [*Coef* = 0.34, *SE* = 0.07, t = 4.94, $p < 0.001$]. There was also a main effect of Interlocutor: speakers produced greater f0 variation in responses directed to the Alexa voice [*Coef* = 0.02, *SE* = 0.01, t = 2.79, $p < 0.01$]. Additionally, there was an effect of Misunderstanding: responses to misrecognitions were produced with greater f0 variation [*Coef* = 0.01, *SE* = 0.01, t = 1.98, $p < 0.05$]. No other effects or interactions were significant in the F0 Variation model.

Participants' Vowel-Level Measurements

Figure 4 displays participants' mean vowel-level values across conditions. Model output tables are provided in **Supplementary Data Sheet 3, Appendices A5 and A6**.

The F1 model testing changes in vowel height (where a smaller F1 values indicate raising) showed no significant intercept; relative to the citation forms, speakers did not change their vowel height. The model revealed only an effect of Vowel Duration: speakers produce lower vowels (higher F1) with increasing duration [*Coef* = 2.1e-04, *SE* = 7.8e-05, t = 2.62, $p < 0.01$]. No other effects or interactions were significant.

The F2 model, testing changes in vowel backness, showed several significant effects. While there was no significant intercept (indicating no general change in vowel backness from citation form), participants produced more backed vowels (i.e., lower F2 values) with increasing vowel duration [*Coef* = −1.8e-04, *SE* = 3.4e-05, t = −5.41, $p < 0.001$]. There was also an interaction between Misunderstanding Condition and Vowel Category. As seen in **Figure 4**, back vowels were produced even farther back (lower F2) in response to a staged word misrecognition [*Coef* = −0.01, *SE* = 1.5e-03, t = −3.46, $p < 0.001$]. No other effects or interactions were observed[4].

DISCUSSION

The current study examined whether participants use a different speech style when talking to an Alexa interlocutor, relative to a human interlocutor, in a computer-mediated interaction (a summary of the main effects is provided in **Table 2**). We systematically controlled functional and socio-communicative pressures in real-time during interactions with both interlocutors who made the same types and rates of staged word misrecognitions, and responded in emotionally expressive and neutral manners. This approach serves to complement studies done with users talking to devices in their home (e.g., Mallidi et al., 2018; Huang et al., 2019) and also pinpoint differences that might be present due to other factors in the situation (e.g., physical distance from the microphone; rate and type of automatic speech recognition (ASR) errors). While TTS methods have advanced in recent years (e.g., Wavenet in Van Den Oord et al., 2016), our participants rated the two talkers as distinct in their human-likeness: Alexa was less human-like than the human voice, consistent with prior work (Cohn et al., 2020b; Cohn and Zellou, 2020).

Overall, we found prosodic differences across Alexa- and human-DS, consistent with *routinized interaction accounts* that propose people have a "routinized" way of engaging with technology (Gambino et al., 2020), and in line with prior

[4]Note that while there is a numerical F2 increase in the Front Vowels in response to Misrecognized Expressive productions, this was not significant in the main model or in a post hoc model (with the subset of Front Vowels).

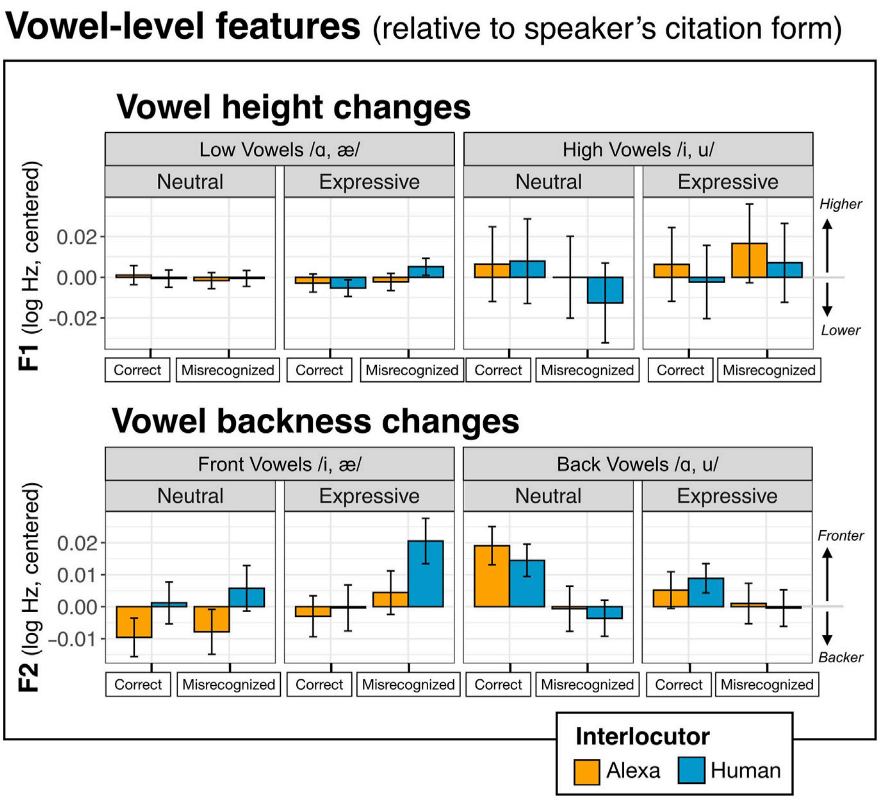


FIGURE 4 | Mean acoustic changes from speakers' citation form productions to the interaction with the Interlocutors (Alexa vs. human) for vowel duration (milliseconds, ms), F1 (log Hertz, Hz), and F2 (log Hertz, Hz). Formant plots are additionally faceted by Vowel Category: F1 (by vowel height: low vs. high vowels) and F2 (by vowel backness: front vs. back vowels). The *x*-axis shows Staged Misunderstanding Condition (correctly heard vs. misrecognized), while Expressiveness Condition is faceted. Values higher than 0.0 indicate an increase (relative to speakers' citation form), while values lower than 0.0 indicate a relative decrease. Error bars depict the standard error.

TABLE 2 | Summary of effects in main analysis, comparing interlocutor acoustics.

		Speaking style changes	Interlocutor acoustics
Sentence-level	Intensity	Louder for misrecognition	--
	Speech rate	Decreased rate in Alexa-DS Decreased rate for misrecognition	Alexa vs. human N.S. Correct vs. misrecognized N.S.
	Mean f0	Higher mean f0 in Alexa-DS Higher mean f0 for misrecognition Higher mean f0 for expressive	Human - higher mean f0 ($p < 0.001$) Correct vs. misrecognized N.S. Expressive - lower mean f0 ($p < 0.01$)
	F0 variation	More f0 variation in Alexa-DS More f0 variation for misrecognition	Human- larger f0 var. ($p < 0.001$) Correct - greater f0 var. ($p < 0.05$)
Vowel-level	F1 (vowel height)	No diff	Alexa vs. human N.S.
	F2 (vowel backness)	Back vowels backed for misrecognition	Alexa vs. human N.S.

studies finding differences in computer and voice-AI speech registers (e.g., Burnham et al., 2010; Huang et al., 2019; Siegert and Krüger, 2021). In the present study, speakers showed a systematic Alexa-DS speech style: when talking to Alexa, speakers produced sentences with a slower rate, higher mean f0, and higher f0 variation, relative to human-DS. These differences align with prior work showing slowed speech rate toward Alexa socialbot (Cohn et al., 2021), increased higher mean f0 in speech toward voice-AI (Raveh et al., 2019), and greater segmental lengthening in computer-DS (Burnham et al., 2010). Furthermore, both an increased mean f0 and f0 variation are consistent with increased vocal effort in response to a presumed communicative barrier; for instance, prior work has reported that speakers produce greater f0 variation in response to a word misrecognition in computer-DS (Vertanen, 2006), as well as higher mean f0 and a larger f0 range in Lombard speech (Brumm and Zollinger, 2011; Marcoux and Ernestus, 2019). Furthermore, in contrast to other work reporting greater intensity in Alexa-DS (Raveh et al., 2019; Siegert and Krüger, 2021), we did not see a difference in intensity in the present study. This might reflect the controlled interaction, where participants were recorded with a head-mounted microphone (such that it was equidistant from their mouths for the entire experiment) and heard amplitude normalized stimuli over headphones. Additionally, the lack of an intensity effect suggests that adjustments in Alexa-DS differ from strict "Lombard" effects (e.g., louder in Brumm and Zollinger, 2011).

While one possibility was that these adjustments reflect alignment toward the Alexa talker, we did not find support for this: acoustic analyses demonstrated that the Alexa productions had lower mean f0 and less f0 variation than the human productions (speech rate did not significantly differ for the Alexa and human productions). Hence, speakers appear to produce more effortful prosodic adjustments in response to an interlocutor with presumed communicative barriers (Clark and Murphy, 1982; Oviatt et al., 1998a; Branigan et al., 2011; Cowan et al., 2015), even while the "actual" misunderstandings were matched across the two talker types.

Do the differences in human- and Alexa-DS reflect *distinct* functionally oriented speech registers? Examining responses to misrecognized utterances suggests that some of these adjustments might be part of a more general speech intelligibility strategy. When either interlocutor "misheard" the word, participants responded by producing many of the same adjustments they did in Alexa-DS, including slower rate, higher f0, and higher f0 variation. These adjustments are in line with proposals that the speech adjustments people make in communicatively challenging contexts are listener-oriented (Lindblom, 1990; Smiljanić and Bradlow, 2009; Hazan and Baker, 2011). Thus, for these particular features, the adjustments made when there is a local communicative pressure parallel those made globally in Alexa-DS, suggesting that speakers make adjustments following misrecognitions and toward Alexa to improve intelligibility.

Yet, we see other adjustments in response to word misrecognitions not seen globally in Alexa-DS: increased intensity and F2 adjustments. These F2 adjustments, in particular, are predicted based on the type of misunderstanding created in the experimental design: when the interlocutor "misheard" the participant, they always produced the correct target word alongside its minimal pair counterpart which differed in backness (e.g., "mask" (front vowel) vs. "mosque" (back vowel)). Producing back vowels further back is consistent with vowel space expansion. In particular, one possibility is that these F2 adjustments are targeted specifically for clarity, making the vowels more distinct from the distractor minimal pair. This aligns with findings from Stent et al. (2008) who found that speakers repaired misrecognitions of high vowels by a dialogue system (e.g., "deed") by producing even *higher* vowels. That the same effect is not seen for front vowels in the current study could come from the dialectal variety of the speakers: participants were California English speakers, a variety with back vowel fronting (Hall-Lew, 2011). Thus, it is possible that there is more room for these speakers to make back vowels more back, rather than to adjust the front vowels, though further work exploring dialect-specific intelligibility strategies can shed light on this question (cf. Clopper et al., 2017; Zellou and Scarborough, 2019). Future work varying vowel height, as well as hyperarticulation of consonants (e.g., flapping vs. /t/ release in Stent et al., 2008) can further explore targeting effects in response to word misunderstandings.

However, if people produce global register differences in speech toward Alexa that parallel those seen in response to misrecognitions, why don't we see *greater* speech adjustments in response to misrecognitions made by Alexa? One possible explanation for the similarities is the rate: in the current study, the interlocutors both had staged word misrecognitions in 50% of trials. Related work has shown that rate of misrecognition can change speakers' global and local adaptations (Oviatt et al., 1998b; Stent et al., 2008); at a high rate of word misrecognitions, speakers might produce more similar intelligibility-related adjustments across interlocutors. Additionally, this high misrecognition rate—as well as random occurrence of the misunderstandings—might be interpreted by the speaker that the listener (human or Alexa) is not benefiting from these adjustments, which might drive similarities. In the current study, speakers might produce a word as clearly as they can and the human/voice-AI listener still misunderstands them half the time. The extent to which these patterns hold at a lower misrecognition rate—or an adaptive misrecognition rate, improving as the speaker produces "clearer" speech—are avenues for future work.

Furthermore, another possible reason for the similar intelligibility adjustments in response to a misunderstanding (in both Alexa- and human-DS) is that the speakers did not have access to information about the source of these perceptual barriers. For example, Hazan and Baker (2011) found that speakers dynamically adjust their speech to improve intelligibility when they are told their listener is hearing them in competing background speakers or as noise-vocoded speech (simulating the auditory effect of cochlear implants), relative to when the listener experienced no barrier. Furthermore, the *type* of adjustments varied according to the type of barrier (e.g., more f0 adjustments when the listener was in "babble" than "vocoded speech"). In the present study, speakers were left to "guess" what the source of the communicative barrier was, based on observed behavior of the human or voice-AI interlocutor. Indeed, when the speaker does not have information about the listener, adaptations might not be advantageous. For example, computer-DS adaptations have been shown in some work to lead to worse outcomes for some ASR systems, leading to a cycle of misunderstanding (e.g., Wade et al., 1992; for a discussion, see Stent et al., 2008; Oviatt et al., 1998b). Future work examining intelligibility for the intended listener (here, a human or ASR system) can further shed light on the extent local intelligibility adjustments in Alexa- and human-DS are equally beneficial.

Another possible factor why we see similar local intelligibility adjustments in response to misunderstandings (across Alexa- and human-DS) is that the experiment was computer-mediated. Recent work has shown differences in linguistic behavior across contexts: for example, participants show stronger style convergence toward their interlocutor in the in-person condition, relative to a (text-based) computer-mediated interaction (Liao et al., 2018). In line with this possibility, Burnham et al. (2010) found similar adjustments in response to a misrecognition made by a computer- and human-DS (but overall differences in computer-DS, paralleling our findings). At the same time, in the current study, the human-likeness ratings for the interlocutors collected at the end of study suggest that the participants found the interlocutors to be distinct. Future work manipulating rate of misunderstanding and embodiment (Staum Casasanto et al., 2010; Cohn et al., 2020a) can investigate what conditions lead to greater targeted intelligibility strategies for distinct interlocutor types.

We also explored whether emotional expressiveness mediates speech styles for Alexa- and human-DS. Here, we found the same speech adjustments in response to expressiveness by both interlocutors: higher mean f0 in response to utterances containing emotional expressiveness. First, speakers' overall higher f0 in their sentences does not appear to reflect an alignment toward the interlocutors (who actually produced lower mean f0 in their expressive productions). One possible explanation for the increased f0 following the expressive responses is that it reflects a positivity bias in reaction to stimuli (but see Jing-Schmidt (2007) for work on biases toward negative valence). Indeed, work has shown that smiling is associated with higher mean f0 (Tartter, 1980; Tartter and Braun, 1994) (but we did not see formant shifts, which are also associated with smiled speech, in response to Expressiveness). Here, one explanation for similarities in response to emotion by both interlocutors is that speakers are applying the social behaviors toward voice-AI as they do toward humans, as proposed by *technology equivalence accounts* (Nass et al., 1994; Nass et al., 1997; Lee, 2008). For instance, here people are reacting to emotional expressiveness by both types of interlocutors similarly. This explanation is consistent with work showing similar affective responses to computers as seen in human-human interaction (e.g., Brave et al., 2005; Cohn et al., 2019a; Cohn and Zellou, 2019).

Additionally, we did not observe differences in how participants adapted their speech following an emotionally expressive or neutral word misrecognition. This contrasts with related work on this same corpus (Zellou and Cohn, 2020) that found greater vowel duration alignment when participants responded to an emotionally expressive word misunderstanding made by a voice-AI system. Thus, it is possible that emotional expressiveness might shape vocal alignment, but it might not influence speech style adjustments. That emotion appears to have an effect on vocal alignment toward humans and voice-AI (e.g., Vaughan et al., 2018; Cohn and Zellou, 2019) could be explained by proposals that alignment is used as a means to communicate social closeness (Giles et al., 1991). While conveying affect is thought to be part of infant- and pet-DS registers (Trainor et al., 2000), listener-oriented speech styles directed toward human adults (non-native speakers, hearing impaired speakers) and computers are generally not associated with increased emotionality. Furthermore, conveying affect is generally not associated with clear speech strategies. Indeed, classic perspectives on clear speech (H&H theory) do not account for emotionality in predicting hyperspeech behavior (e.g., Lindblom, 1990). Yet, one possibility for a lack of difference in the current study is based on how emotion was added in the stimuli: emotional expressiveness was conveyed only in the interjection. Since the time this study was run, there are now more ways to adapt the Alexa voice in terms of positive and negative emotionality (at low, medium, and high levels[5]), which can serve as avenues for future research.

There were also several limitations of the present study which open directions for future work. For instance, one possible factor in the lack of difference detected for emotionality across Alexa- and human-DS is the communicative context: the current study consisted of fully scripted interactions in a lab setting. While this controlled interaction was intentional as we were interested in word misrecognitions (which might otherwise be difficult to control in voice-AI interactions), it is possible that differences based on emotional expressiveness might be seen in a non-scripted conversation with voice-AI, as well as one conducted outside a lab context (e.g., Cohn et al., 2019b). Additionally, the present study used two types of voices; it is possible that other paralinguistic features of those voices might have mediated speech style adjustments. For example, recent work has shown that speakers align speech differently toward TTS voices that "sound" older (e.g., Apple's Siri voices, rated in their 40 and 50s) (Zellou et al., 2021). Furthermore, there is work showing that introducing "charismatic" features from human speakers' voices shapes perception of TTS voices (Fischer et al., 2019; Niebuhr and Michalsky, 2019). The extent to which individual differences in speakers (human and TTS) and participants remain avenues for future research.

While here the findings align with those for another Germanic language (e.g., German in Raveh et al., 2019; Siegert and Krüger, 2021), the extent to which the same effects might be observed with other languages and other cultures is another open question for future work. For example, cultures might vary in terms of acceptance of voice-AI technology, such as due to privacy concerns (e.g., GDPR in Europe: Voss, 2016; Loideain and Adams, 2020). Additionally, cultures vary in terms of their expressions of emotion (Shaver et al., 1992; Mesquita and Markus, 2004; Van Hemert et al., 2007). How emotional expressiveness and "trust" in voice-AI (Shulevitz, 2018; Metcalf et al., 2019) might interact remains an open question for future work.

CONCLUSION

Overall, this work adds to our growing understanding of the dynamics of human interaction with voice-AI assistants—still distinct from how individuals talk to human interlocutors. As these systems and other AI robotics systems are even more widely

[5]https://developer.amazon.com/en-US/docs/alexa/custom-skills/speech-synthesis-markup-language-ssml-reference.html#amazon-emotion

adopted, characterizing these patterns across different timepoints—and with diverse populations of participants—is important in our ability to track the trajectory of the influence of voice-AI on humans and human speech across languages and cultures.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the **Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of California Institutional Review Board (IRB). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MC and GZ contributed to conception and design of the study. MC programmed the experiment and performed the statistical analysis. Both authors contributed to manuscript drafting and revision. Both have read, and approved the submitted version.

FUNDING

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC and an Amazon Faculty Research Award to GZ.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.675704/full#supplementary-material

REFERENCES

Ameka, F. (1992). Interjections: The Universal yet Neglected Part of Speech. *J. Pragmatics* 18 (2–3), 101–118. doi:10.1016/0378-2166(92)90048-g

Ammari, T., Kaye, J., Tsai, J. Y., and Bentley, F. (2019). Music, Search, and IoT. *ACM Trans. Comput.-Hum. Interact.* 26 (3), 1–28. doi:10.1145/3311956

Arnold, R., Tas, S., Hildebrandt, C., and Schneider, A. (2019). "Any Sirious Concerns yet?–An Empirical Analysis of Voice Assistants' Impact on Consumer Behavior and Assessment of Emerging Policy," in *TPRC47: Research Conference on Communications, Information and Internet Policy*, Washington, DC, September 20–21, 2019.

Babel, M. (2012). Evidence for Phonetic and Social Selectivity in Spontaneous Phonetic Imitation. *J. Phonetics* 40 (1), 177–189. doi:10.1016/j.wocn.2011.09.001

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Soft.* 67 (1), 1–48. doi:10.18637/jss.v067.i01

Bell, L., Gustafson, J., and Heldner, M. (2003). "Prosodic Adaptation in Human-Computer Interaction," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Vol. 3, 833–836.

Bell, L., and Gustafson, J. (1999). Repetition and its Phonetic Realizations: Investigating a Swedish Database of Spontaneous Computer-Directed Speech. *Proc. ICPhS* 99, 1221–1224.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., and Lottridge, D. (2018). Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2 (3), 1–24. doi:10.1145/3264901

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Brown, A. (2011). The Role of Beliefs in Lexical Alignment: Evidence from Dialogs with Humans and Computers. *Cognition* 121 (1), 41–57. doi:10.1016/j.cognition.2011.05.011

Brave, S., Nass, C., and Hutchinson, K. (2005). Computers that Care: Investigating the Effects of Orientation of Emotion Exhibited by an Embodied Computer Agent. *Int. J. Human-Computer Stud.* 62 (2), 161–178. doi:10.1016/j.ijhcs.2004.11.002

Brumm, H., and Zollinger, S. A. (2011). The Evolution of the Lombard Effect: 100 Years of Psychoacoustic Research. *Behav.* 148 (11–13), 1173–1198. doi:10.1163/000579511x605759

Burbach, L., Halbach, P., Plettenberg, N., Nakayama, J., Ziefle, M., and Valdez, A. C. (2019). "" Hey, Siri"," Ok, Google"," Alexa". Acceptance-Relevant Factors of Virtual Voice-Assistants," in *2019 IEEE International Professional Communication Conference (ProComm)*, Aachen, Germany, July 2019 (IEEE), 101–111. doi:10.1109/procomm.2019.00025

Burnham, D. K., Joeffry, S., and Rice, L. (2010). "Computer-and Human-Directed Speech before and after Correction," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, Melbourne, Australia, December 14–16, 2010, 13–17. Available at: http://handle.uws.edu.au:8081/1959.7/504796.

Clark, H. H., and Murphy, G. L. (1982). "Audience Design in Meaning and Reference," in *Advances in Psychology*. Editors J. F. Le Ny and W. Kintsch (North-Holland, The Netherlands: Elsevier), Vol. 9, 287–299. doi:10.1016/s0166-4115(09)60059-5

Clark, H. H. (1996). *Using Language*. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/cbo9780511620539

Clopper, C. G., Mitsch, J. F., and Tamati, T. N. (2017). Effects of Phonetic Reduction and Regional Dialect on Vowel Production. *J. Phonetics* 60, 38–59. doi:10.1016/j.wocn.2016.11.002

Cohn, M., Chen, C.-Y., and Yu, Z. (2019a). "A Large-Scale User Study of an Alexa Prize Chatbot: Effect of TTS Dynamism on Perceived Quality of Social Dialog," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden, September 11–13, 2019 (Association for Computational Linguistics), 293–306. Available at: https://www.sigdial.org/files/workshops/conference20/proceedings/cdrom/pdf/W19-5935.pdf.

Cohn, M., Ferenc Segedin, B., and Zellou, G. (2019b). "Imitating Siri: Socially-Mediated Alignment to Device and Human Voices," in *Proceedings of International Congress of Phonetic Sciences*, Melbourne, Australia, August 5–9, 2019, 1813–1817. Available at: https://icphs2019.org/icphs2019-fullpapers/pdf/full-paper_202.pdf.

Cohn, M., Jonell, P., Kim, T., Beskow, J., and Zellou, G. (2020a). "Embodiment and Gender Interact in Alignment to TTS Voices," in *Proceedings of the Cognitive Science Society*, Toronto, Canada, July 31–August 1, 2020, 220–226. Available at: https://cogsci.mindmodeling.org/2020/papers/0044/0044.pdf.

Cohn, M., Liang, K.-H., Sarian, M., Zellou, G., and Yu, Z. (2021). Speech Rate Adjustments in Conversations with an Amazon Alexa Socialbot. *Front. Commun.* 6, 1–8. doi:10.3389/fcomm.2021.671429

Cohn, M., Sarian, M., Predeck, K., and Zellou, G. (2020b). "Individual Variation in Language Attitudes toward Voice-AI: The Role of Listeners' Autistic-like Traits," in *Proceedings of the Interspeech 2020*, Shanghai, China, October 25–29, 2020, 1813–1817. doi:10.21437/Interspeech.2020-1339

Cohn, M., and Zellou, G. (2019). "Expressiveness Influences Human Vocal Alignment toward Voice-AI," in *Proceedings of the Interspeech 2019*, Graz, Austria, September 15–19, 2019, 41–45. doi:10.21437/Interspeech.2019-1368

Cohn, M., and Zellou, G. (2020). "Perception of Concatenative vs. Neural Text-To-Speech (TTS): Differences in Intelligibility in Noise and Language Attitudes," in

Proceedings of the Interspeech, Shanghai, China, October 25–29, 2020, 1733–1737. doi:10.21437/Interspeech.2020-1336
Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., and Beale, R. (2015). Voice Anthropomorphism, Interlocutor Modelling and Alignment Effects on Syntactic Choices in Human−computer Dialogue. *Int. J. Human-Computer Stud.* 83, 27–42. doi:10.1016/j.ijhcs.2015.05.008
De Jong, N. H., Wempe, T., Quené, H., and Persoon, I. (2017). Praat Script Speech Rate V2. Available at: https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2 (Accessed August 30, 2020).
DiCanio, C. (2007). Extract Pitch Averages. Available at: https://www.acsu.buffalo.edu/~cdicanio/scripts/Get_pitch.praat (Accessed May 12, 2019).
Fischer, K., Niebuhr, O., Jensen, L. C., and Bodenhagen, L. (2019). Speech Melody Matters—How Robots Profit from Using Charismatic Speech. *ACM Trans. Human-Robot Interaction (Thri)* 9 (1), 1–21. doi:10.1145/3344274
Gambino, A., Fox, J., and Ratan, R. A. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Commun.* 1 (1), 5. doi:10.30658/hmc.1.5
Giles, H., Coupland, N., and Coupland, J. (1991). "Accommodation Theory: Communication, Context, and Consequence," in *Contexts of Accommodation* (Cambridge, United Kingdom: Cambridge University Press), Vol. 1, 1–68.
Goffman, E. (1981). *Forms of Talk*. Philadelphia, PA: University of Pennsylvania Press, 78–122. Response Cries.
Habler, F., Schwind, V., and Henze, N. (2019). "Effects of Smart Virtual Assistants' Gender and Language," in *Proceedings of Mensch und Computer 2019*, Hamburg, Germany, September 8–11, 2019, 469–473. doi:10.1145/3340764.3344441
Hall-Lew, L. (2011). "The Completion of a Sound Change in California English," in *Proceedings of ICPhS XVII*, Hong Kong, China, August 17–21, 2011, 807–810.
Hazan, V., and Baker, R. (2011). Acoustic-phonetic Characteristics of Speech Produced with Communicative Intent to Counter Adverse Listening Conditions. *The J. Acoust. Soc. America* 130 (4), 2139–2152. doi:10.1121/1.3623753
Hazan, V. L., Uther, M., and Granlund, S. (2015). "How Does Foreigner-Directed Speech Differ from Other Forms of Listener-Directed clear Speaking Styles?," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, United Kingdom, August 10–14, 2015.
Hoffmann, L., Krämer, N. C., Lam-Chi, A., and Kopp, S. (2009). "Media Equation Revisited: Do Users Show Polite Reactions towards an Embodied Agent?," in *International Workshop on Intelligent Virtual Agents*, Amsterdam, The Netherlands, September 14–16, 2009, 159–165. doi:10.1007/978-3-642-04380-2_19
Huang, C.-W., Maas, R., Mallidi, S. H., and Hoffmeister, B. (2019). "A Study for Improving Device-Directed Speech Detection toward Frictionless Human-Machine Interaction," in *Proceeding of the Interspeech 2019*, Graz, Austria, September 15–19, 2019, 3342–3346. doi:10.21437/Interspeech.2019-2840
Jing-Schmidt, Z. (2007). Negativity Bias in Language: A Cognitive-Affective Model of Emotive Intensifiers. *Cogn. Linguist*. 18(3), 417–443. doi:10.1515/COG.2007.023
Junqua, J.-C. (1996). The Influence of Acoustics on Speech Production: A Noise-Induced Stress Phenomenon Known as the Lombard Reflex. *Speech Commun.* 20 (1–2), 13–22. doi:10.1016/s0167-6393(96)00041-6
Junqua, J. C. (1993). The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers. *J. Acoust. Soc. America* 93 (1), 510–524. doi:10.1121/1.405631
Knoll, M. A., Johnstone, M., and Blakely, C. (2015). "Can You Hear Me? Acoustic Modifications in Speech Directed to Foreigners and Hearing-Impaired People," in *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, September 6–10, 2015. Available at: https://www.isca-speech.org/archive/interspeech_2015/i15_2987.html.
Lee, K. M. (2008). "Media Equation Theory," in *The International Encyclopedia of Communication* (Malden, MA: John Wiley & Sons, Ltd), Vol. 1, 1–4. doi:10.1002/9781405186407.wbiecm035
Liao, W., Bazarova, N. N., and Yuan, Y. C. (2018). Expertise Judgment and Communication Accommodation in Linguistic Styles in Computer-Mediated and Face-To-Face Groups. *Commun. Res.* 45 (8), 1122–1145. doi:10.1177/0093650215626974
Lindblom, B. (1990). "Explaining Phonetic Variation: A Sketch of the H&H Theory," in *Speech Production and Speech Modelling*. Editors W. J. Hardcastle and A. Marchal (Dordrecht, The Netherlands: Springer), Vol. 55, 403–439. doi:10.1007/978-94-009-2037-8_16
Loideain, N. N., and Adams, R. (2020). From Alexa to Siri and the GDPR: the Gendering of Virtual Personal Assistants and the Role of Data protection Impact Assessments. *Comp. L. Security Rev.* 36, 105366. doi:10.1016/j.clsr.2019.105366
Lopatovska, I. (2020). "Personality Dimensions of Intelligent Personal Assistants," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, Vancouver, BC, Canada, March 14–18, 2020, 333–337. doi:10.1145/3343413.3377993
Lunsford, R., Oviatt, S., and Arthur, A. M. (2006). "Toward Open-Microphone Engagement for Multiparty Interactions," in *Proceedings of the 8th International Conference on Multimodal Interfaces*, New York, NY, November 2–4, 2006, 273–280. doi:10.1145/1180995.1181049
Mallidi, S. H., Maas, R., Goehner, K., Rastrow, A., Matsoukas, S., and Hoffmeister, B. (2018). *Device-directed Utterance Detection*. Hyderabad, India: Interspeech 2018ISCA. doi:10.21437/interspeech.2018-1531
Maniwa, K., Jongman, A., and Wade, T. (2009). Acoustic Characteristics of Clearly Spoken English Fricatives. *J. Acoust. Soc. America* 125 (6), 3962–3973. doi:10.1121/1.2990715
Marcoux, K. P., and Ernestus, M. T. C. (2019). "Differences between Native and Non-native Lombard Speech in Terms of Pitch Range," in *Proceedings of the 23rd International Congress on Acoustics*, Berlin, Germany, September 9–13, 2019, 5713–5720. doi:10.18154/rwth-conv-239240
Mayo, C., Aubanel, V., and Cooke, M. (2012). "Effect of Prosodic Changes on Speech Intelligibility," in *Thirteenth Annual Conference of the International Speech Communication Association*, Portland, OR, September 9-13, 2012, 1706–1709. Available at: https://isca-speech.org/archive/archive_papers/interspeech_2012/i12_1708.pdf.
Mesquita, B., and Markus, H. R. (2004). "Culture and Emotion," in *Feelings and Emotions:The Amsterdam Symposium*. Editors A. S. R. Manstead, N. Frijda, and A. Fischer (Cambridge, United Kingdom: Cambridge University Press), 341–358. doi:10.1017/cbo9780511806582.020
Metcalf, K., Theobald, B.-J., Weinberg, G., Lee, R., Jonsson, M., Webb, R., et al. (2019). "Mirroring to Build Trust in Digital Assistants," in *Proc. Interspeech*, Graz, Austria, September 15–19, 2019, 4000–4004. doi:10.21437/Interspeech.2019-1829
Nass, C., Moon, Y., Morkes, J., Kim, E.-Y., and Fogg, B. J. (1997). Computers Are Social Actors: A Review of Current Research. *Hum. Values Des. Comp. Tech.* 72, 137–162.
Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers Are Social Actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston, MA, April 24–28, 1994, 72–78. doi:10.1145/259963.260288
Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels*. Bloomington, IN: Indiana University Linguistics Club.
Niebuhr, O., and Michalsky, J. (2019). "Computer-Generated Speaker Charisma and its Effects on Human Actions in a Car-Navigation System Experiment - or How Steve Jobs' Tone of Voice Can Take You Anywhere," in *International Conference on Computational Science and Its Applications*, Saint Petersburg, Russia, July 1–4, 2019, 375–390. doi:10.1007/978-3-030-24296-1_31
Oviatt, S., Levow, G.-A., Moreton, E., and MacEachern, M. (1998a). Modeling Global and Focal Hyperarticulation during Human-Computer Error Resolution. *J. Acoust. Soc. America* 104 (5), 3080–3098. doi:10.1121/1.423888
Oviatt, S., MacEachern, M., and Levow, G.-A. (1998b). Predicting Hyperarticulate Speech during Human-Computer Error Resolution. *Speech Commun.* 24 (2), 87–110. doi:10.1016/s0167-6393(98)00005-3
Picheny, M. A., Durlach, N. I., and Braida, L. D. (1985). Speaking Clearly for the Hard of Hearing I. *J. Speech Lang. Hear. Res.* 28 (1), 96–103. doi:10.1044/jshr.2801.96
Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005). Lexical Frequency and Acoustic Reduction in Spoken Dutch. *J. Acoust. Soc. America* 118 (4), 2561–2569. doi:10.1121/1.2011150
Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. (2017). "'Alexa is My New BFF': Social Roles, User Satisfaction, and Personification of the Amazon Echo," in *Proceedings of the 2017 CHI Conference Extended*

Abstracts on Human Factors in Computing Systems, Denver, CO, May 6–11, 2017, 2853–2859. doi:10.1145/3027063.3053246

Raveh, E., Steiner, I., Siegert, I., Gessinger, I., and Möbius, B. (2019). "Comparing Phonetic Changes in Computer-Directed and Human-Directed Speech," in *Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, Dresden, Germany, March 3–8, 2019, 42–49.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., et al. (2014). FAVE (Forced Alignment and Vowel Extraction). *Program Suite v1. 2.2.* doi:10.5281/zenodo.9846 (Accessed July 13, 2020).

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., and Brenier, J. (2007). "An Acoustic Study of Real and Imagined Foreigner-Directed Speech," in *Proceedings of the International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6–10, 2007, 2165–2168. Available at: http://icphs2007.de/conference/Papers/1673/1673.pdf.

Scarborough, R., and Zellou, G. (2013). Clarity in Communication: "Clear" Speech Authenticity and Lexical Neighborhood Density Effects in Speech Production and Perception. *J. Acoust. Soc. America* 134 (5), 3793–3807. doi:10.1121/1.4824120

Shaver, P. R., Wu, S., and Schwartz, J. C. (1992). "Cross-cultural Similarities and Differences in Emotion and its Representation," in *Emotion* (Thousand Oaks, CA: Sage Publications, Inc), 175–212. doi:10.21236/ada248148

Shulevitz, J. (2018). "Alexa, How Will You Change Us," in *The Atlantic*, Vol. 322, 94–104. Available at: https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/.

Siegert, I., Krüger, J., Egorow, O., Nietzold, J., Heinemann, R., and Lotz, A. (2018). "Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction Using Amazon's ALEXA," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May 7–12, 2018. Available at: http://lrec-conf.org/workshops/lrec2018/W20/pdf/13_W20.pdf.

Siegert, I., and Krüger, J. (2021). "Speech Melody and Speech Content Didn't Fit Together"-Differences in Speech Behavior for Device Directed and Human Directed Interactions," in *Advances in Data Science: Methodologies and Applications*. Editors G. Phillips-Wren, A. Esposito, and L. C. Jain (Switzerland: Springer), Vol. 189, 65–95. doi:10.1007/978-3-030-51870-7_4

Siegert, I., Nietzold, J., Heinemann, R., and Wendemuth, A. (2019). "The Restaurant Booking Corpus–Content-Identical Comparative Human-Human and Human-Computer Simulated Telephone Conversations," in *Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, 126–133.

Smiljanić, R., and Bradlow, A. R. (2009). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Lang. Linguist Compass* 3 (1), 236–264. doi:10.1111/j.1749-818X.2008.00112.x

Staum Casasanto, L., Jasmin, K., and Casasanto, D. (2010). "Virtually Accommodating: Speech Rate Accommodation to a Virtual Interlocutor," in *32nd Annual Meeting of the Cognitive Science Society (CogSci 2010)*, Portland, OR, August 11–14, 2010, 127–132. Available at: http://hdl.handle.net/11858/00-001M-0000-0012-BE12-A.

Stent, A. J., Huffman, M. K., and Brennan, S. E. (2008). Adapting Speaking after Evidence of Misrecognition: Local and Global Hyperarticulation. *Speech Commun.* 50 (3), 163–178. doi:10.1016/j.specom.2007.07.005

Swerts, M., Litman, D., and Hirschberg, J. (2000). "Corrections in Spoken Dialogue Systems," in *Sixth International Conference on Spoken Language Processing*, Beijing, China, October 16–20, 2000. Available at: https://www.isca-speech.org/archive/icslp_2000/i00_2615.html.

Tartter, V. C., and Braun, D. (1994). Hearing Smiles and Frowns in normal and Whisper Registers. *J. Acoust. Soc. America* 96 (4), 2101–2107. doi:10.1121/1.410151

Tartter, V. C. (1980). Happy Talk: Perceptual and Acoustic Effects of Smiling on Speech. *Perception & Psychophysics* 27 (1), 24–27. doi:10.3758/bf03199901

Trainor, L. J., Austin, C. M., and Desjardins, R. N. (2000). Is Infant-Directed Speech Prosody a Result of the Vocal Expression of Emotion? *Psychol. Sci.* 11 (3), 188–195. doi:10.1111/1467-9280.00240

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you Speak E-NG-L-I-SH? A Comparison of Foreigner- and Infant-Directed Speech. *Speech Commun.* 49 (1), 2–7. doi:10.1016/j.specom.2006.10.003

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv*. ArXiv Prepr. ArXiv160903499.

Van Hemert, D. A., Poortinga, Y. H., and van de Vijver, F. J. R. (2007). Emotion and Culture: A Meta-Analysis. *Cogn. Emot.* 21 (5), 913–943. doi:10.1080/02699930701339293

Vaughan, B., De Pasquale, C., Wilson, L., Cullen, C., and Lawlor, B. (2018). "Investigating Prosodic Accommodation in Clinical Interviews with Depressed Patients," in *International Symposium on Pervasive Computing Paradigms for Mental Health*, Boston, January 9–10, 2018, 150–159. doi:10.1007/978-3-030-01093-5_19

Vertanen, K. (2006). "Speech and Speech Recognition during Dictation Corrections," in *Ninth International Conference on Spoken Language Processing*, 1890–1893.

Voss, W. G. (2016). European union Data Privacy Law Reform: General Data protection Regulation, Privacy Shield, and the Right to Delisting. *The Business Lawyer* 72 (1), 221–234.

Wade, E., Shriberg, E., and Price, P. (1992). "User Behaviors Affecting Speech Recognition," in *Second International Conference on Spoken Language Processing*, Banff, Alberta, Canada, October 13–16, 1992. Available at: https://www.isca-speech.org/archive/archive_papers/icslp_1992/i92_0995.pdf.

Zellou, G., Cohn, M., and Ferenc Segedin, B. (2021). Age- and Gender-Related Differences in Speech Alignment toward Humans and Voice-AI. *Front. Commun.* 5, 1–11. doi:10.3389/fcomm.2020.600361

Zellou, G., and Cohn, M. (2020). "Social and Functional Pressures in Vocal Alignment: Differences for Human and Voice-AI Interlocutors," in *Proceeding of the Interspeech 2020*, Shanghai, China, October 25–29, 2020, 1634–1638. doi:10.21437/Interspeech.2020-1335

Zellou, G., and Scarborough, R. (2019). Neighborhood-conditioned Phonetic Enhancement of an Allophonic Vowel Split. *J. Acoust. Soc. America* 145 (6), 3675–3685. doi:10.1121/1.5113582

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fcomp.2021.685277

Check for updates

Development of an Instrument to Measure Conceptualizations and Competencies About Conversational Agents on the Example of Smart Speakers

Carolin Wienrich[1]* *and Astrid Carolus*[2]

[1]*Institute Human-Computer-Media, Human-Technique-Systems, Julius-Maximilians-University, Wuerzburg, Germany,* [2]*Institute Human-Computer-Media, Media Psychology, Julius-Maximilians-University, Wuerzburg, Germany*

OPEN ACCESS

Edited by:
Stefan Hillmann,
Technical University Berlin, Germany

Reviewed by:
Patricia Martin-Rodilla,
University of A Coruña, Spain
Ronald Böck,
Otto von Guericke University Magdeburg, Germany
Bernd Ludwig,
University of Regensburg, Germany

***Correspondence:**
Carolin Wienrich
carolin.wienrich@uni-wuerzburg.de

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

Received: *24 March 2021*
Accepted: *12 July 2021*
Published: *02 August 2021*

Citation:
Wienrich C and Carolus A (2021) Development of an Instrument to Measure Conceptualizations and Competencies About Conversational Agents on the Example of Smart Speakers. Front. Comput. Sci. 3:685277. doi: 10.3389/fcomp.2021.685277

The concept of digital literacy has been introduced as a new cultural technique, which is regarded as essential for successful participation in a (future) digitized world. Regarding the increasing importance of AI, literacy concepts need to be extended to account for AI-related specifics. The easy handling of the systems results in increased usage, contrasting limited conceptualizations (e.g., imagination of future importance) and competencies (e.g., knowledge about functional principles). In reference to voice-based conversational agents as a concrete application of AI, the present paper aims for the development of a measurement to assess the conceptualizations and competencies about conversational agents. In a first step, a theoretical framework of "AI literacy" is transferred to the context of conversational agent literacy. Second, the "conversational agent literacy scale" (short CALS) is developed, constituting the first attempt to measure interindividual differences in the "(il) literate" usage of conversational agents. 29 items were derived, of which 170 participants answered. An explanatory factor analysis identified five factors leading to five subscales to assess CAL: storage and transfer of the smart speaker's data input; smart speaker's functional principles; smart speaker's intelligent functions, learning abilities; smart speaker's reach and potential; smart speaker's technological (surrounding) infrastructure. Preliminary insights into construct validity and reliability of CALS showed satisfying results. Third, using the newly developed instrument, a student sample's CAL was assessed, revealing intermediated values. Remarkably, owning a smart speaker did not lead to higher CAL scores, confirming our basic assumption that usage of systems does not guarantee enlightened conceptualizations and competencies. In sum, the paper contributes to the first insights into the operationalization and understanding of CAL as a specific subdomain of AI-related competencies.

Keywords: artificial intelligence literacy, artificial intelligence education, voice-based artificial intelligence, conversational agents, measurement

INTRODUCTION

Digitalization offers new opportunities across various aspects of our lives—in work-related and private environments. New technologies are increasingly interactive revealing multifarious potentials on both an individual and a societal level. Digital voice assistant systems, for example, have grown in popularity over the last few years (Hernandez 2021; Meticulous 2021), offering an intuitive way of use: Simply by talking to the device, the user can operate it. Although today's usage scenarios are still limited and voice-based assistants in private households are rather used as remote controls (e.g., to play music or turn on the lights) or for web searches (e.g., for the weather forecast), future usage scenarios suggest that voice-based systems could be omnipresent and ubiquitous in our future lives. Multiple new and more complex use cases will result in more complex interactions involving heterogeneous user groups. For example, both children and older people could benefit from the ease of use in a private environment, at school, or in a nursing or health care context (e.g., ePharmaINSIDER, 2018; The Medical Futurist, 2020). Moreover, in a work-related environment, differently qualified employees could interact with these systems, providing knowledge and support to solve specific problems (Baumeister et al., 2019). Application areas of voice-based assistants are further discussed in (Kraus et al., 2020).

Across all the (future) scenarios, ease of use is one of the most promising features of a voice-based system. The increasingly intuitive usage results in decreasing requirements regarding the users' technical knowledge or capacities. In contrast, the complexity of the technological systems and their engineering increases. Attributable to the penetration of easy-to-use voice-based systems, the gap between usage and knowledge increases and gains importance. In turn, unknowing and illiterate users tend to fundamental misconceptions of the technology, e.g., regarding functional principles, expectations, beliefs, and attitudes towards the technological system. Misconceptions or limited knowledge about digital technologies constrain the effective, purposeful, and sovereign use of technology skills (e.g., Chetty et al., 2018). The user remains a somewhat naive consumer of easy-to-use applications, who tends to interact with the systems mindlessly, blindly trusting their device, unthoughtfully sharing private data, or expecting human-like reactions from the device. The latter might be particularly relevant for voice-based systems directly interacting with users, such as conversational agents (short CA). Referring to a societal level, limited competencies and misconceptions contribute to the biased public debate, which focuses either on the risks or glorifies the use of digital technology (Zhang and Dafoe, 2019; Kelley et al., 2019). As a result, users are far from becoming informed and critical operators who understand the opportunities digitization offers in general, and conversational agents in particular, who know how and when to use them, or who know when to refrain from usage (e.g., Fast and Horvitz, 2017). The *Eliza-Effect* and the *Tale-Spin Effect* are two prominent examples of misconceptions (Wardrip-Fruin, 2001). When a system uses simple functions that produce effects appearing complex, i.e., *Eliza-Effect*, users might overestimate the capabilities of the system. For example, when a speech-based recommender system gives an advice the user will follow this advice without further verification as because he/she trusts in the correctness of the device. In contrast, when a system uses complex functions that produce effects appearing less complex, i.e., *Tale-Spin-Effect*, users might underestimate the capabilities of the system. Such a recommender system might elicit less credibility and users would disregard (correct) advices from the system.

One approach to address this gap is to develop more self-explanatory systems to provide services for which the user needs no prior knowledge. For a detailed discussion of explainable AI systems refer to (Doran, Schulz, and Besold 2018; Goebel et al., 2018). Another approach focuses on the detection of users' misconceptions or limited competencies to learn about the users and to derive design user-centered learning or training programs. Due to the omnipresence and increasing penetration of conversational agents, one key factor of successful digitalization is yielding users with appropriate conceptualizations about conversational agents and competencies to operate them. Consequently, we decided to focus on user's conceptions and competencies. However, which particular conceptualization and competencies are relevant for a "literate" usage of conversational agents? How can such conceptualization and competencies be measured? To answer these questions, in a first step, a conceptualization of "AI literacy" is transferred to the context of conversational agent literacy. The present study specified conceptualization and competencies recently reported as relevant for developing "AI literacy" in general (Long and Magerko, 2020), for one voice-based conversational agent proxy, i.e., smart speakers. The subdomain is defined as the "conversational agent literacy" (short CAL). Then two methodological parts follow. In the first part, the "conversational agent literacy scale" (short CALS) is developed, constituting a first attempt to measure interindividual differences in the "(il)literate" usage of conversational agents. With our focus on smart speakers, we derived 29 items. 170 participants answered these items. An explanatory factor analysis identified five factors leading to five subscales to assess CAL. Subscales and items were analyzed regarding reliability and student's CAL. In the second part, insights into construct validity and impacts of interindividual characteristics have been tested with a subsample of 64 participants. Thus the present study contributes to a first understanding of CAL as a specific subdomain of AI-related conceptualizations and competencies, which allows a sovereign use of conversational technology to unfold the full potential of digitization (e.g., Burrell, 2016; Fast and Horvitz, 2017; Long and Magerko, 2020). Long and Magerko (2020) offered a collection of competencies that are important for AIL. At least to the best knowledge of the authors, the present paper is the first approach to developing an operationalization of this collection.

RELATED WORK

Conversational Agents

A conversational agent is a computer system, which emulates a conversation with a human user (McTear et al., 2016). The dialogue system manages the recognition of speech input, the analysis, the processing, the output, and the rendering on the basis of AI-related methods such as natural language processing, natural language understanding, and natural language generation (Klüwer, 2011; McTear et al., 2016). CAs employ one or more input and output modalities such as text (i.e., chat agents), speech (voice agents), graphics, haptics, or gestures. Various synonyms such as conversational AI, conversational interfaces, dialogue systems, or natural dialogue systems result in conceptional blurring (Berg 2015). In the following, we refer to "conversational agents" and consider the modality of speech. To be more specific, we focused on smart speakers. Smart speakers allow users to activate the device using an intend-word or wake-word such as "Alexa". After the activation, it records what is being said and sends this over the *internet* to the main processing area. The voice recognition service decodes the speech and then sends a response back to the smart speaker. For example, the speech file is sent to Amazon's AVS (Alexa Voice Services) in the cloud for the Amazon system. Amazon published the underlying speech recognition and natural language processing technology with the service of Amazon Lex. Please refer to that service for more technical details. We refer to conceptualizations and competencies essential for the interaction with and understanding of voice-based conversational agents, specifically smart speakers. Since our target group mainly knows the products of Amazon and Google, the study results are closely linked to these devices. However, the basic principles of the approach presented could also be transferred to the modality of text-based systems, as large parts of the underlying operations are similar for both systems.

Media-Related Competencies: From Digital Literacy to Conversation Agent Literacy

In our modern information society, knowledge about digitization processes and digital technologies becomes increasingly relevant. For about a decade, the responsible and reflected use of digital media has been discussed as a **new cultural technique**, which exceeds literacy and numeracy e.g., (Belshaw, 2011). However, it is not easy to provide an exact and distinct definition of digital competencies as different authors have introduced various meanings and definitions (Baacke et al., 1999; Groeben and Hurrelmann, 2002; Güneş and Bahçivan, 2018; Janssen et al., 2013). For example, Gallardo-Echenique et al., 2015 identified a wide range of concepts and approaches associated with digital competence in a literature review, i.e., digital literacy, digital competence, eLiteracy, e-skills, eCompetence, computer literacy, and media literacy. Early concepts such as computer literacy primarily referred to the ability to use a text-processing program or to search the WWW for information (Shapiro and Hughes 1996). Information literacy focused on the individual's more profound cognitive processes of information processing, such as the ability to understand, evaluate, and use information effectively regardless of its multimedia form (Oxbrow 1998). More recently, frameworks of digital competencies neglect operation skills and refer to a broader set of abilities, including technical and non-technical skills e.g., Chetty et al., 2018).

The rise of AI requires a further extension of the concept of literacy. In this sense, Long and Magerko (2020, p. 2) define "artificial intelligence literacy" (AIL) as a 'set of competencies that enables individuals to evaluate AI technologies critically; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace'. Long and Magerko (2020) presented a literature review analyzing 150 studies and reports to derive a conceptual framework of AIL. Their framework involved five themes, each characterized by a set of 17 competencies and 15 design considerations (**Supplementary Table S8**), describing multifaceted aspects of AIL. The identified five main themes or guiding questions are: 1) What is AI?, 2) What can AI do?, 3) How does AI work?, 4) How should AI be used?, and 5) How do people perceive AI? The framework offers a collection of competencies essential for AIL but lacks an operationalization of AIL, allowing a valid and reliable measure of the AIL aspects. Although Long and Magerko (2020, p. 10), themselves argue that "…there is still a need for more empirical research to build a robust and accurate understanding", their descriptive framework constitutes a good starting point for research on AI-related conceptualizations and competencies. Since no other frameworks or theoretical concepts of AIL exist, at least to the authors' best knowledge, the AIL-framework serves as a basis for our scale development.

Since AI is employed in many different applications and systems, the present paper focuses on a subgroup of AI-based systems, which have recently become increasingly important in many human-AI interactions: voice-based conversational agents, specifically smart speakers. To the best knowledge of the authors, conceptualizations, and competencies essential for the interaction and understanding of (voice-based) conversational agents have not been considered yet. Thus, we introduce the conversational agent literacy (CAL) as a subdomain of AIL. CAL employs conceptualizations (e.g., perceptions, attitudes, mental models) and competencies (e.g., knowledge, interactions skills, critical reflection skills) about the CA itself and the interaction with the CA. From an HCI perspective, identifying and monitoring CAL is of utmost importance because future usage scenarios suggest (voice-based) conversational agents to be omnipresent and ubiquitous in our lives and involve more heterogeneous user groups (Baumeister et al., 2019). Equipping users with appropriate conceptualizations and competencies with regard to digital technology will allow sovereign interactions with digital technologies (e.g., Burrell, 2016; Fast and Horvitz, 2017; Long and Magerko, 2020) and (voice-based) conversational agents. However, monitoring CAL requires measurements, which provide the individual assessments of CAL and indicate development potentials.

Measuring Media-Related Competencies: From Digital Literacy to Conversation Agent Literacy

In accordance with the various interpretations of digital competencies, the measures are multifaceted and standardized instruments are missing. Jenkins (2006) developed a twelve-factorial tool assessing general handling with media. Among others, it includes play ("When I have a new cell phone or electronic device, I like to try out . . . "), performance ("I know what an avatar is."), and *multitasking* ("When I work on my computer, I like to have different applications open in the same time.") (see also Literat, 2014). Porat et al. (2018) used a six-factorial instrument measuring digital literacy competencies. It includes, for example, *photo-visual literacy* ("understanding information presented in an illustration."), *information literacy* ("Identifying incorrect or inaccurate information in a list of internet search results."), and *real-time-thinking literacy* ("Ignoring ads that pop up while looking for information for an assignment."). Other approaches used open-ended questions to assess *information* ("judging its relevance and purpose"), *safety* ("personal and data protection"), and *problem-solving* ("solve conceptual problems through digital means") (e.g., Perdana et al., 2019). Another instrument for measuring digital literacy comes from Ng (2012), who distinguished a *technical dimension* (technical and operational skills for learning with information and communication technology and using it in everyday life), a *cognitive dimension* (ability to think critically about searching, evaluating, and creating digital information) and a *social-emotional dimension* (ability to use the internet responsibly for communication, socializing and learning). In addition, the scale also measures attitudes towards the use of digital technologies. The social-emotional dimension of this approach extends previous measures by considering the interactivity of digital technology. However, the development of Ng (2012) scale is not based on any conceptional or empirical foundation leading to difficulties regarding a valid, reliable, and comparable use of the instrument.

In sum, the approaches aiming to measure digital competencies are rather limited in terms of their conceptual range revealing (for a review: Covello, 2010). Moreover, instruments measuring digital competencies have rarely referred to artificial intelligence literacy or conceptualizations and competencies relevant for the sovereign use of (voice-based) conversational agents. The few studies in the field aim for the assessment of associations and perceptions about AI, referring only to the fifth theme (How do people perceive AI?) of the conceptional AIL-framework mentioned above (e.g., AI in general: Eurobarometer, 2017; Zhang and Dafoe 2019b; Kelley et al., 2019; voice-based conversational agent-specific: Zeng et al., 2017; Lau et al., 2018; Hadan and Patil, 2020). However, an analysis of the items of the instruments revealed that the majority referred to either digitalization in general or specific embodiments such as robots (e.g., Eurobarometer 2017). Consequently, and regarding the latent variable, it remains somewhat unclear what exactly the items measured. Besides, the studies report neither the underlying conceptual framework nor criteria of goodness (e.g., reliability, validity). Alternatively, they used single items instead of validated scales. In sum, the quality of measurements available remained unclear. In the area of AIL-related competencies, and particularly regarding conceptualizations and competencies of voice-based CAs, the development of measures and instruments is still in its very early stages. Until today, the literature review reveals no valid and reliable instrument to assess CAL resulting in a research desideratum the present study focuses on.

In sum, digital competencies are associated with a wide range of concepts and measures. The rise of voice-based conversational agents requires further extend the idea of digital literacy. Knowledge has been shown to be a key factor for using new technologies competently. Assessing the state of knowledge allows the implementation of precisely fitting training and transformation objectives (for an overview: Chetty et al., 2018). In this sense, Long and Magerko, (2020) introduced a broad conceptional framework of literacy in the context of AI. However, research so far has not presented tools or instruments allowing an assessment of interindividual levels of AIL or related subdomains such as CAL. Aiming for a first attempt to close this gap, the present study focuses on.

1) the development of an empirically founded measuring instrument to assess CAL and
2) the investigation of first insights into the validation and the impact of interindividual differences.

Overview of the Present Work

With our focus on smart speakers as a proxy of voice-based CAs, we derived 29 items portraying 16 of the original 17 competencies and four of the original five themes introduced by (Long and Magerko, 2020). 170 participants answered these items. An explanatory factor analysis identified five factors suggesting a different structure than the original framework. Items were assigned to five subscales to assess (voice-based) conversational agent literacy (CAL). Subscales and items were analyzed regarding reliability and student's performances (Part I). Finally, preliminary insights into construct validity and the analyses of interindividual characteristics offered insights into CAL of a subsample of 64 participants (Part II). In sum, the present paper presents the first attempt to quantify and measure CAL as a sub-domain of AIL using an empirically based instrument, which follows the conceptional framework introduced by Long and Magerko (2020).

PART I: CONSTRUCTION OF SCALES, FACTOR ANALYSIS, RELIABILITY ANALYSIS, AND STUDENT'S PERFORMANCES

Methods

Development of the Items

Items were derived from the 16-dimensional framework by Long and Magerko (2020). For each dimension, four researchers

TABLE 1 | Item Pool: Example items of the 16 dimensions of (voice-based) conversational agent literacy on the example of a smart speaker.

1. Dimension what is AI?	
3. Sub-dimension: Interdisciplinarity	
What do you think: Which of the following technical disciplines plays no role in the development of a smart speaker?	A- computer science B- Psychology C - Pharmacy D - mechanical engineering E - sociology
2. Dimension: What can AI do?	
5. Sub-dimension: AI's strength and weakness	
What do you think: Which of the following areas of knowledge can a person answer better than a smart speaker?	A- concrete factual knowledge B- abstract concepts C - past events D - interhuman communication E - reasons for factual facts
3. Dimension: How does AI work?	
12. Learning from data	
What do you think: What role do previous requests play in the smart Speaker's response to a current request?	A- past requests are irrelevant B- past requests enable the smart speaker to better understand current requests C -past requests allow to create a profile about personal preferences D - past requests are not stored and therefore do not play a role for current requests E - past requests only play a role if they were spoken in the same pitch
4. Dimension: How should AI be used?	
16. Sub-dimension: Ethics	
What do you think: Is the data from the interaction with a smart speaker shared with advertising companies?	A- yes, to display more relevant ads based on personal interests B- No, because such data is not meaningful enough C - No, because this is a privacy issue. It is not allowed D - yes, because such data provides a detailed insight into the habits of the user E - yes, but only if the users agree to this in the smart speaker settings

individually 1) comprehended the meaning of the dimensions, and 2) developed and phrased items consisting of a question and answering options. Then, the four researchers collaboratively analyzed their individual initial item pools regarding content validity, redundancy, and comprehensibility of the items. Furthermore, to account for non-expert respondents, the final items and their responses were as simple and as unambiguous as possible. In sum, 29 items were derived, with each dimension being referred to by at least one item. Every item consisted of a question and five answers from which the participant needed to select the correct ones. Each of the five answers could be correct or incorrect and received equal weight when summed to arrive at the final score for each of the 29 items. Across these questions, different numbers of the answers could be correct. Therefore, the aggregate score can be from 0 (no correct answers) to five correct answers per item, with higher scores indicating better-founded competencies. **Table 1** presents an extract of the items used in the process of questionnaire development.

Subjects

One hundred seventy participants voluntarily engaged in an online survey. They were between 16 and 55 years old (M = 21.06, SD = 5.04), 82.7% were female. Regarding the highest educational qualification, 83.53% have finished secondary school, 4.71% reported vocational training of 4.71% had a bachelor's, and 1.18% a master's degree. 11.12% owned an Amazon Echo device, 4.12% owned a Google Home.

Procedure

Participants were briefly instructed about the general purpose of the study, with the procedure following the ethical guidelines laid out by the German Psychological Association. Participants answered the 29 questions referring to conversational agent literacy.

Results

The following section presents the factor analysis results and the scale, item analysis, and student's performances on the CAL scores.

Factor Analysis

We conducted a factor analysis of the conceptualizations and competencies questions to gain deeper insights into the factorial structure of (voice-based) conversational agent literacy. Factor analysis is a "multivariate technique for identifying whether the correlations between a set of observed variables stem from their relationship to one or more latent variables in the data, each of which takes the form of a linear model" (Field, 2018, p.1016). There were two possibilities for the analysis: an exploratory factor analysis (EFA) or a confirmatory factor analysis (CFA). The CFA tests specific associations between items and latent variables, which a model or a framework hypothesizes. Without any a priori assumptions, the EFA searches for associations between the items indicating underlying common factors which explain the variation in the data (Field, 2018). Referring to the framework postulated by (Long and Magerko, 2020), a CFA would aim to verify the postulated 4-themes factorial structure or the 16-competencies factorial structure. However, we decided to conduct an EFA as a first step of the empirical analysis of the conceptualization of CAL. An explorative analysis of the factorial solution provides the opportunity to detect factorial solutions deviating from the postulated factorial structure. Nevertheless, if our empirical data reflected the postulated factors, an EFA would reveal a four- or 16-factorial solution.

A principal component analysis was conducted. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO >0.5 (KMO = 0.744; Kaiser and Rice, 1974), and Bartlett's test of sphericity was significant ($\chi^2_{(406)} = 1{,}314.29$, $p <$ 0.000). On the level of KMO values of the individual items, one item was lower than the acceptable limit of 0.5 (Kaiser and Rice, 1974) With reference to (Bühner, 2011), we decided to keep the items in the analysis. In the results part, we will come back to this item again. Anticipating the results section, the factor analysis will reveal that this item will not be part of the final scales. Consequently, the problem can be neglected. To identify the number of meaningful factors, a parallel analysis was conducted, resulting in ambiguous solutions: on the one hand, a three-factorial solution was indicated, on the other hand solutions with four, five and six factors were also justifiable. The additional analysis of the scree plot (detection of a break between the factors with relatively large eigenvalues and those with smaller eigenvalues) could not clarify this ambiguity. As an ambiguous basis of decision-making is a typical challenge during the process of an EFA, content-related and mathematical argumentation needs to be combined for informed choices (Howard, 2016). Therefore, we decided to discard the three-factorial solutions and maintained the five and six factors. Three factors would have meant an inadequate reduction of the postulated four- or 16-dimensional framework (content-related argument) and a relatively small amount of explained variance (three-factor solution: 33.577%; five: 44.170; six: 51.125; mathematical argument). Because the resulting factors were hypothesized to be intercorrelated, two promax (oblique) rotations—on five and on six factors—were conducted. Then, the rotated solutions were interpreted following both content criteria (conceptual fitting of the items loading on the factor; conceptual differences between items of different factors) and statistical criteria (factor loadings below 0.2 were excluded, loadings above 0.4/0.5 indicated relevance). Negative loadings were also excluded, as they would indicate reverse item coding, which would not work for conceptualizations and competencies items. The two rotated factor patterns of pattern loadings resulting from the promax rotation of five and six factors are presented in (**Supplementary Table S9**).

When analyzing the two sets of factor loadings, factors 1, two and three were almost identical in both solutions, factors 4 and 5 were roughly the same. The analysis of the sixth factor of the 6-factor solution reveals that most of its items are either absorbed by the third or fourth factor in the 5-factorial solution or are below the threshold of 0.2 factor loadings. Moreover, this factor 6 involves the item with the non-acceptable MSA-value. Consequently, the sixth factor could hardly bring added exploratory value resulting the rejection of the 6-factorial solution and the acceptance of the 5-factorial solution.

To interpret the meaning of the resulting five factors, the items with the highest loadings served as reference values. Then, additional items were added following statistical criteria of factor loadings and content criteria asking for the fit with the overall meaning of the other items. As a result, five subscales with three to six items were derived to reflect CAL. However, we must be careful with the interpretation of the resulting scales, as they result from a very first attempt of scale development in conversational agent literacy. We, therefore, regard the present scale development more as a kind of work-in-progress report, the limitations of which need to be discussed in the discussion section. **Table 2** gives an overview of the preliminary version of the CALS, its subscales, and their items. Additionally, the table presents the original concept by Long and Magerko (2020) as well as the assignment of this study's items derived from it.

Scale and Item Analysis

Table 3 presents the results of scale and item analyses. Cronbach's α (internal consistency) was computed for each subscale (factor) and the total scale. Values were between 0.34 and 0.79, indicating mixed results. With reference to the fact that this paper provides a first insight into the operationalization of CAL (and very first insights into the more general concept of AI literacy), we argue with Nunnally (1978) that "in the early stages of research [. . .] modest reliability" is acceptable. Thus, we carefully regard the internal consistency of subscales 1 to 3 as acceptable [see also (Siegert et al., 2014), who discuss the challenge of inter-rater reliability in the context of emotion annotation in human-computer interaction]. The fourth and fifth subscale, however, reveal more questionable values (0.34 and 0.35). In sum, we preliminarily maintain these subscales and argue for future optimization (see discussion for more detailed considerations). The average item's *difficulty index p* is 68.84, with mean difficulties ranging from 60.31 (factor 1) to 79.98

TABLE 2 | Empirically derived CALS-subscales and their items compared to the original dimension by Long and Magerko (2020).

CALS-subscales CALS	CAL-items	Original dimension Long and Magerko (2020)	Original assignment of CAL-items
1 storage and transfer of the smart speaker's data input CALS	20, 21, 24, 26, 27, 28	1 what is AI?	1–3
2 smart speaker's functional principles CALS	5, 8, 9, 11, 15, 16	2 what can AI do?	4–7
3 smart speaker's intelligent functions, learning abilities CALS	1, 4, 12, 14, 19	3 how does AI work?	8–26
4 smart speaker's reach and potential CALS	2, 6, 22	4 how should AI be used?	27–29
5 smart speaker's technological (surrounding) infrastructure	10, 13, 25		

TABLE 3 | Scale and Item analysis of the CAL.

CAL subscale	Internal consistency: Cronbach's α	Item difficulty: Mean (range)	Item discrimination: Mean (range)
1 storage and transfer of the smart speaker's data input	0.788	60.31 (41.56–66.83)	0.55 (0.28–0.66)
2 smart speaker's functional principles	0.690	79.98 (60.12–89.88)	0.43 (0.25–0.52)
3 smart speaker's intelligent functions, learning abilities	0.623	76.83 (67.76–85.92)	0.36 (0.22–0.60)
4 smart speaker's reach and potential	0.337	62.75 (58.47–66.12)	0.22 (0.15–0.32)
5 smart speaker's technological (surrounding) infrastructure	0.345	64.32 (50.59–71.26)	0.21 (0.20–0.22)
Total scale	0.780	68.84	0.35

Note: Items were not corrected in terms of chance score (chose the correct options by guessing).

TABLE 4 | Intercorrelations of the CALS subscales.

CALS subscales	1	2	3	4	5
1 storage and transfer of the smart speaker's data input	—	—	—	—	—
2 smart speaker's functional principles	0.235**	—	—	—	—
3 smart speaker's intelligent functions, learning abilities	0.290**	0.406**	—	—	—
4 smart speaker's reach and potential	0.210**	0.225**	0.314**	—	—
5 smart speaker's technological (surrounding) infrastructure	0.331**	0.253**	0.225**	0.084	—
Total scale	0.727**	0.678**	0.695**	0.525**	0.515

(factor 2). Corrected item scale correlation indicates item discrimination ranging from 0.22 to 0.541.

Table 4 presents the intercorrelations of the five CAL subscales with each other and with total CAL-score revealing significant correlations across all scales (only exception: subscales four x 5). With the exception of the correlation between the fourth and fifth scales, the correlations indicate medium effects (Cohen 1988). Subscale five does not correlate significantly with the total scale demanding detailed attention in the upcoming steps. Assessment of smart speaker's technological (surrounding) infrastructure (subscale 5) seems to be independent of the conceptualizations and competencies about conversational agents.

Conversational Agent Literacy: Conversational Agent Literacy-Scores

To provide coherent results, which allow interpretation of the participants' level of CAL, we needed to adjust the CAL scores to correct for guessed answers. Each item consisted of a question and five answers resulting in five potentially correct answers and five points to gain. For each answer, participants had to decide whether the answer was right or wrong. Consequently, participants had a 50–50 chance to choose the correct response simply by guessing, resulting in 2.5 points correctly guessed points, on average. To control for an overestimation of the performance, the values of the CALS need to be corrected for chance level. Correct answers have been counted as +1, incorrect answers as −1, resulting in a total range per item from −5 (all options per item were answered incorrectly) to +5 (all options per item were answered correctly). **Table 5** gives an overview of the initial and the corrected CAL-values. As the initial and the corrected CAL scores only differ in absolute values but are comparable regarding relative values, they both indicate conversational agent literacy on an intermediate level—with highest values for subscale one and lowest for subscale 5.

PART II: (INITIAL STEPS OF) VALIDATION AND IMPACT OF INTERINDIVIDUAL CHARACTERISTICS

Method

Subjects

Sixty four participants (71.9% women), between 18 and 55 years old ($M = 23.64$, $SD = 7.13$) participated additionally in the second

TABLE 5 | CAL-scores of the student sample.

CALS	Corrected values mean (SD)	Corrected values range: 5.00 to +5.00	Initial valuesMean (SD)	Initial values range: 0.00 to 5.00
1 storage and transfer of the smart speaker's data input	2.76 (1.31)	−0.67 to +5.00	3.01 (1.20)	0.83 to 5.00
2 smart speaker's functional principles	3.20 (1.17)	−1.33 to +4.67	4.00 (1.06)	0.17 to 5.00
3 smart speaker's intelligent functions, learning abilities	3.14 (1.13)	0.20 to +4.60	3.84 (1.16)	0.40 to 5.00
4 smart speaker's reach and potential	1.68 (1.92)	−2.33 to +5.00	3.14 (1.33)	0.48. to 5.00
5 smart speaker's technological (surrounding) infrastructure	1.85 (1.57)	−2.33 to +5.00	3.22 (1.11)	0.70 to 5.00
Total scale	2.70 (0.85)	+0.13 to +4.30	3.44 (1.17)	0.52 to 5.00

TABLE 6 | The survey structure of PART II.

Survey part	Construct	Sub-dimensions
1. validation	Competency-relation	Affinity for technology, technology commitment
	Attitudes towards smart speakers	Attitude towards smart speakers
2. Impact of interind. Characteristics	Individual characteristics	Self-efficacy in smart speaker usage, general self-efficacy
	Demographical data	Gender, age, ownership of voice assistants

part. 70.3% were students, 3.1% have finished secondary school, 10.9% reported vocational training, 12.5% had a bachelor's, and 3.1% a master's degree. 12.5% owned an Amazon Echo, 1.6% a Google Home.

Procedure

To learn more about underlying characteristics and interindividual differences between participants, we assessed additional variables via standardized questionnaires, including competency-related measures and attitudes towards digital technologies and smart speakers, psychological characteristics, and demographical data (an overview is given in **Table 6**).

Measures

Conversational Agent literacy (CAL) was assessed by the newly developed scale (see Part I), incorporating 23 items (see **Table 1** for example items), which were summarized to five factors. CAL values as an indicator of participants conversational agent literacy were then correlated with the following items.

Competency-related constructs were measured by the German version of the *affinity for technology interaction scale* (short *ATI* scale) by Franke et al. (2019). The 6-point Likert scale includes nine items (e.g., "I like to occupy myself in greater detail with technical systems.") ranging from 1 "completely disagree' to 6 "completely agree'. Items were averaged so that higher values indicated higher levels of technique affinity. The internal consistency of the scale was $\alpha = 0.90$.

In addition, the German version of the *commitment for technology short scale* (short *CT*) by Neyer et al. (2012) was used. Twelve items were answered on a 5-point Likert scale ranging from 1 "completely disagree" to 5 "completely agree". A *total score* and three subscales can be calculated: *technique acceptance*, *technique competence beliefs*, and *technique control beliefs*. Items were averaged with higher values indicating higher levels of technique commitment. The internal consistency of the scale was $\alpha = 0.84$.

Attitudes towards smart speakers (short *ATSS*) were measured by using an adapted version of the *Negative Attitude toward Robots Scale* by Nomura et al. (2006). Across all items, we replaced "robots" with "smart speakers". Since the original items measured negative attitudes, we modified the wording as it was already done in (Wienrich and Latoschik 2021). In correspondence to each emotionally coded word used in the original items, we created a semantic differential scale so that the positive and negative emotions would be linked to high and low values, respectively. For example, we built the semantic differential "nervous/relaxed" instead asking "I would feel nervous operating a smart speaker in front of other people." Analogously to the original scale, the adapted scale consists of 14 items, which are answered on a 5-point Likert scale ranging from 1 "strongly disagree" to 5 "strongly agree" on a semantic differential. The *total score* and three subscales can be calculated: *attitudes toward (a) situations of interaction with smart speakers*, *the social influence of smart speakers*, and 3) *emotions in interaction with smart speakers*. Again, items were averaged, and higher values indicated higher levels of positive attitudes. Internal consistency of the scale was $\alpha = 84$.

Psychological characteristics were measured by an adaptation of the *self-efficacy in the human-robot interaction scale* (short *SE-HRI*) by Pütten and Bock (2018). We replaced "robots" with "smart speakers" (short *SES*). Ten items were answered on a 6-point Likert scale ranging from 1 "strongly disagree" to 6 "strongly agree" with higher values of averaged items indicating higher levels of self-efficacy in smart speaker interaction. The internal consistency of the scale was $\alpha = 0.91$. In addition, general self-efficacy was measured using the *general self-efficacy scale* (short *SEG*) by Schwarzer et al. (1997). Ten items (e.g., "For each problem, I will find a solution.") were answered on a 6-point Likert scale ranging from 1 "completely disagree" to 6 "completely agree". Items were averaged, and higher values indicated higher levels of general self-efficacy. The internal consistency of the scale was $\alpha = 0.82$.

TABLE 7 | Correlations between CALS subscales and technology-related competencies and attitudes.

CALS	ATI	CT: Total	CT: Acceptance	CT: Competence	CT: Control	ATSS: Total	ATSS: Situations	ATSS: Influence	ATSS: Emotions
1 storage and transfer of the smart speaker's data input	0.278[a]	0.254[a]	0.058	0.354[b]	0.132	–0.334[b]	–0.290[a]	–0.333[b]	–0.245
2 smart speaker's functional principles	0.297[a]	0.269[a]	0.200	0.254[a]	0.127	–0.332[b]	–0.222	–0.457[b]	–0.195
3 smart speaker's intelligent functions, learning abilities	0.405[b]	0.176	0.070	0.258[a]	0.046	–0.070	–0.015	–0.161	–0.014
4 smart speaker's reach and potential	0.079	0.044	0.021	0.173	–0.111	0.138	0.170	0.137	0.031
5 smart speaker's technological (surrounding) infrastructure	–0.077	0.032	-0.019	0.148	–0.069	–0.284[a]	–0.247[a]	–0.310[a]	–0.178
Total	0.404[b]	0.305[a]	0.147	0.434[b]	0.068	-0.297[a]	–0.193	–0.397[b]	–0.195

[a]p < .05.
[b]p < .001; ATI = Affinity for Technology Interaction; CT = Commitment to Technology (subscales: technique acceptance, competence beliefs, control beliefs.); ATSS = Attitudes toward Smart Speaker (subscales: attitudes toward situations of interaction with smart speakers, social influence of smart speakers, emotions in interaction with smart speakers).

Demographical data included gender ("female", "male", "diverse"), age, knowledge of German language, level of education, field of study, and the previous experiences with smart speaker (from 1 "never" to 5 "very often") as well as the ownership of smart speakers (e.g., "Alexa/Amazon Echo" or "Google Home").

Expectations for the Validation of the Questionnaire

Besides the analysis of the reliability, the validation of a newly developed questionnaire is a crucial challenge to meet. To evaluate construct validity, CALS subscales are embedded into a "nomological net" of similar constructs. To gain first insights, we analyze the associations between CALS subscales, scales measuring competency-related constructs and attitudes towards technology (Cronbach and Meehl, 1955). Because the competency-related scales are supposed to measure similar constructs, they shall positively correlate with positive attitudes towards technology (ATI, CT). Being more literate in the area of conversational agents should be associated with increased technology affinity. However, attitudes are not to be equated with competencies, so on the one hand, we do not assume perfect correlations, and on the other hand - depending on the type of attitude - we expect negative and positive correlations.

Results: (Initial Steps of) Validation

Correlations with scales measuring similar constructs establish the first steps towards a nomological net of CAL. **Table 7** reveals correlations between the total score of CAL, the factors, and the chosen variables. As expected, Affinity for Technology (ATI) and commitment to technology (CT) correlated positively with the total CAL scale (CALS Total), indicating overlapping concepts. Furthermore, the medium to large positive correlation between CALS Total and the CT-subscale competence can be interpreted as a first indicator of the validity of CALS. The non-significant correlation of CALS four and CALS five underlines this conclusion as both scales do not explicitly refer to competencies.

Remarkably, the analysis revealed that participants' attitudes toward smart speakers (ATSS) correlated negatively with CALS Total as well as with the subscales CALS 1, CALS 2, and CALS 5. As with commitment before, CALS four did not correlate significantly with attitudes, the same applies to CALS 3. Thus, while participants' competencies in terms of their knowledge about "storage and transfer of the smart speaker's data input" (CALS 1), its "functional principles" (CALS 2) and its technological (surrounding) infrastructure (CALS 5) are negatively associated with their attitudes towards the devices, the participants' knowledge of the intelligent functions and learning abilities of smart speakers are not associated to their attitudes.

Regarding the subscales of ATSS, attitudes toward situations of interaction with smart speakers and social influence of smart speakers are also negatively correlated with CALS 1, 2, and 5. However, the ATSS-subscale referring to emotions occurring when interacting with smart speakers reveals lower and non-significant correlations.

Results: participants' conversational agent literacy and associated psychological and demographic characteristics.

The final step of our analyses aims for first insights into associations between conversational agent literacy and psychological and demographic characteristics. To analyze the impact of interindividual differences, we compared female and male participants and analyzed the effects of age, smart speaker ownership, smart speaker self-efficacy, and general self-efficacy.

The analysis of gender differences revealed that male participants achieved slightly higher scores on CALS-items (total scale: M = 3.39, SD = 0.41) than woman (M = 3.76, SD = 0.38). However, differences across all subscales were small and not significant. Furthermore, our sample's gender ratio was not balanced, so this result should be viewed with caution.

Regarding age, the only significant correlation was found for CALS 5 (*smart speaker's reach and potential*), which correlated significantly positively with age ($r = 0.379$; $p = 0.02$). However, the small age range must be considered when interpreting this result.

Interestingly, the ownership of a smart speaker voice assistant did not result in significantly different CAL scores.

The general self-efficacy did not show any significant correlation with CALS-scales. Similarly, participants' self-efficacy in terms of smart speaker interaction did not correlate with the CAL scales. The only exception is CALS 2 (*smart* *speaker's functional principles*), which does correlate significantly positively ($r = 0.341$; $p = 0.006$).

DISCUSSION AND OUTLOOK

The concept of digital literacy has been introduced as a new cultural technique complementing the former predominant focus on media literacy and numeracy (e.g., Belshaw, 2011). With reference to the increasing importance of computer technology and particularly AI, literacy concepts need to be extended to meet new technological and AI-related developments. In this sense, Long and Magerko (2020) proposed a conceptional framework of AI literary (AIL), constituting a good starting point for research on conceptualizations and competencies in the area of digital and AI technologies. Encouraged by their conceptional work, the present paper aims for the development of measurement to empirically assess the postulated conceptualizations and competencies. To avoid the reference point of rather abstract AI applications, this study focuses on voice-based conversational agents. Consequently, the original conceptualization of "AI literacy" is transferred to the context of "conversational agent literacy" (CAL). With the goals of developing a measurement tool for CAL and constituting first steps towards an assessment of interindividual differences in the " (il)literate" usage of conversational agents, part I develops the "conversational agent literacy scale" (short CALS). Part II reveals first insights into the "nomological net" of constructs similar to CAL. Moreover, the first associations between CAL and psychological characteristics are analyzed. The results contribute to the operationalization and quantification of CAL to assess individual conceptualizations and competencies in terms of (voice-based) conversational agents.

Comparison: Empirical Vs Conceptional Framework of Aritifical Intelligence literary

In their conceptional framework of AIL, (Long and Magerko, 2020), postulated five themes (What is AI?; What can AI do?; How does AI work?; How should AI be used?; How do people perceive AI?) and 17 competencies. Four of these AIL-themes and 16 competencies could be transformed into the initial set of 29 items (see **Supplementary Table S10** for the complete list of items; APPENDICES). Although the explorative factor analysis of these 29 items revealed five factors, the factors represent a different content structure than the original AIL-framework has postulated. The overlaps and differences between the original and the empirical factorial structure are discussed below.

Following the newly developed CALS, we begin with **CALS 1** (storage and transfer of the smart speaker's data input; $\alpha = 0.788$). CALS one comprises six items, four of which initially belonging to the dimension "How does AI work?". This question summarizes large parts of the AI-literacy concept and encompasses nine out of 17 competencies, accordingly. CALS one covers three competencies (Learning from Data; Action and Reaction; Sensors). The two remaining items of CALS one are from the fourth dimension (How should AI be used?), asking for ethical considerations. However, our operationalization of ethics was limited to the passing of data from the interaction with the smart speaker. Consequently, CALS one refers to data input, storage, and transfer of the data input the user generates when interacting with the data. Considering the broad scope of ethics and anticipating this study's discussion, our limitation to data sharing seems to be too limited. Future studies will need to differentiate ethical considerations. **CALS 2** (smart speaker's functional principles; $\alpha = 0.690$) comprises six items, five of them belonging to the "How does AI work"-dimension, again. In terms of the 17 competencies, CALS two involves three: Knowledge Representation, Decision Making and Human Role in AI. The sixth item, however, belongs to the "What can AI do?"-Dimension, asking for areas of knowledge in which humans are superior. In sum, CALS two refers to two main aspects: the resemblance of human information processing (representation of knowledge reasoning, decision making) and the human role in terms of the development of the systems and their possible superiority. **CALS 3** (smart speaker's intelligent functions, learning abilities; $\alpha = 0.623$) includes five items, covering three dimensions and four competencies. One item belongs to the "What is AI"-dimension (competency: Recognizing AI), one to the "What can AI do?"-dimension (AI's Strength and Weakness), and three to the "How does AI work?"-dimension (Machine Learning Steps and Learning From Data). Summarizing this scale, CALS three covers competencies and conceptualizations referring to "intelligent" characteristics of smart speakers, their differentiation from standard speakers, their learning features (machine learning and learning from data), with one item also referring to the possible involvement of humans in data analysis. While CALS 1, 2, and three involve five to six items and show good to acceptable internal consistencies (particularly for this early stage of questionnaire development), **CALS 4** and **CALS 5** are of questionable numerical quality. To avoid false conclusions, we want to emphasize the approach of this study again: the aim was to gain the very first insights into the possibility of making AI-related conceptualizations and competencies measurable. Although we are aware of the shortcomings of the scales, we present the entire process as a first attempt to develop and to use a measuring tool of smart speaker-related literacy. The three items of **CALS 4** (smart speaker's reach; $\alpha = 0.337$) cover three competencies of three original dimensions: Interdisciplinarity of the "What-is-AI"-dimension refers to the multiple disciplines involved in the development of smart speakers; Imagine Future IA of the "What can AI do"-dimension asks for future features of smart speakers; and, Critically Interpreting Data of the "How does AI work?"-dimension enquiring if smart speakers can distinguish their users to process their inputs differently. In sum, CALS four asks for the potential of smart speakers by assessing the multiple facets contributing to their development, their capability in terms of adapting to current users, and their future capabilities. Finally, the three items of **CALS 5** (smart speaker's technological [surrounding] infrastructure; $\alpha = 0.345$) refer to the technological infrastructure, and smart speakers are embedded in. All items are part of the "How does AI work?"-dimension, covering the competencies Decision Making (dependence on internet connection), Machine Learning Steps (hardware used for data storage) and Sensors (sensor hardware). As a result, CALS five asks for the conceptualization of the technological infrastructure, smart speakers depend on.

To summarize the factorial analysis, referring to the original themes, the empirical structure does not seem to reflect the postulated structure closely. However, on the level of the 16 competencies, one might arrive at a different result: of the 16 competencies (29 items) that were entered into the factorial analysis, 11 competencies (23 items) are included in the final scales. Therefore, one can cautiously conclude that the newly developed CALS reflects the original framework, which has been transferred from AI to smart speaker literacy, quite convincingly. Except for Machine Learning, items of a certain competency are only ever taken up by one of the five CALS-scales. Concerning the very early stage of this process and the cautious interpretation of results, we carefully present the first attempt to operationalize and quantify conversational agent literacy.

First Insights into Construct Validity and Students' Performance in Conversational Agent Literacy Scale

When analyzing the correlations of CALS-scales and instruments measuring technology-related attitudes, results revealed substantial overlaps indicating first signs of a nomological net, the newly developed instrument is embedded: Affinity with Technology was significantly positively associated with CALS 1, 2 and 3, Commitment to Technology (total scale) with CALS one and CALS 2, and Attitudes towards Smart Speakers (total) was significantly negatively correlated with CALS 1, 2, and 5. Knowledge about smart speakers seems to be negatively associated with attitudes about these devices, indicating that a positive view of technology does not guarantee technological competencies—but quite the opposite, perhaps. Correlations of CALS Total confirm the significant correlations of the CALS-subscales. CALS 4, however, did not significantly correlate with any other construct, which could be interpreted as another indicator of its questionable quality. Or, it might point to the different scope of CALS 4. With its focus on the (future) potential and the reach of smart speakers, CALS four might be less close to attitudes towards technology. Future studies should widen the nomological net and incorporate more diverse constructs and variables into the analysis, such as experiences with technology, technological competencies, or psychological variables associated with technology-related competencies (among others: underlying motivations of usage and non-usage, personality traits such as openness to new experiences, or curiosity) (e.g., Jenkins, 2006; Literat, 2014; Porat et al., 2018; Perdana et al., 2019). Moreover, future studies should also consider associations with behavioral indicators. Additionally, the fifth AIL dimension ("How do people perceive AI?") could be considered in future CALS-versions.

Since the present data reveal correlations only, future studies should investigate hypotheses about predictors, moderators, and mediators of CAL.

After the CAL-scores were corrected for guessing, our sample ($N = 170$) revealed an intermediated level of conversational agent literacy. The minor interindividual differences seem to mirror the homogeneity of our predominantly student sample. A more heterogeneous sample would probably reveal more detailed interindividual differences. Remarkably, participants who own a voice assistant did not score higher in CAL, indicating that ownership does not guarantee competencies. Smart speakers are easy and intuitive to use and therefore accessible for broad user groups. However, this easiness might create rather positive attitudes and a deceptive impression of an "innocent" technology discouraging users from education. More informed and more critical operators would understand the opportunities digitization offers and would know how and when to use it or when to refrain from usage (e.g., Fast and Horvitz, 2017; Long and Magerko, 2020). Within the other application areas of more complex AI-related systems such as automatic driving, this gap between easy-to-use and underlying technological complexity might even increase. However, the interpretations of our results are still on a speculative level and call for more empirical data.

Limitations and Future Attempts

Conceptional work on AIL and the development of corresponding measuring instruments are in their early stages. This paper presents the first step towards a reliable and valid measure of CAL to allow very first insights into the more general concept of AI literacy. Future studies in this area should consider the following limitations.

The present paper investigates conceptualizations and competencies about smart speakers as one representative of voice-based conversational agents. Although the basic technological operation and interaction principles are transferable to further variations such as text-based CAs, these preliminary results are limited to smart speakers. Future studies should also involve further (voice-based) CAs or AI-related applications. When presenting the 29 items to our sample, we did not differentiate various devices and applications or a specific context but simply referred to smart speakers. Consequently, we do not know the participants' exact reference points (e.g., specific agents, specific domains such as medicine, commerce, assistance). Future studies should differentiate the devices, applications, and contexts of usage.

Moreover, different methodological approaches can be used to learn about user's conceptualizations and competencies, offering potentials for future studies. For example, participants can be observed when interacting with smart speakers in specific use cases. However, direct observations have limitations. First, researchers create arbitrary user interactions in a laboratory—particularly the usage of conversational agents in controlled lab studies is artificially and covers limited use cases. Second, the results are limited for the specific situation presented to the participants. Third, participants must show up in the lab limiting access for many user groups, which is an even severe issue in a pandemic. The aim of the present approach was a different one. We focus on getting insights about different aspects corresponding with the usage of conversational agents in general.

In the past, digital literacy was researched a lot. However, our analyses showed a lack of research regarding AI literacy—a concept with increasing importance. Other approaches have addressed this gap by making systems more self-explanatory. For a detailed discussion about the explainable AI systems, refer to (Doran et al., 2018; Goebel et al., 2018). In contrast, our approach refers to detecting misconceptions and lacks of competencies to better understand users and design user-centered learning or training programs.

The analysis of scales and items revealed potentials for improvement. Along with the five CALS scales, future studies should develop and test further complementing items. Particularly, the fourth and the fifth subscale indicate questionable internal consistency. Future re-analyses and improvements of CALS four should elaborate the principles of the reach, and the potential smart speakers have and will have in the future more profoundly. Therefore, additional items should be developed and tested. Moreover, our operationalization of ethical aspects was far too limited (see CALS 1) and needs to be substantially expanded.

Finally, the conceptualization of our items resulted in a 50–50 chance to simply guess the correct answer (on average: 2.5 points of five points). Future work should consider different conceptualizations and response formats such as multiple-choice questions instead of correct/incorrect questions. The interpretation of CALS scores offers first insights into different areas of conversational agent literacy and the performance levels in the different areas, which could be precisely addressed by training programs explainable AI approaches. Future studies could aim to analyze the performance levels of certain (user) groups such as children, older adults, technology enthusiasts, or skeptics to derive standard or norm values. In sum, more data need to be collected to improve both the scale itself and the use of the resulting performance levels.

In sum, as the starting point of this study was a specific conceptional framework (Long and Magerko, 2020) we neglected other conceptualizations of digital competencies, which involve additional definitions and domains. Thus, future studies should consider different concepts of digital literacy, AIL, and CAL (e.g., Jenkins, 2006; Literat, 2014; Vuorikari et al., 2016; Porat et al., 2018; Perdana et al., 2019). The approach presented by Ng (2012) might be promising since the social-emotional dimension extended previous measures by considering the interactivity of digital technology. Following the recent recommendations of the G20, future conceptualizations should include a more diverse set of skills of technical but also non-technical competencies (Lyons and Kass-Hanna, 2019).

Contribution

Artificial Intelligence will transform the way we work and live—involving other human beings and machines (e.g., Chetty et al., 2018; Lyons and Kass-Hanna, 2019; Long and Magerko, 2020). A recent concept paper of the "future of work and education for the digital age" think tank of the G20 stated: "Standardized assessment tools are essential to consistently measure digital literacy, identify gaps and track progress towards narrowing them, especially for the most vulnerable populations" (p.1). Furthermore: "The G20 is well-positioned to lead this process of developing comprehensive definitions, strategies, and assessment tools for measuring digital literacy. These efforts would include the diverse set of skills—technical and non-technical—that are and will be needed in the future" (Lyons and Kass-Hanna, 2019, p. 11). Similar statements of the EU, and other national governments emphasize the aim of the present paper to develop a first attempt of the empirically founded measuring instrument of (voice-based) conversational agents as an increasingly popular representative of an AI-related application. The deductive developmental procedure of the present paper ensures a theoretical embedding of the instrument as the underlying conceptional framework by Long and Magerko (2020) integrates findings of 150 recently published scientific articles and reports on the topic of AIL.

From an HCI perspective, standardized measurements allow us to gain deeper insights into various individual competencies and attributes to monitor the effects of digitization and the effects of the digital divide. The understanding and the conceptualization of the required competencies are presented as the first steps towards the conceptualization of "literate users" compared to the "illiterate users". To conceptualize these different user types, their different levels of technological competencies need to be analyzed and understood. Moreover, to distinguish between differently literate users, their competence levels need to be operationalized to allow standardized measures. As in other scientific areas, which refer to interindividual differences in competencies or attributes such as cognitive capacities, emotional states, or behavioral tendencies, for example, this study argues for the operationalization and quantification of CAL to allow the assessment of individual competencies in terms of (voice-based) conversational agents. Finally, standardized measurements can accompany user-centered evaluations of the rapidly growing numbers of platforms, which address competencies referring to digital, AI-related, or conversational technologies but lack a scientific standard of quality regarding underlying conceptions, measurements, and conclusions. Finally, reliable and valid diagnoses allow the implementation of user-centered training measures to develop users' digital competencies be it CAL or AIL (Chetty et al., 2018).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CW: concept and structure, theoretical background, scale development, conduction of the study, interpreting results. AC: concept and structure, theoretical background, scale development, analysis of the study, interpreting results.

FUNDING

The university library pay partly the open access publication fees.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2021.685277/full#supplementary-material

REFERENCES

Baacke, D., Kornblum, S., Lauffer, J., Mikos, L., Günter, A., and Thiele, A. (1999). *Handbuch Medien: Medienkompetenz. Modelle Und Projekte*. Bonn, Germany: MEDIENwissenschaft: Rezensionen/Rev. doi:10.17192/ep2000.2.2792

Baumeister, J., Sehne, V., and Wienrich, C. (2019). *A Systematic View on Speech Assistants for Service Technicians*. Berlin, Germany: LWDA. doi:10.1136/bmjspcare-2019-huknc.228

Belshaw, D. A. J. (2011). "*What Is "digital Literacy? A Pragmatic InvestigationTese (Doutorado). Departamento de Educação*. Durham: Universidade de Durham.

Berg, M. M. (2015). "NADIA: A Simplified Approach towards the Development of Natural Dialogue Systems." in Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Passau, Germany, Editors C. Biemann, S. Handschuh, A. Freitas, F. Meziane, and E. Métais (Springer-Verlag), 144–150. doi:10.1007/978-3-319-19581-0_12

Bühner, M. (2011). *Einführung in Die Test-Und Fragebogenkonstruktion*. Munich, German: Pearson Deutschland GmbH.

Burrell, J. (2016). How the Machine 'thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data Soc.* 3 (1), 205395171562251–12. doi:10.1177/2053951715622512

Chetty, K., Liu, Q., Gcora, N., Josie, J., Li, W., and Fang, C. (2018). "Bridging the Digital Divide: Measuring Digital Literacy." *Economics: The Open-Access. Open-Assessment E-Journal* 12 (23), 1–20. doi:10.5018/economics-ejournal.ja.2018-23

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences. Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, N.J: Lawrence Eribaum Associates. doi:10.4324/9780203771587

Covello, S. (2010). *A Review of Digital Literacy Assessment Instruments*. Syracuse, New York: Syracuse University, 1–31.

Cronbach, L. J., and Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychol. Bull.* 52 (4), 281–302. doi:10.1037/h0040957

Doran, D., Schulz, S., and Besold, T. R. (2018). "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives." in CEUR Workshop Proceedings, Vol. 2071. Bari, Italy: CEUR-WS.

Eurobarometer (2017). "*Special Eurobarometer 460: Attitudes towards the Impact of Digitisation and Automation on Daily Life*". Brussels: European Commission.

Fast, E., and Horvitz, E.. 2017. "Long-Term Trends in the Public Perception of Artificial Intelligence." 31st AAAI Conference on Artificial Intelligence, AAAI 2017, San Francisco, California, USA: Association for the Advancement of Artificial Intelligence.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. 5th ed. London: Sage Publications.

Franke, T., Attig, C., and Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *Int. J. Human-Computer Interaction* 35 (6), 456–467. doi:10.1080/10447318.2018.1456150

Gallardo-EcheniqueEliana, E., Minelli de Oliveira, J., Marqués-Molias, L., Esteve-Mon, F., Wang, Y., et al. (2015). Digital Competence in the Knowledge Society. *MERLOT J. Online Learn. Teach.* 11 (1), 1–16.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., et al. (2018). *Explainable AI: The New 42?* Cham, Switzerland: Springer, 295–303. doi:10.1007/978-3-319-9974010.1007/978-3-319-99740-7_21

Groeben, N., and Hurrelmann, B. (2002). *Medienkompetenz: Voraussetzungen, Dimensionen, Funktionen*. Weinheim and Munich: Juventa-Verlag.

Güneş, E., and Bahçivan, E. (2018). "A Mixed Research-Based Model for Pre-service Science Teachers' Digital Literacy: Responses to 'Which Beliefs' and 'How and Why They Interact' Questions. *Comput. Educ.* 118 (December 2017), 96–106. doi:10.1016/j.compedu.2017.11.012

Hadan, H., and Patil, S. (2020). "Understanding Perceptions of Smart Devices," in *Lecture Notes In Computer Science (Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics) 12063 LNCS*. Editors M. Bernhard, A. Bracciali, L. Jean Camp, S. Matsuo, A. Maurushat, P. B. Rønne, et al. (Cham, Switzerland: Springer), 102–121. doi:10.1007/978-3-030-54455-3_8

Hernandez, A. (2021). The Best 7 Free and Open Source Speech Recognition Software Solutions. Available at: https://www.goodfirms.co/blog/best-free-open-source-speech-recognition-software (Accessed March 17, 2021).

Howard, M. C. (2016). A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve?. *Int. J. Human-Computer Interaction* 32 (1), 51–62. doi:10.1080/10447318.2015.1087664

Janssen, J., Stoyanov, S., Ferrari, A., Punie, Y., Pannekeet, K., and Sloep, P. (2013). Experts' Views on Digital Competence: Commonalities and Differences. *Comput. Educ.* 68, 473–481. doi:10.1016/j.compedu.2013.06.008

Jenkins, H. (2006). "Confronting the Challenges of Participatory Culture: Media Education for the 21st Century," in *An Occasional Paper on Digital Media and Learning*. Editors D. John and T. Catherine (Chicago, Illinois: MacArthur Foundation).

Kaiser, H. F., and Rice, J. (1974). Little Jiffy, Mark IV. *Educ. Psychol. Meas.* 34 (1), 111–117. doi:10.1177/001316447403400115

Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., et al. (2019). *"Happy and Assured that Life Will Be Easy 10years from Now: Perceptions of Artificial Intelligence in 8 Countries*†*. New York, NY: ArXiv. Availablat: https://arxiv.org/abs/2001.00081.

Klüwer, T. (2011). From Chatbots to Dialog Systems. *IGI Glob.*, 1–22. doi:10.4018/978-1-60960-617-6.ch001

Kraus, M., Ludwig, B., Minker, W., and Wagner, N. (2020). "20 Assistenzsysteme," in *Handbuch Der Künstlichen Intelligenz*. Editors G. Gö, U. Schmid, and T. Braun (Berlin: De Gruyter), 859–906. doi:10.1515/9783110659948-020

Lau, J., Zimmerman, B., and Schaub, F. (2018). Alexa, Are You Listening?. *Proc. ACM Hum.-Comput. Interact.* 2, 1–31. doi:10.1145/3274371

Literat, I. (2014). Measuring New Media Literacies: Towards the Development of a Comprehensive Assessment Tool. *J. Media Literacy Educ.* 6 (1), 15–27.

Long, D., and Magerko, B. (2020). "What Is AI Literacy? Competencies and Design Considerations." in Conference on Human Factors in Computing Systems - Proceedings, Honolulu, HI, USA, 1–16. New York, NY: Association for Computing Machinery. doi:10.1145/3313831.3376727

Lyons, A. C., and Kass-Hanna, J. (2019). *The Future of Word and Educaton for the Digital Age Leaving No One behind: Measuring the Multidimensionality of Digital Literacy in the Age of AI and Other Transformative Technologies*.

McTear, M., Callejas, Z., Griol, D., McTear, M., Callejas, Z., and Griol, D. (2016). *The Conversational Interface*. Basel, Switzerland: Springer. doi:10.1007/978-3-319-32967-3_17

Meticulous (2021). "No Tit Available at:" Meticulous Market Research. Available at: http://www.meticulousresearch.com/. (Accessed March 15, 2021).

Neyer, F. J., Felber, J., and Gebhardt., C. (2012). Entwicklung Und Validierung Einer Kurzskala Zur Erfassung von Technikbereitschaft. *Diagnostica* 58 (2), 87–99. doi:10.1026/0012-1924/a000067

Ng, W. (2012). Can We Teach Digital Natives Digital Literacy? *Comput. Educ.* 59 (3), 1065–1078. doi:10.1016/j.compedu.2012.04.016

Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006). Measurement of Negative Attitudes toward Robots. *Interaction. Stud.* 7 (3), 437–454. doi:10.1075/is.7.3.14nom

Nunnally, J. (1978). *Psychometric Theory*. 2nd ed.. New York: McGraw-Hill. doi:10.2307/1161962

Oxbrow, N. (1998). Information Literacy - the Final Key to an Information Society. *Electron. Libr.* 16, 359–360. doi:10.1108/eb045661

Pütten, A. R-V. D. E. R., and Bock, N.. 2018. "Development and Validation of the Self-Efficacy in Human-Robot-Interaction Scale (SE-HRI)" 7 (3), 1–30. doi:10.1145/3139352

Perdana, R., Riwayani, R., Jumadi, J., and Rosana, D. (2019). Development, Reliability, and Validity of Open-Ended Testto Measure Student's Digital Literacy Skil. *Int. J. Educ. Res. Rev.* 4 (4), 504–516. doi:10.24331/ijere.628309

Porat, E., Blau, I., and Barak, A. (2018). Measuring Digital Literacies: Junior High-School Students' Perceived Competencies versus Actual Performance. *Comput. Educ.* 126 (July), 23–36. doi:10.1016/j.compedu.2018.06.030

Schwarzer, R., Bäßler, J., Kwiatek, P., Schröder, K., and Zhang, J. X. (1997). The Assessment of Optimistic Self-Beliefs: Comparison of the German, Spanish, and Chinese Versions of the General Self-Efficacy Scale. *Appl. Psychol.* 46 (1), 69–88. doi:10.1111/j.1464-0597.1997.tb01096.x

Shapiro, J. J., and Hughes, S. K. (1996). "Information Literacy as a Liberal Art Enlightenment Proposals for a New Curriculum." *Teaching Uncc.Edu* 31 (2), 1–6. Available at: https://teaching.uncc.edu/sites/teaching.uncc.edu/files/media/article-books/InformationLiteracy.pdf (Accessed March 15, 2021)

Siegert, I., Böck, R., and Wendemuth, A. (2014). Inter-Rater Reliability for Emotion Annotation in Human-Computer Interaction: Comparison and Methodological Improvements. *J. Multimodal User Inter.* 8 (1), 17–28. doi:10.1007/s12193-013-0129-9

Vuorikari, R., Punie, Y., Gomez, S. C., and Van Den Brande, G. (2016). *DigComp 2.0: The Digital Competence Framework for Citizens. Update Phase 1: The Conceptual Reference Model*. Luxembourg: Publications Office of the European Union.

Wardrip-Fruin, N. (2001). "Three Play Effects – Eliza, Tale-Spin, and SimCity. *Digital Humanities* 1–2. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.105.2025&rep=rep1&type=pdf.

Wienrich, C., and Latoschik, M. E. (2021). EXtended Artificial Intelligence: New Prospects of Human-AI Interaction Research. Available at: http://arxiv.org/abs/2103.15004. (Accessed March 15, 2021).

Zeng, E., Mare, S., Roesner, F., Clara, S., Zeng, E., Mare, S., et al. (2017). "End User Security and Privacy Concerns with Smart Homes This Paper Is Included in the Proceedings of the End User Security & Privacy Concerns with Smart Homes." in Thirteenth Symposium on Usable Privacy and Security (SOUPS), no. Soups: 65–80, Santa Clara, CA, USA, Berkeley, CA: USENIX Association.

Zhang, B., and Dafoe, A. (2019). Artificial Intelligence: American Attitudes and Trends. *SSRN J.*, 1–110. doi:10.2139/ssrn.3312874

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fcomp.2022.791704

Check for updates

"Alexa, You're Really Stupid": A Longitudinal Field Study on Communication Breakdowns Between Family Members and a Voice Assistant

Lina Mavrina[1]*, *Jessica Szczuka*[2], *Clara Strathmann*[2], *Lisa Michelle Bohnenkamp*[1], *Nicole Krämer*[2] *and Stefan Kopp*[1]

[1] Social Cognitive Systems Group, Centre for Cognitive Interaction Technology (CITEC), Bielefeld University, Bielefeld, Germany, [2] Social Psychology, Media and Communication, University of Duisburg-Essen, Duisburg, Germany

OPEN ACCESS

Edited by:
Roberto Therón,
University of Salamanca, Spain

Reviewed by:
Ronald Böck,
Otto von Guericke University Magdeburg, Germany
Michela Patrizi,
Sapienza University of Rome, Italy

***Correspondence:**
Lina Mavrina
lvaronina@techfak.uni-bielefeld.de

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

Received: *08 October 2021*
Accepted: *04 January 2021*
Published: *27 January 2022*

Citation:
Mavrina L, Szczuka J, Strathmann C, Bohnenkamp LM, Krämer N and Kopp S (2022) "Alexa, You're Really Stupid": A Longitudinal Field Study on Communication Breakdowns Between Family Members and a Voice Assistant. Front. Comput. Sci. 4:791704. doi: 10.3389/fcomp.2022.791704

In this paper, we present the results of our long-term study on use of a voice assistant (Amazon Alexa via Amazon Echo Dot) in nine families with children and no previous experience with this technology. The study was conducted over the course of 5 weeks during which the families could interact with the device freely. Three house visits were made to collect empirical data from the adult participants in form of questionnaires. Additionally, conversational data from log files of the voice assistant were obtained. These data were annotated and analyzed with a focus on communication breakdowns during human-assistant interaction. We investigate user behavior for both adults and children in such situations, its reasons and consequences for user satisfaction. This article provides qualitative analysis of three particularly interesting breakdown cases, as well as statistical analysis along several hypotheses and research questions combining empirical and conversational data. Described cases of communication breakdown illustrate findings from existing literature on the topic. The statistical analysis paints a mixed picture, however, it helped us identify further avenues for research, some of which can be explored with our data set in the future. We found a significant negative effect of the number of abandoned failed requests on user satisfaction, contrary to the number of successfully repaired requests that had no influence on user satisfaction. We discovered that users are more inclined to use reformulation as repair strategy when they do not perceive the emergence of miscommunication as their fault. We could not identify a significant effect of internal reasons for the choice of other strategies, so we suggest that situational clues such as the immediate response of the voice assistant are more important for the choice of repair strategy. Our results also hint that users distinguish between repair strategies differently, as the self-perceived frequency of repetitions and abortions of requests were found to be positive predictors for the use of reformulation-based strategies. With regards to the long-term aspect of the study, use of repetition as a repair strategy by both children and adults significantly decreased with time, no other changes were found for other strategies. Additionally, no significant impact of age on the choice of repair strategy was found, as well as no interaction effect between age and time.

Keywords: voice assistants, conversation analysis, miscommunication, communication breakdowns, human-machine interaction, repair strategies

1. INTRODUCTION

Voice assistants (VAs) continue to become part of the daily life for more and more people. Apart from their integration in smartphones, the sales of smart speakers are high and predicted to rise even more in the future. Tenzer (2021), for instance, puts the amount of smart speakers to be sold worldwide in the year 2025 at 205 million devices. It is therefore important to understand the phenomenon of voice assistants, as for many people it is or will be their major experience in speech-based human-machine interaction that will influence their perception, expectations and behavior toward this technology. Analyzing current patterns of use, problems and shortcomings of voice assistants can also help us build truly conversational agents for a variety of tasks beyond question answering. It is also important to investigate these aspects with regards to various demographic groups, such as children or the elderly, as they have specific qualities and requirements that demand specialized approaches to technology design.

As shown in Szczuka et al. (2021), children specifically can be counted as a vulnerable group of users who might lack knowledge and understanding to be sufficiently informed about the functionality of voice assistants and various legal and ethical consequences of their use with regards to topics such as data security and processing. They also often face technical difficulties during interactions, as children's speech contains more pauses, repetitions, speaker-initiated repairs, ungrammatical utterances and other inconsistencies than adult speech. Additionally, children have higher-pitched voices due to a shorter vocal tract. All of these features decrease the performance of common speech recognition systems, as shown in Kennedy et al. (2017). While there are efforts to increase speech recognition performance to account for children's acoustic variability, such as Dubagunta et al. (2019), Gale et al. (2019), Wu et al. (2019), or Shivakumar and Georgiou (2020), other above-mentioned inconsistencies can still prevent speech-based systems from understanding or fulfilling the communicative intent of a child in spite of the low word recognition error rate, which could be observed in the study by Lovato et al. (2019). Despite all these issues, a lot of children have access to and engage with voice assistants in their everyday lives. According to a survey cited in Kats (2018), more than nine in ten children aged 4–11 have access to a VA and out of these, 26 per cent engage with it between 2 and 4 h a week and 20 per cent more than 5 h. This study also found that an overwhelming majority of children used voice assistants on smart speakers whereas only about a half of teenagers did that. These figures highlight the relevance of research on children interacting with voice assistants and the impact these interactions have on their lives.

One has to additionally consider that when things go wrong in an interaction with a voice assistant, the burden to repair the communication breakdown and to ensure the understanding of own utterances generally falls to the users. Commercially available VAs can be seen as "black boxes" that provide little indication as to the source of the breakdown and force users to rely on their own experiences, intuition and expectations in order to find a solution to the problem, which is supported by various studies such as Luger and Sellen (2016), Myers et al. (2018), Porcheron et al. (2018), Beneteau et al. (2019), Cho and Rader (2020).

We have conducted a long-term study during which families with no previous VA ownership interacted with one (Amazon Alexa) over the course of 5 weeks. Google (Kleinberg, 2018) reports that parents use smart speakers more than non-parents as part of their daily routine and for multitasking, which supports our idea that families with children are an important environment where voice assistant technology is and will be used in the future and therefore represents an important subject for scientific research. During the course of our study, we acquired both conversational data as recorded in Alexa log files and empirical data from questionnaires filled out by the participants. By combining these data we would like to understand what internal factors might influence the users' choice of behavior in miscommunication situations in the under-informed context of interaction with a "black box", as in our opinion there is still a research gap there. Moreover, there is little literature discussing the possible differences between children and adults in the context of miscommunication in long-term interactions with VAs, something that we, too, can address with our data. This is important as the communicative experience, patterns of use and perception of voice assistants vary between children and adults, as can be seen, for example, in Garg and Sengupta (2020). They report that while adults primarily use smart speakers for listening to music and automating tasks, children are more likely to seek knowledge, play or engage in small talk or emotional conversations with the device. These latter kinds of interactions might lead to communication breakdowns more frequently, as they can require capabilities beyond simple request-response format, such as context-awareness and memory over past interactions. The authors also found younger children to have a tendency to ascribe human-like characteristics to the device which might further influence their communicative behavior.

Additionally, we are also interested in the impact that communication breakdowns and the success of their resolution have on user satisfaction with the assistant. As will be shown in the next section, over time users often reduce their interactions with a VA to a minimum (Cho et al., 2019) and one of the reasons for it may be the mismatch between the expectations of the users with regards to the conversational abilities of the devices and the reality. By understanding where exactly these capabilities fail and how communication between the user and the VA can be improved, the feeling of disappointment can be reduced in users which may lead to a better adoption of the technology in the future.

In the next section, we will explain how all these research questions tie in with the existing research on voice assistants and how our specific research questions emerge from it. Further on, we will provide a more detailed description of our study and the process of data preparation for statistical analysis and then present the results of both qualitative and quantitative analysis. Finally, we will discuss our results and their implications for the design of voice-based interaction technology.

2. RELATED WORK AND RESEARCH QUESTIONS

2.1. Expectations With Regards to Conversational Abilities of VAs and Their Influence on User Satisfaction

Based on the results of their long-term study of Amazon Echo usage in various households and pre-existing models of technology adoption, Cho et al. (2019) identified five phases of Amazon Alexa use consisting of pre-adoption, adoption, adaptation, stagnation, and acceptance. These phases describe a pattern of user's journey from having vague and mostly inaccurate ideas about the VA pre-adoption over exploration ending in disappointment, lower expectations, and negative role recognition of the assistant up to the loss of interest resulting in minimal use for simple functions or, in extreme cases, abandonment of the device. Among possible reasons for the establishment of this pattern is the mismatch between the expectations for conversational abilities of the voice assistant and its actual capabilities. Luger and Sellen (2016) and Cho et al. (2019) report that users with more experience with technology have more realistic ideas about the assistant, but those with less experience would mostly draw their expectations from familiar characteristics of human-human interaction due to insufficient information provided by the system itself, which eventually leads to disappointment at the lack of human-like conversational abilities in the assistant, such as its inability to acknowledge context across several temporally adjacent turns or understand more natural colloquial ways of communication. It could be argued, for example, based on findings of Porcheron et al. (2018) that voice assistants are currently not designed to act as conversation partners, but rather provide one-shot request-response interactions which users embed into conversational situations within human-human interaction domain, e.g., in the family context. It leads to a differing treatment of voice assistants and characterization of such interactions as can be seen from the interviews in Clark et al. (2019) where users mostly describe concepts relating to conversations with virtual agents in a functional way in contrast to human-human conversations that are described in more social terms. Along with these differences in perception, users may also change the way they speak to the assistant by, for instance, simplifying their utterances, removing excess words that do not function as keywords, or altering their prosody, a phenomenon Luger and Sellen (2016) call the "economy of language". Siegert and Krüger (2018) investigated in their study differences in speaking style between human-human and human-assistant communication. All of their participants reported changes to their speaking style at some point during the interaction with the VA with regards to loudness, intonation and rhythm of speech. Analysis of the objective features of speech in general supported these self-assessments, yet differences in the amount of characteristics altered could be observed between types of tasks, suggesting overall variability of user speech style adaptation.

Even then, communication breakdowns still occur during interaction. Beyond the initial playful exploration phase, users are not particularly forgiving of failed interactions, which negatively affects the frequency of assistant use, as was reported in Luger and Sellen (2016). However, they also state that the more technically savvy users were more tolerant in cases of miscommunication and more persistent in their attempts to accomplish their tasks. They were also more likely to identify the causes of communication breakdowns, unlike less experienced users who were more likely to blame themselves and experience negative feelings as a result. The self-attribution of blame was also observed by Cho and Rader (2020) in cases where the voice assistant provided no clues as to the cause of miscommunication by giving the user a neutral non-understanding reaction (e.g., "Sorry, I don't know how to help with that yet, but I'm still learning").

Our hypothesis **H1** assumes the existence of these observations in our data. With the help of statistical methods, we would like to investigate how a set of particular internal characteristics of the user may influence user's success in the resolution of communication breakdowns.

H1*: User's affinity for technology, the party they attribute the emergence of a communication breakdown to, and the emotions they experience when such a breakdown occurs serve as predictors for the number of abandoned failed requests (**H1a**) and successfully repaired requests (**H1b**). Hereby, we expect users with lower affinity for technology, users that attribute communication breakdowns to own mistakes and users experiencing negative emotions during breakdowns to abandon requests more frequently and achieve successful resolution of miscommunication less frequently.*

Other studies with adult participants such as Jiang et al. (2015), Kiseleva et al. (2016) and the analysis of Amazon Echo reviews by Purington et al. (2017) show that various aspects connected to communication breakdowns, such as the quality of speech recognition, occurrence of technical errors or effort required to accomplish a task have an effect on user satisfaction with speech-based human-machine interactions. Purington et al. (2017) report that users who mentioned technical issues with the device in their reviews were significantly less satisfied with it. Jiang et al. (2015) found a significant positive correlation between intent recognition quality and user satisfaction across various tasks and a lower, but also a significant positive correlation between speech recognition quality and user satisfaction for certain types of tasks, such as device function and web search. The article of Kiseleva et al. (2016) also suggests that impact of certain aspects on user satisfaction varies based on the type of the underlying task, as they report a low correlation between user satisfaction and task accomplishment in use cases of information seeking, but high correlation in use cases of device control. Additionally, they found a strong negative correlation between perceived effort spent accomplishing a task and user satisfaction. Correspondingly, our hypothesis **H2** assumes an impact of communication breakdown resolutions on user's satisfaction with the voice assistant.

H2: *The number of abandoned failed requests negatively (**H2a**) and successfully repaired requests positively (**H2b**) influence user's satisfaction with the voice assistant.*

Hereby, we assume abandoned requests to have a negative influence on user satisfaction as they correspond with unresolved issues in VA's understanding of user intent and potentially mean that the user deemed the cost of repairing the breakdown too high. We also hypothesize that successfully repaired requests have a positive effect on user satisfaction as they correspond with successfully accomplished tasks, especially if the user deemed the effort of repairing the breakdown acceptable in these situations.

2.2. Types of Communication Breakdowns in Speech-Based Human-Machine Interaction and Corresponding User Behavior

Various papers have addressed communication breakdowns in speech-based human-machine interaction in general and voice assistants in particular, for example, Stent et al. (2008), Jiang et al. (2013), Myers et al. (2018), Beneteau et al. (2019), Cho and Rader (2020), Motta and Quaresma (2022). These studies investigate the relationship between different strategies that users employ in the event of miscommunication and the types of underlying errors or system's responses to the user. System's responses are widely recognized as an important resource for the users to identify communication breakdowns and a fitting course of action.

The results of these studies are often consistent with each other, as Stent et al. (2008), Jiang et al. (2013), Myers et al. (2018) and Motta and Quaresma (2022) observe hyperarticulation or prosodic changes as one of the most common strategies employed by users, especially when they are faced with errors in speech recognition. Other commonly seen strategies mentioned in Jiang et al. (2013), Myers et al. (2018), Porcheron et al. (2018), Beneteau et al. (2019) and Motta and Quaresma (2022) include simplification of utterances, variations of the amount of information given to the system, semantic and syntactical adjustments to queries, repetition of commands. Users tend to explore these varied strategies when faced with errors seemingly not directly related to speech recognition, such as the system being unable to recognize the communicative intent of the user or to follow through on the recognized intent or performing a task that is different from the requested one.

Two of the above-mentioned studies have a unique focus. Beneteau et al. (2019) investigate interactions with a voice assistant as a joint family activity and look at various types of discourse scaffolding employed by family members to support each other during resolution of communication breakdowns, such as giving instructions or redirecting the interaction back to the desired conversation topic. Unfortunately, we cannot systematically examine this phenomenon with the data available through our study, since it requires extra recordings of conversations between family members that are not documented by the assistant. However, we could observe some cases of joint miscommunication repair in the conversational data, examples of which will be shown in the section 4.1.

Cho and Rader (2020) investigate the responses Google Home gives in miscommunication situations with regards to their advancement of conversational grounding between the system and the user and their helpfulness for the achievement of user goals. For this, they do not focus on specific repair strategies, but rather categorize user utterances into "advancing" and "backtracking" depending on whether the utterance seems to move the conversation closer to task completion or not. They find that the most common type of response observed from the Google Home, which can be categorized as the "Cannot Help" response (e.g., "Sorry, I don't know how to help with that yet, but I'm still learning"), provides the least clues to the user for the resolution of the communication breakdown, despite being a correct response in case of an error. On the contrary, "Unrelated" responses that are not connected to user's request and often occur when the system acts on own misunderstanding may provide more clues to the user for the cause of miscommunication or invite experimentation with regards to repair strategies.

While out of scope of this paper, another type of miscommunication can occur with the voice assistant, namely, accidental activation of the device and, subsequently, a false assumption about an unrelated user utterance being a request. This may lead to issues, especially, as modern VAs have extensive capabilities to make purchases, contact people via calls, etc. We have seen examples of falsely assumed requests that Alexa acted upon in our data, however, we do not focus on the analysis of such situations here. Work by other researchers is concerned with identifying potential reasons for accidental activation of the VA and how the amount of such cases can be decreased through improved addressee detection mechanisms. For example, Siegert (2021) analyzed examples of audio recordings of utterances incorrectly triggering Alexa and found that higher variety in intonation leads to more accidental activations.

Researchers are in agreement that users require better guidance from voice assistants in order to successfully accomplish tasks and resolve communication breakdowns, for example, Porcheron et al. (2018), Beneteau et al. (2019), Cho and Rader (2020) and Motta and Quaresma (2022), that currently available state-of-the-art systems are often unhelpful and act as "black boxes", leaving users under-informed. We pose our first research question **RQ1** to investigate the impact of internal factors such as personal estimation of causes for communication breakdowns and perception of own preferred cause of action in these situations on the actual behavior of users. While different papers mentioned in this section use slightly different ways to describe repair strategies in interaction with VAs, in our research, we decided to base our classification on the system used in Beneteau et al. (2019). Hereby, we grouped breakdown repair strategies into three categories, namely reformulation (group A), repetition (group B) and changes in articulation (group C) that are used for the research questions **RQ1a-c** respectively. More information on the repair strategies we identified in the conversational data can be found in section 3.2.

RQ1: *Do user's attribution of communication breakdowns to specific factors and issues, along with their perception of their own behavior when such situations occur predict their choice of repair strategy (**RQ1a-c**) or abandonment of the query before attempting any repair in the first place (**RQ1d**)?*

2.3. Children and Communication Breakdowns in Interactions With Voice Assistants

Prior research such as Gallagher (1977) suggests that children in general are persistent when it comes to establishing understanding and attempting communicative repairs. While adults were not the primary users in the study conducted by Cheng et al. (2018), but rather supported their children and thus did not have an immediate interest in pursuing task accomplishment, they gave up on the repair attempts faster than the children, perhaps because they were quicker to attribute the communication breakdown to a technical failure that could not be resolved. Further research is required to study potential differences in repair persistence between children and adults in equal scenarios.

When it comes to interactions with voice assistants, children employ a variety of repair strategies. Studies such as Lovato and Piper (2015), Druga et al. (2017), Cheng et al. (2018) and Yarosh et al. (2018) agree that the most commonly employed strategies are repetition of the initial query and increase in volume. However, children will also reword their requests or supplement them with additional context, especially if their initial repair attempts have failed. Lovato and Piper (2015), Druga et al. (2017), Cheng et al. (2018), Beneteau et al. (2019) and Garg and Sengupta (2020) also discuss the importance of discourse scaffolding mentioned in the previous subsection with regards to children and their repair strategies, as adults can encourage children to try out different approaches through suggestion or modeling and reinforce specific behaviors by giving their approval. However, Garg and Sengupta (2020) report that in their long-term study children learnt the interaction principles and required less help when using voice assistants after 2–3 months. It is also suggested that some of these scaffolding functions could be relegated to the assistant itself, the successful realization of which can be seen in Xu and Warschauer (2019) where the system was able to provide re-prompts that would constrain children's response options in case of miscommunication or to use follow-up questions to scaffold correct pronunciation of words and prevent communication breakdowns in the future.

Our research question **RQ2** ties in with this research and investigates the effect of age (children vs. adults) and time spent using a voice assistant on user behavior and choice of repair strategies in case of communication breakdowns. Here again, we operate within the three categories of repair strategies A to C for research questions **RQ2a-c** respectively.

RQ2: *Does user's age and the length of time they have interacted with a voice assistant predict their choice of repair strategy (**RQ2a-c**) or abandonment of the query before attempting repair (**RQ2d**)?*

3. STUDY DESCRIPTION AND DATA ANALYSIS

In order to investigate our hypotheses and research questions, we conducted a study during which families with children and no previous ownership of voice assistants interacted with an Amazon Alexa over the course of 5 weeks. This section describes the exact qualities of the data sample we obtained and the measures that were calculated from both empirical and conversational data.

3.1. Sample

In our study, ten families with twelve children received an Alexa Echo Dot device for 5 weeks. The study was conducted in Germany between mid-January and end of February 2020. Unfortunately, for one family, no log file data were retrieved. Therefore, calculations were done with nine complete datasets. Recruiting was carried out via local Facebook groups, the distribution of flyers and personal contacts. Families who wanted to participate must not have had previous voice assistant experience and must at least have one child between 6 and 12 years of age living in the household. Out of the nine families, one father was a single parent, whereas the remaining eight lived in heterosexual relationships, which adds up to a total of 17 adults. On average, the parents were 41.17 years old (SD = 5.37, *Range*: 31–48). In total, seven people had a degree below the German Abitur (A-levels, formerly 9 years of secondary school), whereas nine people finished their Abitur. One person had obtained a university degree. Thus, gender and education amongst adults were well-balanced and largely representative of the German population. Across the families, there was a total of 11 children, who were on average 8.91 years old (SD = 1.70, *Range*: 6–11). Further information on the age and gender of the children can be found in **Table 1**.

Throughout the 5 weeks, three home visits were carried out. The first one included the device's installation, encouragement of a first interaction, and running of the first questionnaire. Here, the families also provided their consent for participation and log file retrieval after they had been briefed appropriately. Afterwards, the parents filled in an online questionnaire including sociodemographic data and affinity for technology. In the second and third session they rated questions regarding situations when misunderstandings with the voice assistant occurred as well as their satisfaction with the device. If not differently stated, participants were asked to rate the questions as a mean for the entire household. These questionnaires can be found in **Supplementary Materials** to this paper. Further questionnaires which are of no relevance to this work were

TABLE 1 | Gender and age of children participating in the study.

Gender	Frequency
Female	7
Male	4
Age	**Frequency**
6	1
7	2
8	1
9	2
10	3
11	2

included but will not be described here in more detail. The study procedure was approved by the University of Duisburg-Essen ethics committee.

3.2. Annotation of Conversational Data

With participants' consent, we acquired access to the log files of their Amazon Echo devices. These files provide audio recordings of user queries, textual representations of these queries as recognized by the voice assistant and system's responses to the queries. These data were annotated using ELAN 6.0 annotation software (Hellwig and Sloetjes, 2021), a screenshot of which can be found in **Figure 1**. There was a total of two annotation cycles (done by one and two annotators, respectively).

In the first annotation cycle, the audio recordings of user queries were manually transcribed to facilitate the comparison with the textual representation given by Alexa, and then each request was annotated with the information about the speaker as perceived from the audio. Thus, the annotation tier "AnnT"contains the transcription of the annotator and the tier "SysT"—the transcription of the same query as provided by Amazon Alexa. Regarding the speaker, the tier "Spr" contains the perceived gender of the person, represented by M (male) or W (female), and the age distinction between an adult (E) and a child (K). The speakers are also numbered within their gender-age category, e.g., KW1 and KW2 for "female child #1" and "female child #2". In addition to this, the tier "AnnS" was used to indicate the certainty of the annotator regarding the classification of the speaker on a scale from 1 (least certain) to 10 (absolutely certain). Alexa's responses were recorded in the tier "SysA". The tier "TS" was used to annotate the timestamp of the interaction.

The second annotation cycle was focused solely on communication breakdowns. For this, conversational data were segmented into temporally adjacent and thematically consistent interaction episodes (blocks consisting of one or more request-response pairs). Out of these, only episodes containing communication breakdowns were selected. Each of these episodes started with a request-response pair where miscommunication occurred and ended with a request-response pair corresponding to either a successful resolution of the breakdown and fulfillment of the user's request or the abandonment of repair attempts by the user. These episodes were then subjected to the second cycle of annotation. The paper by Beneteau et al. (2019) served as basis for the annotation scheme here, as it provided a comprehensive overview over repair strategies seen in conversational data of an experiment with conditions similar to ours.

Again, multiple tiers were defined during the annotation. The tier "SprAct" was used to annotate the characteristics of the speaker's request. A full overview can be found in **Table 2**. Characteristics describing conversational repair strategies were partially taken from Beneteau et al. (2019) and additional ones were defined to accurately represent the observations in our data. To facilitate calculations concerning our research questions, we then combined repair strategies into three supercategories, namely reformulation (including lexical, syntactical and semantic adjustments, termed "group A"), repetition (termed "group B"), and changes in articulation (including increased volume, prosodic changes and overarticulation, termed "group C"). These groups are denoted by different colors in **Table 2**. The tier "SysAct" was used to annotate how the Echo Dot responded to the speaker. For this tier the response types "acting on misunderstanding," "neutral clarification response," and "specific clarification response" from Beneteau et al. (2019) were used. Two more response types were added: "no response" and "proper response." A description of these labels can be found in **Table 3**. The tier "Skill" was used to indicate whether the user was interacting with a third-party Alexa skill or not.

As the second annotation cycle involved two annotators, inter-annotator agreement had to be determined. For this, a set of conversational data including two families was annotated by both annotators and then average Fleiss' κ for these families was calculated, once for the annotation of speaker's speech acts ($\kappa = 0.685$) and once for the annotation of system response types ($\kappa = 0.815$). According to Viera and Garrett (2005), these values indicate substantial agreement between annotators.

3.3. Measures

3.3.1. Failed and Successfully Repaired Request Blocks

First, we assessed the number of requests a family made during the 5 weeks of the study by investigating the respective log files (M = 518.90, SD = 275.08, *Range*: 166–1,087). Then, as mentioned in the section 3.2, these requests were grouped into interaction episodes, consisting of one or many request-response pairs connected by time and topic. For hypotheses **H1a** and **H1b**, we defined successfully repaired request blocks as the number of successful interaction episodes wherein at least one repair attempt has been done. Hence, requests which were successful directly are not included. Failed blocks were defined as those interaction episodes where a request block ends with an unsuccessful request whereupon the user abandons (further) repair attempts. To make the number of failed and successfully repaired blocks comparable among the families, we calculated the number of request blocks per 100 requests (successfully repaired blocks: $M = 1.95$, $SD = 0.85$; failed blocks: $M = 10.82$, $SD = 3.55$).

3.3.2. Repair Strategies

For research questions **RQ1a-c** we wanted to investigate the repair strategies and the way in which they may be influenced by

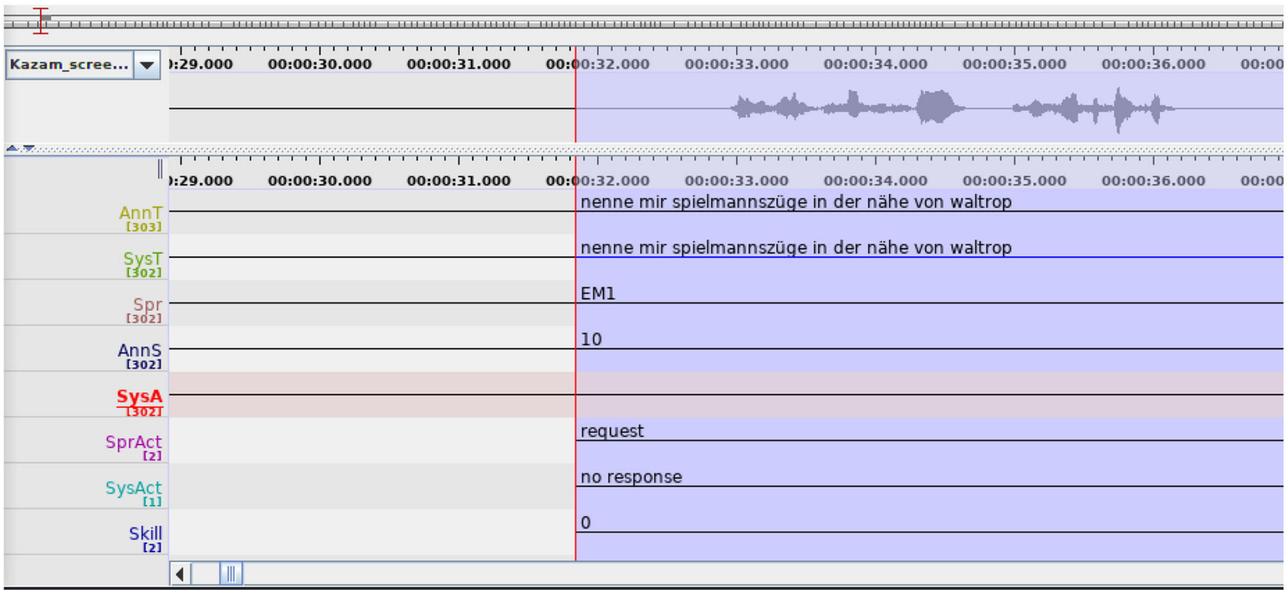


FIGURE 1 | Screenshot of ELAN annotation software from the second annotation cycle. All layers mentioned in section 3.2 can be seen here.

the perceived causes of errors and participants' reactions to them. For these calculations, repair strategies were not considered separately, but in three groups described in section 3.2 and **Table 2**. Hereby research question **RQ1a** corresponds to the repair strategy group A and so on, respectively. For the research question **RQ1d** we consider failed interaction episodes where no repair was undertaken by the user in the first place. Again, to make the number of times a specific strategy was used more comparable, we calculated a number of uses for each strategy per 100 "miscommunication and repair requests." These include all unsuccessful requests which required a repair strategy (M = 197.78, SD = 110.74, *Range*: 90–405). The descriptive values reveal that strategy A was most frequently used (M = 11.21, SD = 7.89, *Range*: 2.13–27.78), followed by strategy B (M = 1.92, SD = 1.25, *Range*: 0–3.39) and strategy C (M = 1.49, SD = 1.91, *Range*: 0–4.92). On average, 1.95 out of 100 miscommunication situations were left unrepaired (SD = 1.08, *Range*: 0.26–4.13). For research questions **RQ2a-d** we furthermore made a distinction between children and adults as well as the two phases of measurement, i.e., between the first and second (MP1), and between the second and third (MP2) house visit (for descriptive values see **Table 4**).

3.3.3. Affinity for Technology

To measure participants' affinity for technology, we used three items inspired by two subscales of the TA-EG by Karrer et al. (2009), namely enthusiasm for technology and technical competencies: "*We like integrating new electronic devices into our everyday family life*", "*Members of my family know most of the functions of the electronic devices we own (to the extent they can understand them based on their age)*", and "*Compared to our social environment, we are more open to the use of electronic devices in everyday family life*". Testing for reliability, Cronbach's Alpha yielded a score of 0.74, which indicates a good scale reliability (Streiner, 2003). Participants rated the items on a five-point Likert scale during the first home visit, ranging from 1 = *not true at all* to 5 = *fully true*. To have a value for hypothesis testing, we calculated a mean score ($M = 3.67$, $SD = 0.82$).

3.3.4. Reasons for Errors

In the questionnaire, to describe possible reasons for errors during interaction with the VA six items concerned with common problems were generated: "*I/The other person spoke too unclearly or too quietly*", "*I expressed myself/The other person expressed themself too ambiguously (e.g., by using a word that was misunderstood by the voice assistant)*", "*I/The other person did not activate the voice assistant correctly*", "*I have/The other person has made a request incorrectly*", "*My language skills/The language skills of the other person were not sufficient*" and "*I/The other person used a formulation/dialect which the voice assistant did not understand*". On a five-point Likert scale, participants rated how often these problems were reason for miscommunication during the second and third home visit (from 1 = *never* to 5 = *very often*). We first calculated a mean across the two sessions before we ran a factor analysis which indicated a one-factor solution including all items with a very good reliability (Cronbach's α = 0.87). A scale mean was retrieved for further calculations (M = 2.04, $SD = 0.68$).

3.3.5. Error Attribution

Furthermore, we used Nass and Moon's (Nass and Moon, 2000) consideration that people can either direct their social reactions to a system directly or to a person behind the system (such as a programmer) to generate three items concerning the attribution of blame when an error occurs. The participants were asked to think of an error that had occurred multiple times, and in

TABLE 2 | Annotations for request characteristics.

Request characteristic	Definition
Wake word	Alexa
Request	A request where miscommunication occurred
Falsely assumed request	An utterance misunderstood by Alexa as a request and reacted upon
Incorrect request	Faulty request where user has made a mistake
Follow-up request	A response to a follow-up question posed by Alexa
Semantic adjustment	Rewording of the previous request, so that it still concerns the topic of the original request, but carries different meaning
Lexical adjustment	Rewording of the previous request, so that it still carries the same meaning, but is expressed in different words
Syntactical adjustment	Modifying sentence structure of the original query
Repetition	Exact repetition of the previous request
Prosodic changes	Adjustments to the cadence of speech
Overarticulation	Hyperarticulation, i.e., exaggerated pronunciation of sounds
Increased volume	Raising the volume of the speaker's voice
Speaker Change	change of the active speaker compared to the previous request

Non-black colors indicate characteristics belonging to specific repair strategy groups. Blue—group A: reformulation-based strategies, green—group B: repetition, red—group C: articulation-based strategies.

TABLE 3 | Annotated system response types.

Response type	Definition
Acting on misunderstanding	Action or response based on misunderstood user input
Neutral clarification response	Generalized response indicating non-understanding
Specific clarification response	Clarification request related to the latest query
No response	Absence of an expected response from Alexa
Proper response	Correct response to the latest user query

what way they would attribute that error to the following parties: themself (*"I have made an error"*), the programmer (*"The voice assistant has not been satisfactorily programmed"*) and the voice assistant (*"The voice assistant itself has committed an error"*). Participants rated the items on a five-point Likert scale (from 1 = *not true at all* to 5 = *fully true*) during the second and third home visit. We calculated mean scores for each item across both sessions for further computations (self: $M = 2.40$, $SD = 0.70$; programmer: $M = 3.40$, $SD = 1.13$; voice assistant: $M = 2.55$, $SD = 0.93$).

TABLE 4 | Descriptive statistics of repair strategies by time and age group (MP1, first phase of the study; MP2, second phase of the study).

		MP1			MP2		
		M	*SD*	*Range*	*M*	*SD*	*Range*
Children	Strategy A (reformulation)	20.20	10.67	8–37	9.56	11.50	0–37
	Strategy B (repetition)	12.78	10.04	1–27	3.78	4.18	0–14
	Strategy C (articulation)	3.67	3.97	0–12	1.22	1.92	0–6
	No Repair	13.33	9.57	3–29	5.44	7.06	0–23
Adults	Strategy A (reformulation)	16.44	13.37	3–48	4.67	5.15	0–15
	Strategy B (repetition)	3.44	2.35	1–7	1.00	1.58	0–5
	Strategy C (articulation)	2.33	3.94	0–12	0.78	1.39	0–4
	No Repair	18.80	16.55	1–57	8.50	10.41	0–33

3.3.6. Emotions When Errors Occur

Based on the Differential Emotions Scale [DES; Izard et al. (1974)] we extracted emotions that may occur due to a communication breakdown with a VA and set up a scale of nine items: anger, disappointment, sadness, surprise, desperation, interest, motivation, annoyance, and amusement. Participants rated to what extent they experienced the respective emotions when an error occurred on a five-point Likert scale (1 = *not at all* to 5 = *very strongly*) both in the second and third home visit. For a factor analysis, mean scores of both sessions were calculated. Factor analysis revealed a two-factor solution, one including anger ($M = 1.55$, $SD = 0.80$), sadness ($M = 1.45$, $SD = 0.80$), and desperation ($M = 1.25$, $SD = 0.43$), and one encompassing the remaining six items. Since the first subscale was more unambiguous in its meaning, we decided to use it for further computations by calculating a mean value for each participant ($M = 1.42$, $SD = 0.56$).

3.3.7. Reaction to Errors

For **RQ1** we also wanted to investigate the predictive power of the reaction to errors for the repair strategy a user picks. Therefore, we generated four distinct items, each rated on a five-point Likert scale from 1 = *never* to 5 = *very often*, asking the participants how often they reacted to an error in the respective way: repetition (*"I repeat the request"*), rephrasing (*"I rephrase the request"*), getting help (*"I get help"*, e.g., from another person or the internet) and abortion (*"I abort the request"*). Again, participants rated these items during the second and third home visit, which is why for further calculations, a mean score was retrieved for each item (repetition: $M = 3.05$, $SD = 1.14$; rephrase: $M = 3.00$, $SD = 1.03$; get help: $M = 1.90$, $SD = 0.78$; abortion: $M = 2.70$, $SD = 1.18$).

3.3.8. Satisfaction With the Voice Assistant

Lastly, we wanted to explore the connection between satisfaction with the device and the number of aborted and successful request blocks. Participants rated their satisfaction with the device on a five-point scale, where a very sad emoji indicated absolute dissatisfaction and a very happy emoji indicated absolute satisfaction with the voice assistant. The scale was rated at the end of the experiment to measure the final level of satisfaction ($M = 3.56$, $SD = 1.01$, *Range*: 2–5).

4. RESULTS

In this section, we present the results of our study. First, three interesting cases of communication breakdown (translated into English from German) that occurred during the interactions within our study will be considered to illustrate the repair strategies and VA responses described in section 3.2. Then, quantitative results of our statistical analysis with regards to the hypotheses and research questions presented in section 2 will be shown.

4.1. Examples of Communication Breakdowns

Many interaction episodes involving miscommunication observed in the conversational data are rather simple and do not follow many steps until either resolution or abandonment. However, these examples were selected because they showcase some of the more complex behaviors, both on the side of the user and the assistant and invite informed speculation about possible reasons for such behaviors.

Consider the following example where an adult male speaker during his second week as a participant in the study attempts various reformulation-based repair strategies, such as semantic, lexical and syntactical adjustment (defined in **Table 2**) in order to get the desired information and accomplish his task.

Example 1

EM1: "Alexa, name me the member countries of the WHO"
A: "I am sorry, I don't know anything about that"
EM1: "Alexa, what is the WHO?"
A: "Browser Helper Objects are computer programmes that extend the functionality of the Internet Explorer. BHOs are used since version 4 of the Internet Explorer. They have direct access to the Document Object Model"
EM1: "Alexa, what is the World Human Organisation?"
A: "The World Human Vehicle Association was founded in 1976 in the US under the name of the International Powered Vehicle Association"
EM1: "Alexa, name me the member countries of the World Human Organisation"
A: *no response*
EM1: "Alexa, name me the member states of the World Human Organisation"
A: *no response*
EM1: "Alexa, what is the WHO?"
A: "Browser Helper Objects are computer programmes that extend the functionality of the Internet Explorer. BHOs are used since version 4 of the Internet Explorer. They have direct access to the Document Object Model"
EM1: "Alexa, the Health Organisation what, Alexa"
A: *no response*
EM1: "Alexa, name me the member states of the World Health Organisation"
A: "Unfortunately, I don't know anything about that"
EM1: "Alexa, what is the World Health Organisation?"
A: *no response*

Here, we can observe how the speaker tries to get information from Alexa with regards to the World Health Organisation (WHO) and its member states. Considering that the WHO currently has 194 member states, as is stated on its official website (World Health Organisation, 2021), this request is not well-answerable via a speech-based interface and a voice assistant would only be able to provide a fraction of the requested information within the interaction and perhaps a link to a website that could be followed by the user on a device such as a mobile phone. However, judging by the transcription of the user request provided in Alexa log files, it seems that the assistant could not fulfil the request due to speech recognition errors as two words, including the actual abbreviation "WHO," were mistranscribed. Yet Alexa only provides a neutral clarification response here, not indicating to the user what the problem might have been.

The first repair attempt the user undertakes is a semantic adjustment of his initial request, showing his willingness to settle for a more generalized type of information that might be easier to retrieve for Alexa. The request is misunderstood by the assistant and it acts on this misunderstanding by providing a definition to an object with a name similar to the requested one. This acting on misunderstanding might have provided additional clues to the speaker, which would be consistent with the findings of Cho and Rader (2020) described in section 2.2, and so the user decides to expand the abbreviation to eliminate it as a possible source of the communication breakdown. However, he gets the term wrong, calling the WHO the "World Human Organisation" and Alexa acts on this incorrect request as best it can, yet the moment the user reformulates his request to be more specific again while keeping the wrong name of the organization, the assistant fails to provide any kind of response.

After the repetition of the first repair attempt with the same result, the user corrects himself and uses the exact full name of the WHO in his next repair attempt. He also briefly tries dropping the sentence structure almost entirely and focusing his request on the keyword, though it could be questioned whether it was a deliberate syntactical adjustment at this point. From the log files we can see that out of the last three repair attempts of this block only the second one ("Alexa, name me the member states of the World Health Organisation") was correctly transcribed by the speech recognition. However, Alexa probably cannot deliver a response to query that specific and so this is the only request out of these three that it reacts to with a neutral clarification response. The other two failed requests are not reacted upon, since speech recognition errors made them into gibberish according to the log files. After these failed breakdown resolution attempts, the user gives up and abandons the request.

Next one is an example of parental help for resolution of miscommunication situations. The importance of such help for child-assistant interaction was mentioned in section 2.3. There are two speakers here in their second week of the experiment, a male child that is aided by a female adult speaker (presumably, his mother).

Example 2

KM1: "Alexa, what can I do together with my mom and sister?"
A: *no response*
EW1: "Alexa, events in [city name] today"
A: "Unfortunately, I am not sure how to respond"
EW1: "Alexa, what is happening today in [city name]?"
A: "'Today' means the present, the current times (the times in which we live)"
EW1: "Alexa, events in [city name]"
A: *starts a skill offering event recommendations in [city name]*

Here, the child expresses his request in a rather colloquial fashion, more appropriate for human-human conversation where additional context is either known to the conversation partner or can be acquired through additional questions. Alexa does not provide a response to this request, despite the fact that it was correctly transcribed by the speech recognition. That is when the adult person comes to help and attempts a semantic adjustment of the original request by creating a more precisely-worded query-style utterance and providing required context such as the name of the city and the fact that they are looking for events specifically and not, for example, ideas on what to do at home. It is not clear why Alexa cannot provide a response to this, as it only gives a neutral clarification to the user, though the request is transcribed correctly. Considering the last request-response pair in this block, this request might be overspecified and cannot be matched to the corresponding skill due to the use of the word "today". The user tries further repair strategies by adjusting the lexical content and the syntactic form of the request, yet it is misunderstood by Alexa and this misunderstanding is acted upon. In the next step, the user returns back to the query-style utterance consisting only of keywords which is ultimately successful and resolves the communication breakdown. This ties in with the phenomenon of the "economy of language" which is a special modification of linguistic qualities of speech employed by users of voice assistants as mentioned in Luger and Sellen (2016).

And here is the third example of two children, male and female, who are collaborating to receive information as to which animal on Earth is the most agile. This example comes from the third week of the experiment.

Example 3

KM1: "Alexa, what is the most agile animal in the world?"
A: "The sea wasp which is a kind of box jelly is the most venomous animal in the world"
KW1: "Alexa, what is the most agile animal in the entire world?"
A: "The smartest animal in the world is a dolphin. They can recognize emotions, plan own actions in advance and thus, solve complex problems. The capacity of their brains can potentially equal that of the human brain"
KW1: "Alexa, who is the most agile?"
A: "I am sorry, I don't know anything about that"
KW1: "Alexa, is there alien alien something?"
A: "On WELT, there is a documentary 'Aliens, the Message' tomorrow night at 1 a.m. and two episodes on N24 Doku starting from 1:45. By the way, you can ask me questions such as 'What movies are on tonight?' or 'Is there a crime series on today?' More information can be found in your Alexa app"
KM1: "Alexa, what is the most agile animal in the world?"
A: "The longest animal on Earth is a blue whale that is 27.5 m long"
KM1: "Alexa, what is the most agile animal in the world?"
A: "The smallest known animals in the world are the Myxozoa that belong to the phylum of Cnidaria and do not grow bigger than 20 μm. It is a parasitic lifeform living in the waters. The smallest of the Myxozoa is the Myxabolus shekel with the average size of 8.5 μm. The smallest vertebrate is the frog Paedophryne amauensis with an average size of approximately 7.7 mm"
KM1: "Alexa, what are the most agile animals in the world?"
A: "Hmm, I don't know, but I am still learning"
KM1: "Alexa, what is the most agile animal in the world?"
A: "I like seahorses"

In this example, we see two children supporting each other in the resolution of a communication breakdown without intervention or scaffolding from an adult. Apart from the third request that is unclear and may be incomplete or unintended, the rest are repair attempts utilizing various strategies, such as lexical and syntactical adjustments (although they are minor), speaker change and a repetition at an increased volume as could be recognized from the audio files (the sixth request).

Alexa for the most part, however, keeps providing the wrong kind of information with regards to animals. It appears to be mostly an issue of speech recognition, as in the first, the fifth and the last request-response pairs the transcription substitutes the word "gelenkigste" ("the most agile" in German) with similar-sounding "giftigste" ("the most venomous"), "längste" ("the longest"), and "beliebteste" ("the most beloved"), respectively, and Alexa acts on these misunderstandings by providing the corresponding answers. The seventh request is also mistranscribed, but the resulting utterance is non-sensical, and while the third one features the correct adjective, it lacks the necessary context that animals are the subject of this request. In both cases, Alexa provides a neutral clarification response. The second and the sixth request are transcribed more or less accurately by the system and feature at least the correct adjective, yet Alexa seems to have misunderstood the intention of the users again and delivered wrong results. After all these unsuccessful repair attempts, the children have abandoned their task.

These examples can be seen as an illustration of some of the aspects known from previous research as described in section 2. Sometimes family members help each other resolve communication breakdowns in interactions with voice assistants which happened in examples 2 and 3. Sometimes the assistant acting on misunderstanding instead of delivering a neutral message of non-understanding can provide clues as to what the reason for the breakdown might be, which could have been the case in example 1. Children in these interactions seem to have more difficulty to be understood correctly by the system, either due to errors in speech recognition or lack of context in their requests or unclear communication of their intent. However, these are just a few examples the reconstruction of which is limited by the characteristics of the conversational data that was collected during this study.

4.2. Quantitative Results

4.2.1. Predictive Power of Affinity for Technology, Error Attribution and Emotions on the Number of Failed and Successfully Repaired Requests

H1: *User's affinity for technology, the party they attribute the emergence of a communication breakdown to, and the emotions they experience when such a breakdown occurs serve as predictors for the number of abandoned failed requests (**H1a**) and successfully repaired requests (**H1b**). Hereby, we expect users with lower affinity for technology, users that attribute communication breakdowns to own mistakes and users experiencing negative emotions during breakdowns to abandon requests more frequently and achieve successful resolution of miscommunication less frequently.*

In our study, we were able to combine the results in form of the empirical and conversational data in order to investigate the connections between various variables present in these data sets. For this hypothesis specifically, we analyse the relationship of the measures described in sections 3.3.1, 3.3.3, 3.3.5, and 3.3.6. In order to do that, we calculate two multiple linear regressions where the affinity for technology scale's mean, the error attribution items and the subscale for negative emotions are used as predictors, whereas the number of failed and successfully repaired request blocks per 100 requests serve as criteria. For the first regression (**H1a**), homoscedasticity was not given, which is why a multiple linear regression with bootstrapping (1,000 samples) was calculated. Results revealed that there was no significant relation [$F_{(5,3)} = 1.883$, $p = 0.319$]. For **H1b**, all prerequisites were fulfilled, yet the calculations yielded no significance either [$F_{(5,3)} = 2.272$, $p = 0.266$]. Therefore, **H1a** and **H1b** need to be rejected.

4.2.2. Predictive Power of the Number of Failed and Successfully Repaired Requests on the Satisfaction With the VA

H2: *The number of abandoned failed requests negatively (**H2a**) and successfully repaired requests positively (**H2b**) influence user's satisfaction with the voice assistant.*

To investigate to what extent the satisfaction with the voice assistant (measure described in section 3.3.8) can be predicted from the number of abandoned failed and successfully repaired requests (measure described in section 3.3.1), we conducted a multiple linear regression. The number of failed and successfully repaired requests per 100 requests served as predictors, whereas the satisfaction with the device was the criterion. Heteroscedasticity was given, hence the regression was calculated using bootstrapping (1,000 samples). The model became significant with $F_{(2,6)} = 11.798$ and $p = 0.013$. According to Cohen (1988), the model had a high goodness-of-fit (adjusted $R = 0.76$). Taking a closer look at the predictors, only the number of abandoned failed requests was significant for the prediction of satisfaction ($p = 0.005$, $\beta = -0.309$), whereas the number of successfully repaired requests was not ($p = 0.361$). Thus, we can accept **H2a**: the number of abandoned failed requests has a negative effect on user's satisfaction with the VA.

4.2.3. Predictive Power of Reasons for Errors and Reactions to Errors on the Choice of Repair Strategy and Abortion of Requests

RQ1: *Do user's attribution of communication breakdowns to specific factors and issues, along with their perception of their own behavior when such situations occur predict their choice of repair strategy (**RQ1a-c**) or abandonment of the query before attempting any repair in the first place (**RQ1d**)?*

Combining our empirical and conversational data, we were able to investigate the relationship between the internal perception of reasons for breakdowns (measure described in section 3.3.4), user's own reactions to them (described in section 3.3.7) and the repair strategies they actually used during these situations (described in section 3.3.2). Here, we calculated four multiple regression analyses. For each of them, the mean of all "reasons for errors"-items as well as the four items regarding reactions to errors were included as predictors, whereas the criterion was the number of choices of specific repair behavior per 100 requests. Hereby, reformulation corresponds to the research question **RQ1a**, repetition to **RQ1b**, changes in articulation to **RQ1c** and absence of attempted repairs to **RQ1d**.

RQ1a lacked homoscedasticity, which is why the regression was calculated with bootstrapping (1,000 samples). It yielded significance [$F_{(5,3)} = 17.926$, $p = 0.019$] with an adjusted R of 0.91, which, according to Cohen (1988), indicates a high goodness-of-fit. Looking at the individual predictors, three out of five proved to be significant in predicting the frequency of making reformulations, namely the reasons for errors mean ($p = 0.015$, $\beta = -9.043$), *"I repeat the request"* ($p = 0.024$, $\beta = 3.173$) and *"I abort the request"* ($p = 0.010$, $\beta = 5.480$). The remaining two were not significant (*"I rephrase the request"*: $p = 0.953$; *"I get help"*: $p = 0.051$). **RQ1c** was the only other research question which lacked homoscedasticity, which is why here, too, bootstrapping was used (1,000 samples). For **RQ1b**, **RQ1c** and **RQ1d**, no significant relationship could be shown [**RQ1b**: $F_{(5,3)} = 0.337$, $p = 0.864$; **RQ1c**: $F_{(5,3)} = 3.709$, $p = 0.155$; **RQ1d**: $F_{(5,3)} = 0.200$, $p = 0.942$].

Thus, we could only find a significant relationship regarding **RQ1a** here. The questions about the reasons for miscommunication with a VA are a negative predictor for the number of reformulations that users applied during repairs. These questions are described in section 3.3.4 and can be seen to refer to certain actions of the speaker that might have led to a communication breakdown, such as *"I spoke too unclearly or too quietly"*. Meanwhile, the user's perception of them repeating their requests or aborting them in cases of miscommunication was found to be a positive predictor of them using reformulation.

4.2.4. Impact of Time and Age on Choice of Repair Strategy and Abortion of Requests

RQ2: *Does user's age and the length of time they have interacted with a voice assistant predict their choice of repair strategy (**RQ2a-c**) or abandonment of the query before attempting repair (**RQ2d**)?*

One of the advantages of our study is that we could observe developments over time due to two measuring points (MPs) and one of our goals was to investigate the influence of time on the choice/lack of repair strategy (as described in section 3.3.2),

TABLE 5 | Descriptive statistics for repair strategy B by time and age (MP1, first phase of the study; MP2, second phase of the study).

		N	*M*	*SD*
Repair strategy B per 100 requests (MP1)	Child	9	5.49	4.00
	Adult	9	3.49	1.96
	Total	18	4.49	3.22
Repair strategy B per 100 requests (MP2)	Child	9	2.75	1.81
	Adult	9	1.83	2.25
	Total	18	2.29	2.03

including the distinction between children and adults. Therefore, we calculated four repeated measures ANOVAs, each with the age group (adult/child) as within-subjects factor, and the respective strategy (reformulation: **RQ2a**; repetition: **RQ2b**; changes in articulation: **RQ2c**; no repair attempt: **RQ2d**) per 100 requests as dependent variable. Testing for normal distribution with the Shapiro-Wilk test ($\alpha = 0.05$), we yielded significant results for reformulation (MP1: $p = 0.027$; MP2: $p < 0.001$) and changes in articulation (MP1: $p = 0.009$; MP2: $p < 0.001$). As studies have shown, though, that repeated measures ANOVAs are largely robust against effects of normal distribution violations (Berkovits et al., 2000), we proceeded with our calculations as usual.

For reformulation (**RQ2a**), no significant development could be observed over time [$F_{(1)} = 0.202$, $p = 0.659$], as was the case for changes in articulation (**RQ2c**) [$F_{(1)} = 0.018$, $p = 0.896$] and the absence of repair attempts (**RQ2d**) [$F_{(1)} = 0.160$, $p = 0.694$]. Regarding time's effect on the frequency of repetition as a repair strategy (**RQ4b**), the ANOVA became significant with $F_{(1)} = 4.947$ and $p = 0.047$. For descriptive values, see **Table 5**. Age did not play a significant role [$F_{(1)} = 0.294$, $p = 0.595$]. It needs to be noted that for two research questions, Levene's Test for Equality of Variances revealed a significance, namely for the first measuring point regarding repetition (**RQ2b**) ($p = 0.001$) and the second measuring point regarding changes in articulation (**RQ2c**) ($p = 0.025$), which means that equality of variances is only partially given here. Therefore, **RQ2a**, **RQ2c**, and **RQ2d** could not be answered, whereas the calculations could give us some insight regarding **RQ2b**, namely that the amount of time the user has interacted with the VA has a negative impact on the number of repetitions they use during repairs.

5. DISCUSSION

The study presented in this paper aimed at investigating the situations in which communication breakdowns occur in interactions with a voice assistant in the family context. Hereby we address research gaps concerning (1) the behavior of both child and adult users in situations with communication breakdowns, (2) users' perception of such situations and their repair strategies, and (3) the consequences for user satisfaction with the assistant. In the previous section the results of our investigation were presented. Now we will reflect on these findings with regard to their implications for the state of research on the topic and their meaning for the design of future voice assistants or speech-based agents, more generally.

The examples presented in section 4.1 illustrate the user and VA behavior in situations when communication breakdowns occur. Instances of different repair strategies and system responses could be seen there, including some that seem rather baffling and can be explained only by looking at actual speech recognition results in the conversational data. The users of VAs, however, do not have access to this information, unless they are actively monitoring the history of their interactions online at that moment. The feedback that the assistant uses to signal a miscommunication is usually fourfold, as presented in **Table 3**. Generally, this feedback is considered unhelpful, as most of it puts the burden of finding a solution on the user. However, unexpected system behavior in cases of misunderstanding might still provide the users with clues to finding an adequate repair strategy, as suggested by Cho and Rader (2020). One such case could be seen in the first example in section 4.1.

In the conversational data we occasionally saw instances of family members helping each other, both within and across the two age groups (adults and children). Two examples of such situations were also presented in section 4.1. Unfortunately, as we did not have access to interactions between family members outside of Alexa log files, we could not identify whether more instances of such scaffolding took place during our study, and we were unable to gather sufficient data on this repair strategy for statistical analysis. In the second and the third examples, some of the children's requests lacked crucial contextual information and were therefore not successful. Further, children probably have difficulty recognizing the missing information due to an inaccurate mental model of the capabilities of the VA. In the second example, the adult was able to help the child by providing the missing contextual information to the assistant. In the third example, the children were unable to succeed with the request on their own as the system gave them no direct clues on how to fix the breakdown and their own knowledge about the VA may not have been sufficient to find the cause of the problem and an adequate repair strategy.

The statistical results give a mixed picture with regards to our research questions and hypotheses. Contrary to our expectations, **H1a** and **H1b** had to be rejected as we could find no significant effect of user's affinity for technology, their attribution of blame for communication breakdowns, and emotions they experience in miscommunication situations on either user's abandonment of failed requests or their success in accomplishing communicative repairs. This result suggests that other factors may contribute to the successful resolution of communication breakdowns, for example, the users' level of detail of knowledge about the VA or their speech quality. As mentioned in the literature overview, Siegert (2021) was able to find that accidental activation of Alexa was connected to the intonation variety of the speaker, so perhaps similar features could also be a factor here. Still, user characteristics such as affinity for technology may have an effect on the amount of effort the users are willing to invest into resolution of communication breakdowns or the variety of repair strategies they employ. These questions could be investigated in further research.

With regards to **H2** we could confirm that the number of abandoned failed requests was a significant negative predictor for the satisfaction of the user with the voice assistant (**H2a**) which connects to related findings presented in section 2.1. Errors in speech and intent recognition were found to have a negative impact on user satisfaction and they also can prevent communication breakdowns from being resolved, which might cause the user to abandon their request. Concerning **H2b**, however, no effect of successfully repaired requests on user satisfaction was identified. This might be connected to findings of Kiseleva et al. (2016) that show that correlation between task completion and satisfaction is task-dependent, so no general significant impact could be found across all tasks. The positive impact of task completion might also have been removed by the negative effect of effort to be spent when repairing the communication breakdown.

We were able to gain insights regarding the research question **RQ1a** where three predictors became significant for the choice of reformulation-based strategies: the mean of perceived reasons for errors as a negative predictor (the reasons for errors as can be seen in their description in section 3.3.4 all refer to user's linguistic and communicative shortcomings) and the self-perceived frequency of repetitions and abortions of requests as positive predictors. This suggests that the users tend to vary the semantic, syntactical and lexical contents of their utterance when they do not find reasons for errors in their own communicative behavior, reverting to a general strategy of varying their utterances to overcome internal errors or shortcomings of the VA. The positive predictors here are especially interesting and may suggest that the classification of repair strategies used in this study is not experienced as such by the users, e.g., they might see minor lexical or syntactical adjustments as repetitions of the old request and major semantic adjustments as presenting a new request after the abortion of the old one.

Additionally, we found a significant effect of the time spent interacting with the voice assistant on the amount of repetitions used to repair communication breakdowns (**RQ2b**). In the second half of our experiment, the mean amount of repetitions used by both adults and children per 100 requests decreased significantly. It would be interesting to further investigate whether repetitions were the least successful repair strategy and were therefore used less over time. Unexpectedly, we could not find a significant impact of age and time on other choices of behavior. Perhaps our study was still too short to see any real changes in children's interaction style, as Garg and Sengupta (2020) suggest it might take 2–3 months for them to become more independent in voice assistant use.

Overall, we can summarize our findings and identify avenues for further research as follows:

- As we could not support **H1**, we believe that other factors might contribute to the failure or success of repair attempts, for example, user's speech style. Instead, the internal characteristics such as affinity for technology could have an effect on the amount of effort the users are prepared to spend on resolution of communication breakdowns or the variety of repair strategies they employ.
- We could confirm **H2a** and found a significant negative effect of the number of abandoned requests on user satisfaction.
- We could not confirm **H2b** and believe that further investigation of communication breakdown situations distinguished by task and effort spent on repairs might be necessary to understand the impact of successfully repaired dialogues on user satisfaction. This kind of research may give insight into the interaction quality necessary in various use contexts and identify critical types of tasks where miscommunication is associated with the highest cost for the user. Being able to mitigate these costs can help reduce the risk of abandonment of the device in the long run due to disappointment of the user in conversational capabilities of the system.
- We found that users are more inclined to use reformulation, i.e., variation of semantic, syntactical or lexical content of their requests when they do not perceive the emergence of miscommunication as their fault. We could not find a significant effect of internal reasons for the choice of other strategies, so we suggest that situational clues such as the immediate response of the VA are more important for the choice of repair strategy (**RQ1**) which could be investigated based on our conversational data set in the future.
- Our results suggest that the classification of repair strategies that was created for this paper based on other approaches in the field may be differently perceived by users, as we found the self-perceived frequency of repetitions and abortions of requests to be positive predictors for the use of reformulation-based strategies (**RQ1**). This raises a plethora of questions concerning studies of communication breakdowns in interactions with VA in general, which should be addressed in the future. On what basis should the classification of repair strategies be constructed? Should there be an empirical evaluation of strategy categories? How much does self-perception of own repair behavior matter and what factors is it influenced by? How does it correspond with mental models about the functionality of the VA?
- In our data, use of repetition as repair strategy by both children and adults significantly decreases over time. No changes were found for other strategies (**RQ2**). To understand this finding, further analysis is needed. We hypothesize that repetitions might not lead to successful resolutions of breakdowns and, therefore, are used less with time. Our data set can be used to investigate the relationship between the used repair strategy and the following Alexa response.
- No significant impact of age on the choice of repair strategy was found, as well as no interaction effect between age and time (**RQ2**). We suggest that a study over the course of several months is needed to investigate changes in children's interaction style and the success of their communication with the device over time.

Finally, some limitations of our study need to be pointed out. Our sample of nine families is quite small and despite good balance with regards to the representation of gender and education amongst the demographic in Germany, it may not be representative in terms of voice-assistant-related behavior.

Additionally, we have only used the data sets of families that interacted with Amazon Echo Dot due to the availability of audio log files. While the results of Berdasco et al. (2019) show Alexa and Google Assistant both as significantly better than Cortana and Siri with regards to the quality and correctness of their responses, but not significantly different among each other, inclusion of various voice assistants into the study might have provided a fuller picture and potentially more robust results. Further, the conversational data might have missed some of the dynamics within a family concerning the use of the voice assistant, such as discourse scaffolding between family members. We were able to detect only those instances when the other family member actively addressed the voice assistant to help in miscommunication repair, such as the examples shown in section 4.1.

As far as consequences for the development of voice assistants are concerned, we support the view previously advocated by other researchers that VAs need to give better feedback to users in cases of miscommunication. Based on our results, we suggest that situational factors may play a major role in the choice of repair strategy as opposed to internal characteristics of the user. Finding appropriate repair strategies might reduce the amount of requests that users abandon and thus increase the level of satisfaction with the device. Hereby, the assistant could also gain awareness of the cost of repair a particular user associates with a particular type of task. The VA also could use information about critical and non-critical tasks in terms of repair cost to be more efficient. In the long run, all of these adjustments could help alleviate the problem of device abandonment or reduced usage shown in Luger and Sellen (2016) and Cho et al. (2019).

The clues that the assistant gives in cases of miscommunication, however, should be adapted to specific user groups, accounting for their mental models with regards to VA functionality, e.g., children. Here, specific discourse scaffolding strategies could be employed by the system to address unique challenges presented by children's speech, for example, building up on the research of Xu and Warschauer (2019). Children already are active users of speech-based technology (Yarosh et al., 2018), especially as it can alleviate the limitations presented by the lack of reading and writing skills in younger children, and the design of voice assistants or other speech-based agents has to address their specific needs. In general, however, one has to consider whether this sort of proactive or cooperative system behavior is possible under the paradigm of one-shot request-response interactions that can be seen in commercially available voice assistants today (Porcheron et al., 2018). We thus conjecture that future speech-based agents will require additional capabilities that would allow understanding the current interaction context and the mental states and knowledge level of the user, through some sort of joint co-construction and mentalizing occurring incrementally over the course of the interaction (Kopp and Krämer, 2021).

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Duisburg-Essen, Germany. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

LM responsible for the writing and editing of the most chapters, the general structure of this work and its presentation. LM and JS were additionally responsible for the formulation of the hypotheses and research questions and the direction of the data analysis process. CS was responsible for writing sections 3.1, 3.3, and 4.2. LMB was responsible for writing section 3.2. NK and SK were principal investigators of the project and are responsible for the overall scientific and organizational supervision as well as editing and revision of this article. All authors participated in either the design and execution of the study or in the data analysis process or both.

FUNDING

This research was part of the project IMPACT (*The implications of conversing with intelligent machines in everyday life for people's beliefs about algorithms, their communication behavior and their relationship-building*) and was funded by the Volkswagen Foundation.

ACKNOWLEDGMENTS

The authors would like to thank Benedict Schneider, a former student researcher at the Social Cognitive Systems Group who greatly contributed to the data analysis process necessary for this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2022.791704/full#supplementary-material

Data Sheet S1 | Questionnaire for the first home visit.

Data Sheet S2 | Questionnaire for the second and third home visit.

REFERENCES

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., and Hiniker, A. (2019). "Communication breakdowns between families and Alexa," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow). doi: 10.1145/3290605.3300473

Berdasco, A., Lopez, G., Diaz, I., Quesada, L., and Guerrero, L. A. (2019). "User experience comparison of intelligent personal assistants:

Alexa, Google Assistant, Siri and Cortana," in *Proceedings of 13th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2019)* (Toledo). doi: 10.3390/proceedings2019031051

Berkovits, I., Hancock, G. R., and Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: relative robustness to sphericity and normality violations. *Educ. Psychol. Measure*. 60, 877–892. doi: 10.1177/00131640021970961

Cheng, Y., Yen, K., Chen, Y., Chen, S., and Hiniker, A. (2018). "Why doesn't it work?: Voice-driven interfaces and young children's communication repair strategies," in *Proceedings of the IDC 2018: The 17th International Conference on Interaction Design and Children* (Trondheim). doi: 10.1145/3202185.3202749

Cho, J., and Rader, E. (2020). "The role of conversational grounding in supporting symbiosis between people and digital assistants," in *Proceedings of the ACM on Human-Computer Interaction*. 4, 1–28. doi: 10.1145/3392838

Cho, M., Lee, S.-S., and Lee, K.-P. (2019). "Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten," in *Proceedings of the 2019 ACM Designing Interactive Systems Conference (DIS 2019)* (San Diego, CA). doi: 10.1145/3322276.3322332

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., et al. (2019). "What makes a good conversation? Challenges in designing truly conversational agents," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow). doi: 10.1145/3290605.3300705

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Druga, S., Breazeal, C., Williams, R., and Resnick, M. (2017). ""Hey Google, is it OK if I eat you?": initial explorations in child-agent interaction," in *Proceedings of the IDC 2017: The 16th International Conference on Interaction Design and Children* (Stanford, CA). doi: 10.1145/3078072.3084330

Dubagunta, S. P., Kabil, S. H., and Magimai-Doss, M. (2019). "Improving children recognition through feature learning from raw speech signal," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton), 5736–5740. doi: 10.1109/ICASSP.2019.8682826

Gale, R., Chen, L., Dolata, J., van Santen, J., and Asgari, M. (2019). "Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques," in *Proceedings of Interspeech 2019* (Graz). doi: 10.21437/Interspeech.2019-3161

Gallagher, T. M. (1977). Revision behaviors in the speech of normal children developing language. *J. Speech Hear. Res*. 20, 303–318. doi: 10.1044/jshr.2002.303

Garg, R., and Sengupta, S. (2020). "He is just like me": a study of the long-term use of smart speakers by parents and children. *Proc. ACM Interact. Mobile Wearable Ubiquit. Technol*. 4, 3381002. doi: 10.1145/3381002

Hellwig, B., and Sloetjes, H. (2021). *ELAN - Linguistic Annotator*. Available online at: https://www.mpi.nl/corpus/html/elan/ (accessed September 28, 2021).

Izard, C. E., Dougherty, F. E., Bloxom, B. M., and Kotsch, N. E. (1974). *The Differential Emotions Scale: a Method of Measuring the Subjective Experience of Discrete Emotions*. Nashville, TN: Vanderbilt University Press.

Jiang, J., Awadallah, A. H., Jones, R., Ozertem, U., Zitouni, I., Kulkarni, R. G., et al. (2015). "Automatic online evaluation of intelligent assistants," in *Proceedings of the 24th International Conference on World Wide Web (WWW '15)* (Florence). doi: 10.1145/2736277.2741669

Jiang, J., Jeng, W., and He, D. (2013). "How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)* (Dublin). doi: 10.1145/2484028.2484092

Karrer, K., Glaser, C., Clemens, C., and Bruder, C. (2009). Technikaffinität erfassen - Der Fragebogen TA-EG. *Der Mensch im Mittelpunkt technischer Systeme* 8, 196–201.

Kats, R. (2018). *Are Kids and Teens Using Smart Speakers? The Smart Speaker Series Infographic*. Available online at: https://www.emarketer.com/content/the-smart-speaker-series-kids-teens-infographic (accessed December 18, 2021).

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., et al. (2017). "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)* (Vienna), 82–90. doi: 10.1145/2909824.3020229

Kiseleva, J., Williams, K., Jiang, J., Awadallah, A. H., Crook, A. C., Zitouni, I., et al. (2016). "Understanding user satisfaction with intelligent assistants," in *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)* (Chapel Hill, NC). doi: 10.1145/2854946.2854961

Kleinberg, S. (2018). *Why Parents Are Voice-Assistance Power Users*. Available online at: https://www.thinkwithgoogle.com/marketing-strategies/app-and-mobile/voice-assistance-parent-users/ (accessed December 18, 2021).

Kopp, S., and Krämer, N. (2021). Revisiting human-agent communication: the importance of joint co-construction and understanding mental states. *Front. Psychol*. 12, 580955. doi: 10.3389/fpsyg.2021.580955

Lovato, S., and Piper, A. M. (2015). ""Siri, is this you?": understanding young children's interactions with voice input systems," in *Proceedings of the IDC 2015: The 14th International Conference on Interaction Design and Children* (Medford, MA). doi: 10.1145/2771839.2771910

Lovato, S., Piper, A. M., and Wartella, E. A. (2019). ""Hey Google, do unicorns exist?": conversational agents as a path to answers to children's questions," in *Proceedings of the IDC 2019: The 18th International Conference on Interaction Design and Children* (Boise, ID). doi: 10.1145/3311927.3323150

Luger, E., and Sellen, A. (2016). ""Like having a really bad PA": the gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, CA), 5286–5297. doi: 10.1145/2858036.2858288

Motta, I., and Quaresma, M. (2022). "Users' error recovery strategies in the interaction with voice assistants (VAs)," in *Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021)*, eds N. L. Black, W. P. Neumann, and I. Noy (Springer International Publishing), 658–666. doi: 10.1007/978-3-030-74614-8_82

Myers, C., Furqan, A., Nebolsky, J., Caro, K., and Zhu, J. (2018). "Patterns for how users overcome obstacles in voice user interfaces," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC). doi: 10.1145/3173574.3173580

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). "Voice interfaces in everyday life," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC). doi: 10.1145/3173574.3174214

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. (2017). ""Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, CO). doi: 10.1145/3027063.3053246

Shivakumar, P. G., and Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: evaluation, analysis and recommendations. *Comput. Speech Lang*. 63, 101077. doi: 10.1016/j.csl.2020.101077

Siegert, I. (2021). "Effects of prosodic variations on accidental triggers of a commercial voice assistant," in *Proceedings of INTERSPEECH 2021* (Brno). doi: 10.21437/Interspeech.2021-1354

Siegert, I., and Krüger, J. (2018). How do we speak with ALEXA - Subjective and objective assessments of changes in speaking style between HC and HH conversations. *Kognitive Systeme* 1. doi: 10.17185/duepublico/48596

Stent, A. J., Huffman, M. K., and Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: local and global hyperarticulation. *Speech Commun*. 50, 163–178. doi: 10.1016/j.specom.2007.07.005

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J. Pers. Assess*. 80, 99–103. doi: 10.1207/S15327752JPA8001_18

Szczuka, J., Artelt, A., Geminn, C., Hammer, B., Kopp, S., Krämer, N., et al. (2021). *Konnen Kinder aufgeklärte Nutzer*innen von Sprachassistenten sein*?: *Rechtliche, psychologische, ethische und informatische Perspektiven*. University of Duisburg-Essen, University Library, Essen.

Tenzer, F. (2021). *Prognose zum Absatz von intelligenten Lautsprechern weltweit bis 2025*. Available online at: https://de.statista.com/statistik/daten/studie/1079997/umfrage/prognose-zum-absatz-von-intelligenten-lautsprechern-weltweit/ (accessed September 28, 2021).

Viera, A. J., and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Med*. 37, 360–363.

World Health Organisation (2021). *World Health Organisation - Countries*. Available online at:https://www.who.int/countries (accessed September 29, 2021).

Wu, F., Garcia-Perera, L. P., Povey, D., and Khudanpur, S. (2019). "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proceedings of Interspeech 2019* (Graz). doi: 10.21437/Interspeech.2019-2980

Xu, Y., and Warschauer, M. (2019). "Young children's reading and learning with conversational agents," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow). doi: 10.1145/3290607.3299035

Yarosh, S., Thompson, S., Watson, K., Chase, A., Senthilkumar, A., Yuan, Y., et al. (2018). "Children asking questions: speech interface reformulations and personification preferences," in *Proceedings of the IDC 2018: The 17th International Conference on Interaction Design and Children* (Trondheim). doi: 10.1145/3202185.3202207

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

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 29 April 2022
doi: 10.3389/fpsyg.2022.856283

Check for updates

A Mind in Intelligent Personal Assistants: An Empirical Study of Mind-Based Anthropomorphism, Fulfilled Motivations, and Exploratory Usage of Intelligent Personal Assistants

Cuicui Cao[1], *Yingying Hu*[2] *and Haoxuan Xu*[3]*

[1]School of Management, Huazhong University of Science and Technology, Wuhan, China, [2]School of Information Management and Statistics, Hubei University of Economics, Wuhan, China, [3]School of Business Administration, Zhongnan University of Economics and Law, Wuhan, China

OPEN ACCESS

Edited by:
Benjamin Weiss,
Technical University of Berlin,
Germany

Reviewed by:
Adam Galpin,
University of Salford,
United Kingdom
Guy Laban,
University of Glasgow,
United Kingdom

***Correspondence:**
Haoxuan Xu
haoxuanxu@zuel.edu.cn

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

Received: 17 January 2022
Accepted: 04 April 2022
Published: 29 April 2022

Citation:
Cao C, Hu Y and Xu H (2022) A Mind in Intelligent Personal Assistants: An Empirical Study of Mind-Based Anthropomorphism, Fulfilled Motivations, and Exploratory Usage of Intelligent Personal Assistants. Front. Psychol. 13:856283. doi: 10.3389/fpsyg.2022.856283

Intelligent personal assistants (IPAs) own anthropomorphic features which enable users' perception of anthropomorphism. Adopting the perspective of mind-based anthropomorphism, the purpose of this paper is to investigate how mind-based anthropomorphism influences users' exploratory usage of IPAs. Based on the notion that anthropomorphism can satisfy people's sociality and effectance motivation, we hypothesize that mind-based anthropomorphism can enhance people's social connection with IPAs and IPA self-efficacy, which can in turn influence their exploratory usage of IPAs. Questionnaires were developed and distributed to users who had experience in smart speaker-based IPAs on Wenjuanxing and 551 valid questionnaires were collected to test the research model. The results revealed that cognitive and affective anthropomorphism exerted common and differential impacts on IPA self-efficacy and social connection. Cognitive anthropomorphism versus affective anthropomorphism had stronger influences on IPA self-efficacy, while affective anthropomorphism had stronger impacts on social connection. Both IPA self-efficacy and social connection enhanced users' intentions to explore IPAs. This study enriches previous studies on IPA adoption or post-adoption by investigating exploratory usage which captures how users are deeply engaged with IPAs.

Keywords: IPAs, anthropomorphism, IPA self-efficacy, social connection, intention to explore IPAs

INTRODUCTION

Intelligent personal assistants (IPAs) have emerged as one of the fastest-growing artificial intelligence (AI) applications in recent years, and many giant technology companies have developed their IPAs, such as Siri by Apple, Alexa by Amazon, and TmallGenie by Alibaba in China. The global IPA market size is expected to reach USD 45.1 billion by 2027, expanding at a CAGR of 34.0% (Businessware, 2020). IPA is defined as "*a software agent that acts* *intelligently and uses natural language to provide professional/administrative, technical, and social assistance to human users by automating and easing many day-to-day activities*" (Han and Yang, 2018; Moussawi et al., 2020). As disembodied agents or virtual agents embedded in various devices, IPAs can help manage home automation, complete daily tasks, and communicate with users in a way that resembles interpersonal communication (Han and Yang, 2018). They are permeating our daily lives and offer significant potential capabilities. Take Amazon's Alexa, for example, it owns 80,000 skills and the number of skills is increasing each day (Amazon, 2021). In this case, users' exploratory usage of IPAs (i.e., exploring various IPA skills to provide hedonic and utilitarian value) becomes especially important. For IPA users, they can fully utilize the powerful capabilities provided through IPA skill exploration. If they do not engage in exploration and only use several limited functions, they will soon find IPAs useless and abandon them. For IPA providers such as Amazon, users' exploratory behaviors can help them save cost, acquire more value from users (Pan et al., 2017), and maintain long-term relationships with users (Pan et al., 2017). Thus, it is imperative to find out what factors contribute to users' exploratory usage of IPAs.

Previous studies on adoption or post-adoption of IPAs mostly focus on adoption/acceptance (Park et al., 2018; Yang and Lee, 2018; Mclean and Osei-frimpong, 2019; Moussawi and Benbunan-fich, 2020; Moussawi et al., 2020; Mishra et al., 2021; Vimalkumar et al., 2021) or continuous usage of IPAs (Han and Yang, 2018; Moussawi and Koufaris, 2019; Ki et al., 2020; Hu et al., 2021; Sun et al., 2021), which provide limited knowledge on how users deeply interact and engage with IPAs (Delgosha and Hajiheydari, 2021). Exploratory usage of IPAs belongs to such type of deeply-involved post-adoption behaviors and refers to using newly added skills of IPAs or using some other skills beyond their routine usage (Nambisan et al., 1999; Ahuja and Thatcher, 2005). Such research is important and imperative since there are several examples of IPAs which fail to connect with users in a deeply social manner despite early success to create enthusiasm in user acceptance (Cho et al., 2019).

In exploring factors influencing adoption or post-adoption of IPAs, prior studies emphasize the role of anthropomorphism (i.e., the attribution of human characteristics to nonhuman entities) since IPAs' anthropomorphic features (e.g., voice and humor) generate users' perceptions of anthropomorphism (i.e., humanlike perception), which may affect users' subsequent behavior toward IPAs (Moussawi and Benbunan-fich, 2020; Moussawi et al., 2020; Hu et al., 2021; Mishra et al., 2021). However, the findings regarding the direct effects of perceived anthropomorphism on users' adoption behavior are mixed (Moussawi et al., 2020; Hu et al., 2021). For instance, Hu et al. (2021) found the direct influence of perceived anthropomorphism on the continuous usage of IPAs. Meanwhile, Moussawi et al. (2020) did not discover the direct effect of perceived anthropomorphism on users' adoption intention. Thus, it is necessary to study mediators to avoid over- or under-estimate the role of anthropomorphism.

Driven by practical problems and theoretical gaps, this study aims to investigate the following questions: *How does users' anthropomorphism of IPAs influence their exploratory usage?* To achieve this goal, we adopt a perspective of mind-based anthropomorphism, which refers to "the attribution of unobservable and uniquely human mental capacities to nonhuman entities" (Castelo et al., 2019), to understand how users anthropomorphize IPAs. We propose cognitive and affective anthropomorphism as the two dimensions of mind-based anthropomorphism, operationalized by the humanlike cognitive abilities (i.e., autonomy and interactivity) and affective abilities (i.e., sociability) of IPAs, respectively. Drawing on fulfilled motivations of anthropomorphism, we contend that mind-based anthropomorphism of IPAs will satisfy users' effectance and sociality motivations, which are manifested as IPA self-efficacy and social connection respectively, and users' intentions to explore IPAs will be further enhanced by these fulfilled motivations. More importantly, cognitive and affective anthropomorphism will exert differential impacts on IPA self-efficacy and social connection.

LITERATURE REVIEW AND THEORETICAL FOUNDATIONS

Prior Adoption-Related Studies in IPAs Context

We mainly classify IPA adoption-related studies into two streams. The first stream focuses on the adoption of IPAs, which focuses on whether users adopt IPAs or not (Park et al., 2018; Yang and Lee, 2018; Mclean and Osei-frimpong, 2019; Moussawi and Benbunan-fich, 2020; Moussawi et al., 2020; Mishra et al., 2021; Vimalkumar et al., 2021). On the one hand, these studies apply traditional adoption perspectives like the Unified Theory of Acceptance and Use of Technology model, Uses and Gratification theory, and utilitarian and hedonic value (Yang and Lee, 2018; Mclean and Osei-frimpong, 2019; Vimalkumar et al., 2021). On the other hand, they investigate unique characteristics of IPAs such as anthropomorphism, intelligence, and privacy concern (Mclean and Osei-frimpong, 2019; Moussawi and Benbunan-fich, 2020; Moussawi et al., 2020; Vimalkumar et al., 2021). These studies are limited in offering an understanding of how users engage with IPAs after adoption.

The second stream focuses on the post-adoption of IPAs. These studies mostly focus on continuous usage of IPAs, which focuses on whether users continue to use IPAs after initial adoption (Han and Yang, 2018; Moussawi and Koufaris, 2019; Ki et al., 2020; Hu et al., 2021; Sun et al., 2021). Some of the factors investigated are similar to that of adoption such as anthropomorphism and intelligence (Moussawi and Koufaris, 2019). Further, they go beyond to investigate some other factors related to the para-social relationship (Han and Yang, 2018). In addition, they investigate how service failure influences users' continuance usage of IPAs (Sun et al., 2021). Besides general continuous usage, other researchers focus on a specific application setting of IPAs, such as voice shopping or playful requests

(Maarek, 2018, 2019, Shani et al., 2021). Though insightful, these studies offer limited value in understanding how users deeply engage with IPAs.

Taken together, users' exploratory usage of IPAs in current research remains less explored, which is distinct from the adoption or continuous usage. Considering the important role of anthropomorphism in the context of IPAs, we aim to examine how the anthropomorphism of IPAs is helpful for users' exploratory usage. The conceptualization of anthropomorphism and related research will be discussed in the next section.

Anthropomorphism of IPAs: From the Humanlike Mind Perspective

In the context of IPAs, prior studies have investigated anthropomorphic characteristics (e.g., voice, humor) and perceived anthropomorphism (i.e., humanlike perception). In the current study, we focus on users' humanlike perception of IPAs, namely, perceived anthropomorphism. Perceived anthropomorphism of IPAs has been investigated in prior studies from the perspective of humanlike mind perception (Hu et al., 2021; Li and Sung, 2021), or attributes that are either uniquely or typically human (Moussawi and Benbunan-fich, 2020; Moussawi et al., 2020). In the current study, we adopt the perspective of mind-based anthropomorphism which refers to attributing humanlike mental capacities to nonhuman entities (Waytz et al., 2010a).

According to Castelo et al. (2019), mind-based anthro pomorphism can be classified into cognitive anthropomorphism and affective anthropomorphism. Cognitive anthropomorphism refers to the attribution of humanlike cognitive capacities to nonhuman entities, such as self-control, plan, and cognitive sophistication. It concerns more about the agents' ability to "do" and how they deal with the tasks. Affective anthropomorphism is defined as attributing mental capacities to feel and express emotions, such as emotional responsivity, hunger, and fear, to nonhuman agents. It concerns more about the agents' ability to "feel" and how they deal with others. Different kinds of AI applications own different humanlike cognitive and affective abilities (Castelo, 2019).

In the context of IPAs, prior studies also distinguish the two dimensions and they found the differential effects of these two dimensions (Hu et al., 2021). At the same time, other researchers investigated the humanlike abilities of IPAs (Cao et al., 2019; Wagner et al., 2019; Wagner and Schramm-klein, 2019). Contextualized in our study, cognitive anthropomorphism is manifested as interactivity and autonomy of IPAs. Interactivity refers to the ability to communicate with users in a consecutive way (Sundar et al., 2016; Wagner and Schramm-klein, 2019). Autonomy refers to the capacity to help people autonomously perform tasks, such as controlling smart home devices and setting alarms (Rijsdijk et al., 2007; Wagner and Schramm-klein, 2019). Affective anthropomorphism is mainly manifested as sociability, which refers to the capability of IPAs to carry out sociable behavior (Heerink et al., 2010; Cao et al., 2019).

As for the direct effect of perceived anthropomorphism on adoption or post-adoption behaviors in the context of IPAs, prior studies present some mixed findings. For instance, Li and Sung (2021) posited that the relationship between people's acceptance of AI assistants and perceived anthropomorphism was mediated by psychological distance. However, Hu et al. (2021) found the direct effect of humanlike perception of IPAs on continuous usage of IPAs. Thus, we posit that it is necessary to study the mediators between perceived anthropomorphism and exploratory usage of IPAs.

Taken together, prior studies on IPA anthropomorphism indicate that mind-based anthropomorphism has two dimensions (i.e., cognitive vs. affective) and they may have differential impacts. However, they present some mixed findings between the direct effect of perceived anthropomorphism on adoption or post-adoption behaviors. Thus, we intend to study the mediators from the perspective of fulfilled motivations of anthropomorphism, which will be discussed in the next section.

Fulfilled Motivations of IPA Anthropomorphism

Based on previous research, effectance motivation and sociality motivation are the two motivational factors of users' anthropomorphism (Epley et al., 2007). That is to say, anthropomorphism is a way to satisfy users' effectance motivation and sociality motivation.

Effectance motivation involves humans' motivation to interact with the outside world effectively (White, 1959). As vulnerable creatures, humans have the desire to reduce the uncertainty of the environment and try to understand and predict the agents that inhabit this environment. Anthropomorphism provides such an efficient way to better understand and predict a context by increasing its controllability and predictability and satisfies human's desire to master the environment (Epley et al., 2007; Waytz et al., 2010c). Anthropomorphizing nonhuman entities enhances people's ability to explain the nonhuman entities' actions and accordingly improves users' efficacy in interacting with them. For instance, yelling at a malfunctioned computer may help people ease their burden (Luczak et al., 2003). Similarly, anthropomorphism of IPAs can also satisfy users' effectance motivation (Cao et al., 2019; Chen and Park, 2021; Li and Sung, 2021), which is manifested as IPA self-efficacy in the current study. IPA Self-efficacy refers to users' evaluation of their competence to use IPAs (Compeau and Higgins, 1995) and has been validated as a strong predictor of usage behaviors, especially those that extend beyond the defined usage (Wang et al., 2008; Peng et al., 2018; Tams et al., 2018).

Sociality motivation refers to humans' innate need and desire to build social connections with the outside world (Baumeister and Leary, 1995). Driven by this motivation, people are more likely to actively search for sources of social connections in their environment, and more sensitive to notice and perceive human characteristics of nonhuman agents (Epley et al., 2007). This desire can be satisfied by anthropomorphizing nonhuman entities, such as technological devices and pets when people cannot establish social connections with other people. For example, lonely people anthropomorphize their pets to obtain the social connection they need (Epley et al., 2007). Similarly, anthropomorphism of IPAs can satisfy users' sociality motivation as well (Cao et al., 2019; Chen and Park, 2021; Li and Sung, 2021; Noor et al., 2021), which is manifested by social connection with IPAs in the current study. Social connection refers to users' feeling of closeness with the IPAs (Lee et al., 2001) and has been validated by previous studies to strengthen usage behaviors (Tseng et al., 2018).

When individuals are driven by different motivations (effectance vs. sociality), they tend to prioritize different abilities or attributes of the targets. For example, consumers with a sociality motivation attribute more affective abilities to brands. However, consumers with an effectance motivation attribute more cognitive abilities to brands (Changizi and Hall, 2001; Balcetis and Dunning, 2006; Chen et al., 2013). Based on these arguments, we propose that although two types of mind-based anthropomorphism can influence IPA self-efficacy (i.e., effectance motivation) and social connection (i.e., sociality motivation), their influences may be different. In other words, the two types of mind-based anthropomorphism play different roles in satisfying these two motivations.

RESEARCH MODEL AND HYPOTHESES

We present our research model in **Figure 1**. Based on fulfilled motivations of anthropomorphism, we hypothesize that both cognitive and affective anthropomorphism of IPAs can satisfy users' effectance and sociality motivation, which are represented by IPA self-efficacy and social connection. However, these two types of anthropomorphism also play different roles in satisfying these two motivations. To be specific, we hypothesize that cognitive anthropomorphism exerts stronger effects on IPA self-efficacy, while affective anthropomorphism has stronger effects on social connection. Finally, we hypothesize that IPA self-efficacy and social connection positively influence users' intentions to explore IPAs. We also include several control variables, such as gender (GEN), age, relation status (STS), and frequency of use (FRE) in the research model.

Anthropomorphism and IPA Self-Efficacy (Fulfilled Effectance Motivations)

IPA self-efficacy refers to users' evaluation of their competence to use IPAs (Compeau and Higgins, 1995). Humans have a fundamental need to effectively interact with the outside world and accordingly try to reduce the uncertainty and increase the controllability of the environment (White, 1959). Anthropomorphism is such an effective way to make contact with nonhuman agents (Epley et al., 2007). It is realized based on knowledge about the self or humans; thus, people have confidence in predicting future behaviors of the agents (Epley et al., 2007). Thus, anthropomorphism can enhance people's sense of efficacy toward nonhuman agents.

In the context of IPAs, we believe both cognitive and affective anthropomorphism of IPAs can enhance users' IPA self-efficacy. Firstly, the cognitive intelligence of IPAs is increasing because sophisticated algorithms are continuously updated to improve the IPAs' ability, such as predicting and satisfying users' needs and conducting continuous dialogue with users. Users can contact with IPAs more simply and comfortably with less effort in adapting to the IPAs. It is expected that users who attribute humanlike cognitive capacities to IPAs will feel like they are communicating with humans, and will have a stronger sense of self-efficacy. Secondly, the emotional intelligence of IPAs is also improved in several ways. For example, users' emotions can be identified through changes in tone or words, or some emotional phrases can be used in specific contexts. In addition, the virtual characters of IPAs are usually designed with a sense of humor by telling jokes or witticisms. These techniques can help avoid machines' coldness and also comfort the users

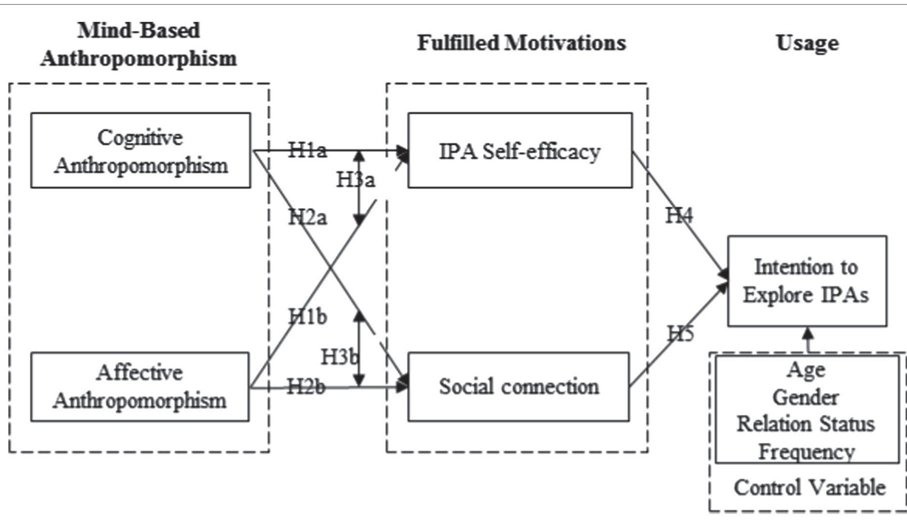


FIGURE 1 | Research model.

when there may be failures during interactions. Users who attribute emotional capacities to IPAs will feel more comfortable communicating with the IPAs and less uncertain about the IPAs because they understand and express emotions as humans do. It has also been validated by prior studies that IPA anthropomorphism decreases the sense of unfamiliarity and brings about a stronger sense of efficacy toward IPAs (Cao et al., 2019; Chen and Park, 2021; Li and Sung, 2021). Thus, we hypothesize:

> *H1a*: Cognitive anthropomorphism of IPAs positively affects users' IPA self-efficacy.
> *H1b*: Affective anthropomorphism of IPAs positively affects users' IPA self-efficacy.

Anthropomorphism and Social Connection (Fulfilled Sociality Motivation)

Social connection refers to users' feeling of closeness with the IPAs (Lee et al., 2001). Humans have a natural desire and tendency to be connected to other humans (Baumeister and Leary, 1995). They can also establish humanlike connections with nonhuman objects by anthropomorphism when the social connection to other people is absent (Epley et al., 2007). For example, lonely people who lack social connection find nonhuman agents, such as dogs and electronic gadgets, to be more humanlike because they can make it up by establishing connections with those nonhuman agents (Epley et al., 2007). Not only chronic loneliness but also social disconnection in some circumstances may activate anthropomorphism, which is more prevalent in daily life for most people (Epley et al., 2007). Thus, anthropomorphism is a way to satisfy people's sociality motivation. Previous research has found that interaction with anthropomorphic products can satisfy social needs and thus alleviate social exclusion effects (Mourey et al., 2017).

In the context of IPAs, such needs for social connection can be fulfilled by both cognitive and affective anthropomorphism of IPAs. Intelligent agents like IPAs become prevalent in our daily life, and they are more and more like friends that we can communicate with without worrying about awkwardness and disturbance. For example, many users anthropomorphize IPAs and create certain social connections with them (Cao et al., 2019; Chen and Park, 2021; Li and Sung, 2021; Noor et al., 2021). As we mentioned in the arguments for hypotheses 1a and 1b, as the cognitive and emotional intelligence of IPAs is gradually improved, users are more likely to attribute humanlike cognitive and emotional ability to IPAs and regard it as a source of social connection. Accordingly, we also expect that IPA anthropomorphism will increase a feeling of social connection. Thus, we hypothesize:

> *H2a*: Cognitive anthropomorphism of IPAs positively affects users' social connection.
> *H2b*: Affective anthropomorphism of IPAs positively affects users' social connection.

Comparative Effects of Cognitive vs. Affective Anthropomorphism on IPA Self-Efficacy and Social Connections

Although cognitive anthropomorphism and affective anthropomorphism can both influence IPA self-efficacy and social connection with IPAs, we believe the degree of their influence will be different. The underlying reason is that when people are influenced by different motivations, they tend to prioritize stimulus objects consciously or unconsciously, to better satisfy their motivation, which has also been validated by previous research (Changizi and Hall, 2001; Balcetis and Dunning, 2006; Chen et al., 2013). This can be explained by the three mechanisms for the motivation-based perception of humans, namely, selective sensitization, perceptual defense, and value resonance (Postman et al., 1948). Selective sensitization means that motivation as a sensitizer lowers the threshold for acceptable stimulus objects. Perceptual defense refers to the increase of the threshold for unsuitable stimulus objects. Value resonance keeps people responding to objects that are valuable to their motivations.

In the context of IPA anthropomorphism, users motivated by sociality are more sensitive to the affective capacities of IPAs which can better directly alleviate their loneliness (selective sensitization), are less sensitive to cognitive capacities, such as autonomy and interactivity, which are not so critical for sociality motivations (perceptual defense), and also attribute affective capacities to IPAs in congruence with their sociality motivations (value resonance). The same reasoning process applies to those users motivated by effectance. Simply put, people will attribute more cognitive capacities, such as the ability to autonomously complete tasks and communicate with users in a contingent way, since doing so helps alleviate their desire for control and predictability. Consequently, we argue that affective anthropomorphism can better satisfy users' sociality motivation compared with cognitive anthropomorphism, and cognitive anthropomorphism can better satisfy users' effectance motivation than affective anthropomorphism. Thus, we hypothesize:

> *H3a*: Cognitive anthropomorphism exerts stronger effects on IPA self-efficacy than affective anthropomorphism.
> *H3b*: Affective anthropomorphism exerts stronger effects on social connection than cognitive anthropomorphism.

Fulfilled Motivations and Intention to Explore IPAs

Different from other traditional IT applications, IPAs are more like ambiguous technology, which relies on users' self-driven exploratory form of learning rather than traditional instruction-based learning (Zhao et al., 2018). Thus, users' self-confidence in interacting with IPAs may help them overcome barriers to explore unfamiliar or hidden functions or use familiar functions innovatively, such as searching for information or trialing new oral commands. Previous research has revealed that individuals with a higher level of self-efficacy are inclined to be more devoted in pursuit of goals (Latham et al., 2000), more persistent in the face of difficulties (Schaefers et al., 1997), and more proactive in information seeking. Recent IS research has also validated the positive impacts of self-efficacy on exploratory, creative, or extended usage of ISs (Wang et al., 2008; Peng et al., 2018; Tams et al., 2018). Thus, we hypothesize:

> *H4*: IPA self-efficacy positively influences users' intentions to explore IPAs.

Though social connection in this study refers to the relationships and connections with IPAs, theories of interpersonal relationships may also be referential to understanding such relationships. For example, for those users who establish social connections with IPAs, IPAs function not simply as a useful instrument but as a friend or family member to them (Purington et al., 2017). One major characteristic or outcome of close interpersonal relationships is the commitment to the partner in the relationship, and such commitment will in turn, positively affect one's feelings and behaviors toward the partner (Rusbult, 1980). We believe when users establish a close connection with IPAs, they will also experience a certain degree of commitment to these intelligent agents and also tend to cherish the possessions which signify social relationships (Richins, 1994), such as making efforts to maintain the relationships with IPAs by exploring more functions of IPAs. Thus, we hypothesize:

> *H5*: Social connection positively influences users' intentions to explore IPAs.

METHODOLOGY

Participants

In the present study, we chose users of smart speakers as the research subjects. Smart speakers are one of the most popular IPAs in China. Despite a short history, smart speakers have permeated many people's daily lives. Examples of popular smart speakers in China include Xiaoaitongxue by Xiaomi, TmallGenie by Alibaba, and Duer by Baidu. Though other IPAs, such as Siri, might also be famous, we did not choose those kinds of IPAs because the user base of smart speakers is larger. We believed it was appropriate to choose the smart speaker as the research object.

An online questionnaire was distributed through a leading online survey distribution platform with 260 million registered users in China. Only users who had experience in smart speakers were invited, and each of them received a monetary award for each questionnaire. The survey began in July 2020 and lasted for 2 weeks. The responses were examined carefully, and invalid responses with missing answers, the same answers to all items, and a completion time of fewer than 6 min were removed. After deleting the invalid responses, 551 valid responses were left. The basic demographic information is listed in **Table 1**. Among all the respondents, 63.5% were male, and 36.5% were female, which was consistent with the overall composition of smart speaker users in China (Aurora Mobile, 2019). Most users were below 35 years old, which was reasonable since smart speakers were quite new in China, and young people tend to be more interested in new IT products.

Measurements

All the measurement items in the current study were adapted from the previous literature. They were measured by seven-point Likert scales. The measurement items for intention to explore IPAs were adapted from Nambisan et al. (1999). The items for assessing social connection with IPAs were adapted from Lee et al. (2001). The items for measuring IPA self-efficacy were adapted from Compeau and Higgins (1995). Affective anthropomorphism was measured by users' perceived sociability of IPAs, whose measurement items were adapted from Heerink et al. (2010). Cognitive anthropomorphism was measured by users' perceived autonomy and interactivity of IPAs, whose measurement items were adapted from Rijsdijk et al. (2007) and Bellur and Sundar (2017). The final questionnaire used in the survey is listed in **Table A1** in **Appendix A**.

Data Analysis Procedure

Following the two-step procedure proposed by Anderson and Gerbing (1988), we analyzed the research model with SPSS 22 and AMOS 24, and the data analysis part was composed of the following two parts: analysis of the measurement model and structural model. In the current study, we chose the covariance-based SEM method.

Measurement Model Testing

First, we conducted the confirmatory factor analysis for the measurement model with AMOS 24. All the fit indices (i.e.,

TABLE 1 | Sample profile ($N = 551$).

Variable	Option	N	Percentage (%)
Gender (GEN)	Male	350	63.5
	Female	201	36.5
Age	<=25	118	21.4
	26–30	189	34.3
	31–35	138	25.0
	36–40	67	12.2
	41–45	30	5.4
	>=46	9	1.6
Education (EDU)	High school or below	28	5.1
	Two-year college	68	12.3
	Four-year college	392	71.1
	Graduate school or above	63	11.4
Frequency of use (FRE)	At least once per day	190	34.5
	4–5 times per week	211	38.3
	2–3 times per week	123	22.3
	Less than once per week	27	4.9
Relation status (STS)	Single	118	21.4
	Just in love	88	16.0
	Married with no children	34	6.2
	Married with children	311	56.4
Years used (YU)	<=3 months	20	3.6
	3–6 months	117	21.2
	6 months–1 year	196	35.6
	1 year–1.5 years	145	26.3
	1.5 years above	73	13.2

CMIN/DF = 1.832, RMSEA = 0.039, NFI = 0.964, CFI = 0.974) met the criterion of each index (i.e., CMIN/DF < 3, RMSEA < 0.08, NFI > 0.9, CFI > 0.9), which indicated acceptable model fit (Bentler and Bonett, 1980).

Then, the construct reliability was evaluated. The construct reliability was all good (i.e., Cronbach's alpha > 0.7; composite reliability > 0.7; Nunnally, 1978) and the details can be found in **Table B1** in **Appendix**. Next, the construct validity was evaluated through AVEs and the comparison of the square root AVEs of each construct with other correlation coefficients. **Table B1** in **Appendix** shows that the AVEs were greater than 0.5 and thus the convergent validity was good (Fornell and Larcker, 1981). Besides, **Table 2** shows that the values on the diagonal (i.e., square root AVEs) were larger than other values on the corresponding rows and columns, indicating good discriminant validity (Fornell and Larcker, 1981).

Next, we assessed the construct validity and reliability of cognitive anthropomorphism according to Petter et al. (2007) since we chose the second-order formative model for cognitive anthropomorphism (Diamantopoulos, 2011; Jarvis et al., 2004). (The reasons can be found in **Appendix C**). First, each first-order construct had a significant path pointing to cognitive anthropomorphism, indicating satisfactory validity. Second, the variance inflation factor (VIF) values of the two first-order constructs were under the recommended value of 3.3, suggesting acceptable reliability (Diamantopoulos and Siguaw, 2006).

Finally, we adopted Harmon's single-factor analysis to examine the common method bias since the data were self-reported. The first factor explained 48.4% of the total variance, which was below the threshold of 50%; thus, no single factor existed, which explained most of the variance (Lindell and Whitney, 2001).

Structural Model Testing

We tested the structural model with the maximum likelihood technique in Amos 24. The model fit was acceptable since all the fit indices (i.e., CMIN/DF = 1.933, RMSEA = 0.041, NFI = 0.948, CFI = 0.974) met the criterion of recommended values (i.e., CMIN/DF < 3, RMSEA < 0.08, NFI > 0.9, CFI > 0.9; Bentler and Bonett, 1980).

The hypothesis testing results are summarized in **Figure 2**. The explained variance of each dependent construct was 52.0, 61.1, and 63.2% for IPA self-efficacy, social connection, and intention to explore IPAs, respectively. Regarding the impacts of cognitive and affective anthropomorphism on IPA self-efficacy and social connection, IPA self-efficacy was significantly affected by both cognitive anthropomorphism ($b = 0.600$, $p < 0.001$) and affective anthropomorphism ($b = 0.150$, $p < 0.05$), thus supporting H1a and H1b. Social connection was positively influenced by affective anthropomorphism ($b = 0.699$, $p < 0.001$) but not for cognitive anthropomorphism; thus, H2b was supported, and H2a was not. IPA self-efficacy also significantly influenced users' intentions to explore IPAs, and the standardized path coefficient was 0.545 ($p < 0.001$); thus, H4 was supported. Social connection had a significantly positive effect on users' intentions to explore IPAs with a standardized path coefficient of 0.364 ($p < 0.001$); thus, H5 was supported. As for the control variables, age negatively influenced users' intentions to explore IPAs, which was reasonable since young people are more tech-savvy and more likely to explore new functions of IPAs. Other control variables (gender, use frequency, and relation status) did not have significant impacts on users' intentions to explore IPAs.

The comparison hypothesis was tested with the pairwise parameter comparisons in AMOS. The results are summarized in **Table 3**. Cognitive anthropomorphism and affective anthropomorphism differed in their impacts on IPA self-efficacy and social connection. Cognitive anthropomorphism had a stronger effect on IPA self-efficacy than affective anthropomorphism, thus supporting H3a. Affective anthropomorphism had a stronger effect on social connection than cognitive anthropomorphism; thus, H3b was also supported.

DISCUSSIONS OF RESULTS

The current study aims to examine how mind-based anthropomorphism, in terms of cognitive anthropomorphism and affective anthropomorphism, influences people's exploratory usage of IPAs. The results depict a high degree of explanatory power for all dependent variables and reveal some significant and interesting findings as well.

Firstly, it is found that cognitive and affective anthropomorphism have common impacts on IPA self-efficacy and social connection. Though cognitive anthropomorphism exerts non-significant impacts on social connection, other supported hypotheses still reveal that users' anthropomorphism of IPAs has a positive influence on self-efficacy and social connection. It is consistent with previous findings that anthropomorphism of IPAs can decrease their sense of unfamiliarity and increase their people's social connection with IPAs (Cao et al., 2019; Chen and Park, 2021; Li and

TABLE 2 | Discriminant validity.

	VIF	SB	INT	AU	ISE	SC	IE
Sociability	2.42	0.79					
Interactivity	2.22	0.62	0.81				
Autonomy	1.97	0.57	0.62	0.78			
IPA self-efficacy	1.91	0.52	0.55	0.53	0.80		
Social connection	2.14	0.66	0.54	0.48	0.35	0.82	
Intention to explore IPAs	2.15	0.60	0.60	0.54	0.62	0.51	0.85

SB, sociability; INT, interactivity; AU, autonomy; ISE, IPA self-efficacy; SC, social connection; IE, intention to explore IPAs.

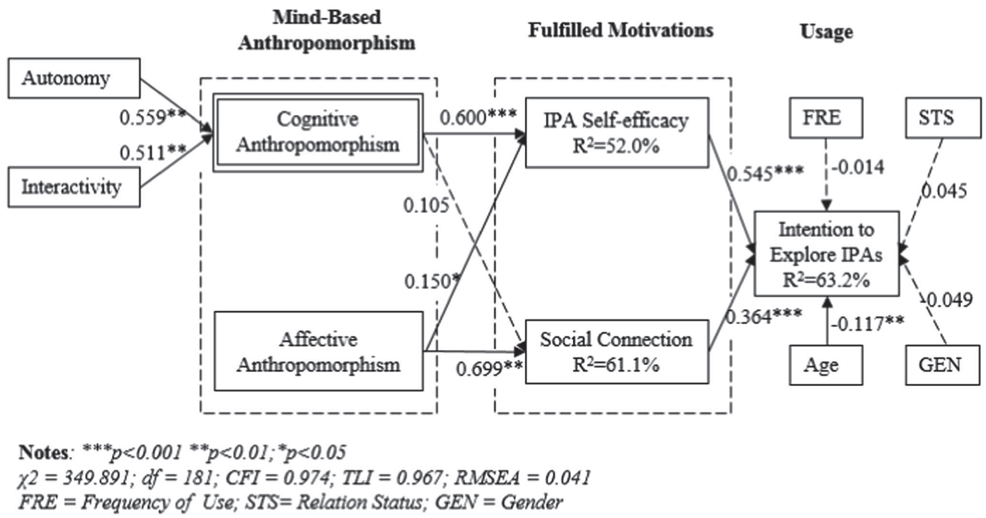


FIGURE 2 | Structural model results.

TABLE 3 | Results of hypotheses test.

Hypothesis	Path coefficient or comparison	C.R.	Hypothesis supported (Y/N)
H1a	$\beta_{CA->ISE} = 0.600$	4.562***	Y
H1b	$\beta_{AA->ISE} = 0.150$	2.185*	Y
H2a	$\beta_{CA->SC} = 0.105$	1.530	N
H2b	$\beta_{AA->SC} = 0.699$	9.508***	Y
H3a	$\beta_{CA->ISE}$ (0.600) > $\beta_{AA->ISE}$ (0.150)	1.786*	Y
H3b	$\beta_{AA->SC}$ (0.699) > $\beta_{CA->SC}$ (0.105)	5.897***	Y
H4	$\beta_{ISE->IE} = 0.545$	11.574***	Y
H5	$\beta_{SC->IE} = 0.364$	8.269***	Y

**p*<0.05; ***p*<0.01; ****p*<0.001 (two-tailed tests, path comparisons used one-tailed tests). CA, cognitive anthropomorphism, AA, affective anthropomorphism, ISE, IPA self-efficacy, SC, social connection, IE, intention to explore IPAs.*

Sung, 2021; Noor et al., 2021). Although most of our hypotheses were supported, the relationship between cognitive anthropomorphism and social connection was proved to be non-significant, which was inconsistent with prior research on anthropomorphism (Cao et al., 2019; Chen and Park, 2021; Li and Sung, 2021; Noor et al., 2021). One possible explanation for the non-significant coefficient between cognitive anthropomorphism and social connection is that when users are majorly driven by sociality motivation, they care less about the cognitive capacities of IPAs. That is, satisfying users' sociality motivation requires fewer cognitive capacities of the IPAs. Meanwhile, different from previous studies which highlight the importance of the cognitive capacities of intelligent agents (Castelo et al., 2019), our study reveals that affective anthropomorphism (affective capacities involved) can significantly affect both IPA self-efficacy and social connection.

Secondly, our results show that cognitive and affective anthropomorphism have differential impacts on IPA self-efficacy and social connection. Specifically, cognitive anthropomorphism exerts stronger impacts on IPA self-efficacy than affective anthropomorphism, while affective anthropomorphism exerts stronger effects on social connection than cognitive anthropomorphism. These results confirm the proposition proposed by Waytz and Young (2014) that different motivations yield a different focus on the different dimensions of the mind attributed to out-groups. Our study validates that in the context of IPAs, cognitive and affective anthropomorphism is motivated by preferential motivations. Further, IPA self-efficacy and social connection exert significantly positive effects on users' intentions to explore IPAs, which are consistent with previous research findings (Wang et al., 2008; Peng et al., 2018; Tams et al., 2018; Tseng et al., 2018).

IMPLICATIONS AND LIMITATIONS

Theoretical Implications

The current study makes several important theoretical implications.

Firstly, our study contributes to IPA research by investigating its exploratory usage. Current studies on IPAs mostly focus on adoption intention (Park et al., 2018; Yang and Lee, 2018; Mclean and Osei-frimpong, 2019; Moussawi and Benbunan-fich, 2020; Moussawi et al., 2020; Mishra et al., 2021; Vimalkumar et al., 2021) and continuous usage of IPAs (Han and Yang, 2018; Moussawi and Koufaris, 2019; Ki et al., 2020; Hu et al., 2021; Sun et al., 2021), which provide very limited knowledge related to how users interact and engage with IPA. Because there are some IPAs whose users only use several basic functions after initial enthusiasm upon adoption, such research is timely and important. The current study is timely to theoretically and empirically examine how users can be deeply engaged with IPAs.

Secondly, our study contributes to prior research on IPA anthropomorphism by validating the common and differential effects of the two dimensions of IPA anthropomorphism on the two fulfilled motivations (IPA self-efficacy and social connection). On the one hand, we found cognitive and affective anthropomorphism can influence users' intention to explore IPAs through IPA self-efficacy and social connection with IPAs. On the other hand, this study empirically validates that the two dimensions of mind-based IPA anthropomorphism can differently satisfy their effectance and sociality motivations (i.e., IPA self-efficacy and social connection). Previous studies indicate that anthropomorphism can enhance efficacy (effectance motivation) and social connection (sociality motivation) with IPAs (Cao et al., 2019; Chen and Park, 2021; Li and Sung, 2021; Noor et al., 2021). However, this study complements by decomposing mind-based anthropomorphism into cognitive and affective anthropomorphism and empirically validates that the two dimensions of mind-based IPA anthropomorphism can differently satisfy their effectance and sociality motivations.

Finally, our study contributes to IPA anthropomorphism by investigating the mediating mechanism between IPA anthropomorphism and the exploratory usage of IPAs. Previous studies identify purposes of using IPAs (utilitarian vs. hedonic), expectation-disconfirmation, and trust as mechanisms for explaining the relationship between anthropomorphism and IPA adoption (Moussawi and Koufaris, 2019; Moussawi and Benbunan-fich, 2020; Moussawi et al., 2020; Mishra et al., 2021). We complement by investigating that IPA self-efficacy and social connection with IPAs can be the mediating mechanisms between anthropomorphism and exploratory usage in the context of IPAs. Considering the mixed findings regarding the relationship between anthropomorphism and adoption-related behaviors (Moussawi et al., 2020; Blut et al., 2021; Hu et al., 2021), our study is timely to empirically investigate these two mediators and future studies are encouraged to study mediators from other perspectives.

Practical Implications

The present study has some practical implications as well.

Firstly, our study provides empirical support for the effectiveness of strategies adopted by service providers to make IPAs more humanlike. Users' mind-based anthropomorphism of IPAs plays an important role in influencing the exploratory usage of IPAs. Considering the great number of functions untapped by users, the service providers need to encourage users' mind-anthropomorphism of IPAs to satisfy their effectance and sociality motivations, which are also important antecedents of intentions to explore IPAs. In other words, adding design features that could increase users' perception of cognitive capacities and affective capacities of IPAs may encourage their in-depth usage. As to the two dimensions of mind-based anthropomorphism, although cognitive and affective anthropomorphism have common effects on IPA self-efficacy and social connection, service providers should also consider their differential effects. Cognitive anthropomorphism facilitates stronger IPA self-efficacy than affective anthropomorphism, while affective anthropomorphism facilitates a stronger social connection than cognitive anthropomorphism. Thus, service providers should pay more attention to the cognitive capacities of IPAs which focus on task efficiency. At the same time, affective capacities should be given more attention to when IPAs are developed for companionship.

Secondly, we found that affective anthropomorphism exerts positive effects on IPA self-efficacy and social connection, while cognitive anthropomorphism only positively affects IPA self-efficacy. This result highlights the importance of the affective capacities that users perceive IPAs to have. Though some studies point out that enhancing the capacities of intelligent agents to "feel" or "experience" might cause users to underestimate their abilities to finish tasks (Castelo et al., 2019), our study reveals that in the IPAs context, the emotional capacities are still important and can result in positive results. This might be because, unlike those intelligent agents designed for specific tasks or contexts, IPAs are used for a wider range of purposes and in more relaxed circumstances. Thus, embedding IPAs with more features that make users perceive that the IPAs can understand their feelings and emotions may be an effective way to enhance their confidence in and connections with them, and further encourage their usage.

Thirdly, our study identifies autonomy, interactivity, and sociability as the specific mental capacities of IPAs, and we believe there are other mental capacities for other types of intelligent agents. Our study can help service providers of IPAs or other intelligent agents to identify specific mental capacities that users highlight in the following two aspects. The first one is that we provide a useful framework (cognitive and affective) to classify these capacities, and service providers can take these two dimensions as overarching guidance. The second one is that the service providers can also make use of the reviews or interviews of users to identify the specific mental capacities as we did in this study. This information provided by users can not only indicate how users use the intelligent agents and what they experience when interacting with these intelligent agents but also provide valuable information about what mental capacities of the intelligent agents' users care about. It provides a bridge that links what the users want and what the designers can do.

Limitations and Future Research

This study also has some limitations. First, smart speaker-based IPAs were chosen as the research objects in our study, and future research is needed to examine whether our research model can be applied to other types of IPAs, which may be used in different contexts with different aims. Second, we use intentions to explore IPAs as the dependent variable in the model, and investigating users' actual exploratory behavior in the future may provide more practical implications. Third, we only investigate the differential effect of two types of anthropomorphism on two different fulfilled motivations. More studies are needed to investigate the differential effects of cognitive anthropomorphism and affective anthropomorphism. For instance, the moral responsibilities of intelligent agents may deserve further investigation (Waytz et al., 2010b). Finally, the specific mental capacities in the current study are posited based on previous studies. Future research is encouraged to apply quantitative content analysis to analyze mental capacities based on new and relevant data such as the latest product reviews of Amazon Echo.

CONCLUSION

The current study sought to investigate how mind-based anthropomorphism of IPAs influences the exploratory usage of IPAs. To this end, we empirically built a research model to investigate the effect of mind-anthropomorphism on the exploratory usage of IPAs through fulfilled motivations of anthropomorphism. The findings reveal that cognitive and affective anthropomorphism exert common and differential impacts on IPA self-efficacy and social connection. Cognitive anthropomorphism versus affective anthropomorphism has stronger influences on IPA self-efficacy, while affective anthropomorphism has stronger impacts on social connection. Both IPA self-efficacy and social connection enhance users' intentions to explore IPAs.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by ethics committees of Huazhong University of Science and Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CC contributed to conceptualization, methodology, software, and writing. YH contributed to the conceptualization and writing of the work. HX contributed the design and writing of the work. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The authors gratefully acknowledge the support of National Natural Science Foundation of China (No. 71701212).

REFERENCES

Ahuja, M. K., and Thatcher, J. B. (2005). Moving beyond intentions and toward the theory of trying: effects of work environment and gender on post-adoption information technology use. *MIS Q.* 29:427. doi: 10.2307/25148691

Amazon. (2021). Amazon Alexa skills. Available at: https://www.amazon.com/alexa-skills/b?node=13727921011 (Accessed October 10, 2021).

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychol. Bull.* 103, 411–423. doi: 10.1037/0033-2909.103.3.411

Aurora Mobile. (2019). Aurora Mobile: Inudustry penetration rate is 2.15%. Available at: http://www.199it.com/archives/872056.html

Balcetis, E., and Dunning, D. (2006). See what you want to see: motivational influences on visual perception. *J. Pers. Soc. Psychol.* 91, 612–625. doi: 10.1037/0022-3514.91.4.612

Baumeister, R. F., and Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychol. Bull.* 117, 497–529. doi: 10.1037/0033-2909.117.3.497

Bellur, S., and Sundar, S. S. (2017). Talking health With a machine: how does message interactivity affect attitudes and cognitions? *Hum. Commun. Res.* 43, 25–53. doi: 10.1111/hcre.12094

Bentler, P. M., and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88, 588–606. doi: 10.1037/0033-2909.88.3.588

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

Businessware. (2020). Global Intelligent Virtual Assistant Market (2020 to 2027). Available at: https://www.businesswire.com/news/home/20200422005328/en/Global-Intelligent-Virtual-Assistant-Market-2020-to-2027---Size-Share-Trends-Analysis-Report---ResearchAndMarkets.com%0A

Cao, C., Zhao, L., and Hu, Y. (2019). "Anthropomorphism of intelligent personal assistants (IPAs): antecedents and consequences". in *Proceedings of the 23rd Pacific Asia Conference on Information Systems*, 187. July 8–12, 2019. Available at: https://aisel.aisnet.org/pacis2019/187 (Accessed October 10, 2021).

Castelo, N. (2019). *Blurring the Line Between Human and Machine: Marketing Artificial Intelligence*. New York, NY: Columbia University.

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *J. Mark. Res.* 56, 809–825. doi: 10.1177/0022243719851788

Changizi, M. A., and Hall, W. G. (2001). Thirst modulates a perception. *Perception* 30, 1489–1497. doi: 10.1068/p3266

Chen, Q. Q., and Park, H. J. (2021). How anthropomorphism affects trust in intelligent personal assistants. *Ind. Manag. Data Syst.* 121, 2722–2737. doi: 10.1108/IMDS-12-2020-0761

Chen, F., Sengupta, J., and Adaval, R. (2013). The effects of Anthropomorphization on brand personality perceptions: a motivational account. *NA – Adv. Consum. Res.* 41, 27–28.

Cho, M., Lee, S., and Lee, K.-P. (2019). "Once a kind friend is now a thing". in *Proceedings of the 2019 on Designing Interactive Systems Conference*, June 23–28, 2019, 1557–1569.

Compeau, D. R., and Higgins, C. A. (1995). Computer self-efficacy: development of a measure and initial test. *MIS Q.* 19, 189–211. doi: 10.2307/249688

Delgosha, M. S., and Hajiheydari, N. (2021). How human users engage with consumer robots? A dual model of psychological ownership and trust to explain post-adoption behaviours. *Comput. Hum. Behav.* 117:106660. doi: 10.1016/j.chb.2020.106660

Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *MIS Q.* 35, 335–358. doi: 10.2307/23044046

Diamantopoulos, A., and Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. *Br. J. Manag.* 17, 263–282. doi: 10.1111/j.1467-8551.2006.00500.x

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

Fornell, C., and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *J Mark. Res.* 18, 39–50. doi: 10.20546/ijcrar.2016.409.006

Han, S., and Yang, H. (2018). Understanding adoption of intelligent personal assistants: A parasocial relationship perspective. *Ind. Manag. Data Syst.* 118, 618–636. doi: 10.1108/IMDS-05-2017-0214

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the Almere model. *Int. J. Soc. Robot.* 2, 361–375. doi: 10.1007/s12369-010-0068-5

Hu, Q., Lu, Y., Pan, Z., Gong, Y., and Yang, Z. (2021). Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants. *Int. J. Inf. Manag.* 56:102250. doi: 10.1016/j.ijinfomgt.2020.102250

Jarvis, C. B., Mackenzie, S. B., and Podsakoff, P. M. (2004). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *J. Consum. Res.* 30, 199–218. doi: 10.1086/376806

Ki, C. C., Cho, E., and Lee, J. (2020). Can an intelligent personal assistant (IPA) be your friend? Para-friendship development mechanism between IPAs and their users. *Comput. Hum. Behav.* 111:106412. doi: 10.1016/j.chb.2020.106412

Latham, G. P., Locke, E. A., and Fassina, N. E. (2000). "The high performance cycle: standing the test of time," in *The Psychological Management of Individual Performance*. ed. S. Sonnentag (Hoboken, NJ: John Wiley & Sons), 201–228.

Lee, R. M., Draper, M., and Lee, S. (2001). Social connectedness, dysfunctional interpersonal behaviors, and psychological distress: testing a mediator model. *J. Couns. Psychol.* 48, 310–318. doi: 10.1037/0022-0167.48.3.310

Li, X., and Sung, Y. (2021). Anthropomorphism brings us closer: the mediating role of psychological distance in user–AI assistant interactions. *Comput. Hum. Behav.* 118:106680. doi: 10.1016/j.chb.2021.106680

Lindell, M. K., and Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *J. Appl. Psychol.* 86, 114–121. doi: 10.1037/0021-9010.86.1.114

Luczak, H., Roetting, M., and Schmidt, L. (2003). Let's talk: anthropomorphization as means to cope with stress of interacting with technical devices. *Ergonomics* 46, 1361–1374. doi: 10.1080/00140130310001610883

Maarek, Y. (2018). "Alexa and her shopping journey." in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, October 22–26, 2018.

Maarek, Y. (2019). "Alexa, can you help me shop?" in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information*, June 21–25, 2019, 1369–1370.

Mclean, G., and Osei-frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Comput. Hum. Behav.* 99, 28–37. doi: 10.1016/j.chb.2019.05.009

Mishra, A., Shukla, A., and Sharma, S. K. (2021). Psychological determinants of users' adoption and word-of-mouth recommendations of smart voice assistants. *I. J. Inf. Manage.* 102413. doi: 10.1016/j.ijinfomgt.2021.102413

Mourey, J. A., Olson, J. G., and Yoon, C. (2017). Products as pals: engaging with anthropomorphic products mitigates the effects of social exclusion. *J. Consum. Res.* 44, 414–431. doi: 10.1093/jcr/ucx038

Moussawi, S., and Benbunan-fich, R. (2020). The effect of voice and humour on users' perceptions of personal intelligent agents. *Behav. Inform. Technol.* 40, 1603–1626. doi: 10.1080/0144929X.2020.1772368

Moussawi, S., Koufaris, M., and Benbunan-fich, R. (2020). How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electron. Mark.* 31, 343–364. doi: 10.1007/s12525-020-00411-w

Moussawi, S., and Koufaris, M. (2019). "Perceived intelligence and perceived anthropomorphism of personal intelligent agents: scale development and validation." in *Proceedings of the Annual Hawaii International Conference on System Sciences*, January 8–11, 2019, 115–124.

Nambisan, S., Agarwal, R., and Tanniru, M. (1999). Organizational mechanisms for enhancing user innovation in information technology. *MIS Q.* 23, 365–395. doi: 10.2307/249468

Noor, N., Rao Hill, S., and Troshani, I. (2021). Artificial intelligence service agents: role of Parasocial relationship. *J. Comput. Inf. Syst.* 00, 1–15. doi: 10.1080/08874417.2021.1962213

Nunnally, J. C. (1978). *Psychometric Theory*. New York, NY: McGraw-Hill.

Pan, Z., Lu, Y., Wang, B., and Chau, P. Y. K. (2017). Who do you think you are? Common and differential effects of social self-identity on social media usage. *J. Manag. Inf. Syst.* 34, 71–101. doi: 10.1080/07421222.2017.1296747

Park, K., Kwak, C., Lee, J., and Ahn, J. H. (2018). The effect of platform characteristics on the adoption of smart speakers: empirical evidence in South Korea. *Telematics Inform.* 35, 2118–2132. doi: 10.1016/j.tele.2018.07.013

Peng, Z., Sun, Y., and Guo, X. (2018). Antecedents of employees' extended use of enterprise systems: an integrative view of person, environment, and technology. *Int. J. Inf. Manag.* 39, 104–120. doi: 10.1016/j.ijinfomgt.2017.11.007

Petter, S., Straub, D., and Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Q.* 31, 623–656. doi: 10.2307/25148814

Postman, L., Bruner, J. S., and McGinnies, E. (1948). Personal values as selective factors in perception. *J. Abnorm. Soc. Psychol.* 43, 142–154. doi: 10.1037/h0059765

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. (2017). "Alexa is my new BFF": social roles, user satisfaction, and personification of the Amazon Echo. in *CHI'17 Extended Abstracts*, May 6–11, 2017, 2853–2859.

Richins, M. L. (1994). Valuing things: the public and private meanings of possessions. *J. Consum. Res.* 21, 504–521. doi: 10.1086/209414

Rijsdijk, S. A., Hultink, E. J., and Diamantopoulos, A. (2007). Product intelligence: its conceptualization, measurement and impact on consumer satisfaction. *J. Acad. Mark. Sci.* 35, 340–356. doi: 10.1007/s11747-007-0040-6

Rusbult, C. E. (1980). Commitment and satisfaction in romantic associations: a test of the investment model. *J. Exp. Soc. Psychol.* 16, 172–186. doi: 10.1016/0022-1031(80)90007-4

Schaefers, K. G., Epperson, D. L., and Nauta, M. M. (1997). Women's career development: can theoretically derived variables predict persistence in engineering majors? *J. Couns. Psychol.* 44, 173–183. doi: 10.1037/0022-0167.44.2.173

Shani, C., Libov, A., Tolmach, S., Lewin-Eytan, L., Maarek, Y., and Shahaf, D. (2021). "Alexa, what do you do for fun?" Characterizing playful requests with virtual assistants. Available at: http://arxiv.org/abs/2105.05571 (Accessed March 10, 2022).

Sun, Y., Li, S., and Yu, L. (2021). The dark sides of AI personal assistant: effects of service failure on user continuance intention. *Electron. Mark.* 2018, 1–23. doi: 10.1007/s12525-021-00483-2

Sundar, S. S., Bellur, S., Oh, J., Jia, H., and Kim, H. (2016). Theoretical importance of contingency in human-computer interaction: effects of message interactivity on user engagement. *Commun. Res.* 43, 595–625. doi: 10.1177/0093650214534962

Tams, S., Thatcher, J. B., and Craig, K. (2018). How and why trust matters in post-adoptive usage: the mediating roles of internal and external self-efficacy. *J. Strateg. Inf. Syst.* 27, 170–190. doi: 10.1016/j.jsis.2017.07.004

Tseng, F. C., Pham, T. T. L., Cheng, T. C. E., and Teng, C. I. (2018). Enhancing customer loyalty to mobile instant messaging: perspectives of network effect and self-determination theories. *Telematics Inform.* 35, 1133–1143. doi: 10.1016/j.tele.2018.01.011

Vimalkumar, M., Sharma, S. K., Singh, J. B., and Dwivedi, Y. K. (2021). "Okay google, what about my privacy?": user's privacy perceptions and acceptance of voice based digital assistants. *Comput. Hum. Behav.* 120:106763. doi: 10.1016/j.chb.2021.106763

Wagner, K., Nimmermann, F., and Schramm-klein, H. (2019). Is it human? The role of anthropomorphism as a driver for the successful acceptance of digital voice assistants. in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, January 8–11, 2019, 1386–1395.

Wagner, K., and Schramm-klein, H. (2019). "Alexa, are you human? Investigating the anthropomorphism of digital voice assistants – a qualitative approach." *Fortieth International Conference on Information Systems*, 1–17.

Wang, W., Butler, J. E., Hsieh, J. P. A., and Hsu, S. H. (2008). Innovate with complex information technologies: A theoretical model and empirical examination. *J. Comput. Inf. Syst.* 49, 27–36. doi: 10.1080/08874417.2008.11645303

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

Waytz, A., Gray, K., Epley, N., and Wegner, D. M. (2010b). Causes and consequences of mind perception. *Trends Cogn. Sci.* 14, 383–388. doi: 10.1016/j.tics.2010.05.006

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., and Cacioppo, J. T. (2010c). Making sense by making sentient: effectance motivation increases anthropomorphism. *J. Pers. Soc. Psychol.* 99, 410–435. doi: 10.1037/a0020240

Waytz, A., and Young, L. (2014). Two motivations for two dimensions of mind. *J. Exp. Soc. Psychol.* 55, 278–283. doi: 10.1016/j.jesp.2014.08.001

White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychol. Rev.* 66, 297–333. doi: 10.1037/h0040934

Yang, H., and Lee, H. (2018). Understanding user behavior of virtual personal assistant devices. *Inf. Syst. e-Bus. Manag.* 17, 65–87. doi: 10.1007/s10257-018-0375-1

Zhao, L., Lu, X., and Hu, Y. (2018). A proposed theoretical model of discontinuous usage of voice-activated intelligent personal assistants (IPAs). in *Twenty-*

Second Pacific Asia Conference on Information Systems 2018 Proceedings, 245. June 26–30, 2018.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX

Appendix A | Instrument measurement items

TABLE A1 | Measurement Instruments.

Construct	Items	Source
Autonomy (AU)	IPAs provide auto-adjusted control IPAs do things semi-autonomously by itself IPAs help the users proactively without human intervention	(Rijsdijk et al., 2007)
Interactivity (INT)	IPAs' responses were related to my earlier responses IPAs took into account my previous interactions with it IPAs gave some smart suggestions based on my responses	(Bellur and Sundar (2017))
Sociability (SB)	I consider the IPAs a pleasant conversational partner I find IPAs pleasant to interact with I think IPAs are nice	(Heerink et al., 2010)
IPA Self-Efficacy(ISE)	I can use most skills of IPAs if there was no one around to tell me what to do as I go I can use most skills of IPAs if I had the tips and experiences from online users for reference. I can use most skills of IPAs if I could call someone for help if I got stuck.	(Compeau and Higgins, 1995)
Social Connection (SC)	I feel close to IPAs I feel socially connected with IPAs I feel related to IPAs	(Lee et al., 2001)
Intention to Explore IPAs (IE)	I intend to explore IPAs for other potential applications I intend to find some new uses of IPAs I intend to spend some time and effort this year in exploring new functions of IPAs	(Nambisan et al., 1999)

Appendix B | Construct reliability and validity

TABLE B1 | Reliability and Validity.

Construct	Items	Loadings	AVE	CR	Cronbach's Alpha
Sociability	SB1	0.733	0.632	0.837	0.832
	SB2	0.797			
	SB3	0.850			
Interactivity	INT1	0.856	0.662	0.854	0.853
	INT2	0.774			
	INT3	0.809			
Autonomy	AU1	0.774	0.604	0.820	0.823
	AU2	0.721			
	AU3	0.832			
IPA Self-Efficacy	ISE1	0.777	0.639	0.841	0.835
	ISE2	0.873			
	ISE3	0.743			
Social Connection	SC1	0.817	0.674	0.861	0.860
	SC2	0.817			
	SC3	0.829			
Intention to Explore IPAs	IE1	0.890	0.725	0.887	0.885
	IE2	0.845			
	IE3	0.817			

Appendix C | Confirmatory factor analysis for cognitive anthropomorphism

TABLE C1 | Confirmatory factor analysis for cognitive anthropomorphism.

Fit index	Cutoff	First-order	Second-order reflective	Second-order formative
CMIN/DF	<3	2.089	2.232	2.088
CFI	>0.9	0.985	0.983	0.985
TLI	>0.9	0.980	0.977	0.980
RMSEA	<0.08	0.045	0.047	0.044

For cognitive anthropomorphism, three models were estimated and compared, namely, the first-order model, second-order reflective model, and second-order formative model. In covariance-based SEM, it is required that a formative construct has two emitting paths so that the model could be identified (Diamantopoulos, 2011; Jarvis et al., 2004). Thus, we included IPA self-efficacy and social connection and made cognitive anthropomorphism point to them since we did not have any reflective indicators for cognitive anthropomorphism. As Table C1 shows, the second-order formative model fitted best among the three models. Therefore, we chose the second-order formative model for cognitive anthropomorphism.

frontiers | Frontiers in Computer Science

ORIGINAL RESEARCH
published: 13 May 2022
doi: 10.3389/fcomp.2022.778367

Speech Assistant System With Local Client and Server Devices to Guarantee Data Privacy

*Hans-Günter Hirsch**

Institute for Pattern Recognition, Niederrhein University of Applied Sciences, Krefeld, Germany

Users of speech assistant systems have reservations about the distributed approach of these systems. They have concerns that people might get access to the transmitted speech data or that somebody is able to access their microphone from outside. Therefore, we investigate the concept of a setup with local client and server systems. This comes along with the requirement of cost-efficient realizations of client and server. We examined a number of different cost-efficient server solutions depending on the required recognition capability of specific applications. A fairly cost-efficient solution is the use of a small computing device for recognizing a few dozens of words with a GMM-HMM based recognition. To perform a DNN-HMM based recognition, we looked at small computing devices with an integrated additional graphical processor unit (GPU). Furthermore, we investigated the use of low-cost PCs for implementing real-time versions of the Kaldi framework to allow the recognition of large vocabularies. We investigated the control of a smart home by speech as an exemplary application. For this, we designed compact client systems that can be integrated at certain places inside a room, e.g., in a standard outlet socket. Besides activating a client by a sensor that detects approaching people, the recognition of a spoken wake-up word is the usual way for activation. We developed a keyword recognition algorithm that can be implemented in the client despite its limited computing resources. The control of the whole dialogue has been integrated in our client, so that no further server is needed. In a separate study, we examined the approach of an extremely energy-efficient realization of the client system without the need of an external power supply. The approach is based on using a special microphone with an additional low-power operating mode detecting the exceeding of a preset sound level threshold only. This detection can be used to wake up the client's microcontroller and to make the microphone switch to normal operating mode. In the listening mode, the energy consumption of the microphone is so low that a client system can be active for months with an energy supply from standard batteries only.

Keywords: speech assistant, local recognition, compact client, keyword recognition, energy efficient client

OPEN ACCESS

Edited by:
Benjamin Weiss, Technical University of Berlin, Germany

Reviewed by:
Stefan Schaffer, German Research Center for Artificial Intelligence (DFKI), Germany
Matthias Wolff, Brandenburg University of Technology Cottbus-Senftenberg, Germany

****Correspondence:***
Hans-Günter Hirsch
hans-guenter.hirsch@hs-niederrhein.de

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

Received: *16 September 2021*
Accepted: *14 April 2022*
Published: *13 May 2022*

Citation:
Hirsch H-G (2022) Speech Assistant System With Local Client and Server Devices to Guarantee Data Privacy. Front. Comput. Sci. 4:778367. doi: 10.3389/fcomp.2022.778367

1. INTRODUCTION

Right now, the way to omnipresent speech assistants is determined by special hardware realizations like the Echo devices by Amazon or the Google home devices as well as by software realizations like Siri by Apple or Cortana by Microsoft. Most of these solutions consist of three components. The first component contains the hardware with microphones and loudspeaker to record and playback speech. This component is referred to as client. The recorded speech signal is preprocessed in the client to reduce the effects of background noise and reverberation. This is usually done by recording the speech with several microphones and applying multi-channel processing techniques. Furthermore, an algorithm is implemented in the client to perform the detection and recognition of a keyword that is used to wake up the assistant.

After wake-up, the preprocessed speech signal is usually transmitted *via* IP through a public network to a speech recognition server, which represents the second component of the entire system. The advantage of this approach is the application of a server configuration with extremely high computational performance, so that powerful recognition algorithms can be applied to enable high recognition performance. However, the data are transported to an external server *via* a public network, which means that it is not clear who gets access to the signal and what the signal could be used for besides its input to the recognition system. Furthermore, users have concerns that somebody can get access to their microphone from outside and can record and analyze audio when the speech assistant is not active (Chung et al., 2017; Lau et al., 2019; Malkin et al., 2019; Hernández Acosta and Reinhardt, 2020). The strength of this concern varies and depends on cultural and country-specific behavior of people. This leads to a high percentage of people in certain countries unwilling to use such systems, although they are not refusing speech technology in general. A number of approaches have been developed and investigated for the case of speech transmission through a public network. An overview about the privacy-by-design technology is given in Nautsch et al. (2019). The encryption of data (Nautsch et al., 2018; Bäckström et al., 2020) is an obvious approach to reduce the concern that somebody else besides the receiver can get access to the speech data or the recognition result. The binarization and protection of i-vectors (Mtbiaa et al., 2021) is an example for a so called cancelable biometric system. Speaker de-identification (Bahmaninezhad et al., 2018) represents another approach to privacy preservation. Furthermore, hardware based techniques can be applied like the software guard extension in Intels processor units (Brasser et al., 2018).

The result of the speech recognition is sent to a third component, e.g., as a text string. The dialogue between user and speech assistant is controlled by this additional server component. This server takes over several tasks. The first task is the interpretation of the received text string to find out what the user wants to know or intends to do. For example, if a user has uttered the sentence "What will be the weather in Krefeld tomorrow?" and if the recognition component has perfectly recognized the sequence of words, the task will be the correct interpretation of this word sequence to enable the search for the desired information. The inquiry can be formulated in many different ways, so that powerful natural language processing is needed at this point. The next task will be the acquisition of the requested information from a data base or another server. Then, a sentence has to be formulated as answer containing the acquired information. The sentence is further transferred to speech with means of a "Text-to-Speech" (TTS) algorithm. Finally, the speech signal is sent to the client to create an acoustic output as feedback to the user's inquiry. The example described before contains the retrieval of information as a frequent task given to a speech assistant. Besides this, users want to apply the speech assistant to control hardware devices at home or at a certain location. In this case, the third component may not need to or does not only have to create acoustic feedback. Its main task is the creation of a command that has to be sent to the hardware device. Often, this is not possible *via* direct communication between the dialogue component of the assistant system and the hardware device. Instead, another server system is needed that has access to and is able to control the hardware components.

In our investigation, we focus on the application of speech assistants to control hardware devices. This can be, for example, the control of a smart home environment. Our approach differs from the behavior of most commercial systems as they have been described before in two respects. First, we are investigating the concept of a setup with local client and server systems, which makes it possible to guarantee users that their speech is not transmitted outside their private networks and that nobody can access their microphones. To realize this concept, several requirements must be fulfilled, the main one being the application of a recognition server system that is affordable for private users on one hand and that guarantees a fairly high recognition performance on the other hand. As a second point, we want to simplify the whole structure of the system. Usually, commercial systems have to include the client and three server systems as described before. The possibility but also the difficulty are presented in Seiderer et al. (2020) to set up such a configuration as a local system with several open source components. Due to developing and integrating the needed components including the speech recognition ourselves we achieve a more compact and more flexible configuration in comparison to combining available open source components (Seiderer et al., 2020). Besides the recognition server, two additional servers are needed, one of them for the dialogue control including the speech interpretation and the other for accessing the hardware components. We combined the dialogue control and the communication with the hardware components in the client system. Thus, we can reduce the system to only two components including a client and a recognition server.

2. SYSTEM OVERVIEW

Figure 1 gives an overview of the goal of our project and the corresponding structure of the system. We want to allow the control of devices like shutters, lights, and any other type of

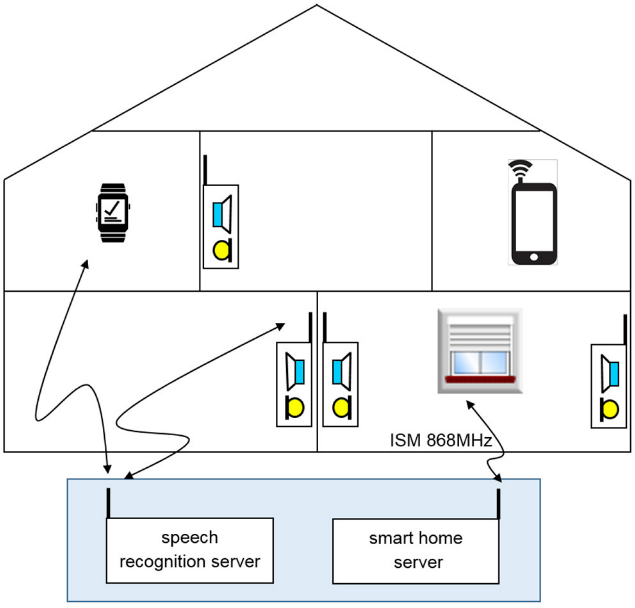


FIGURE 1 | Smart home control with local client and server systems.

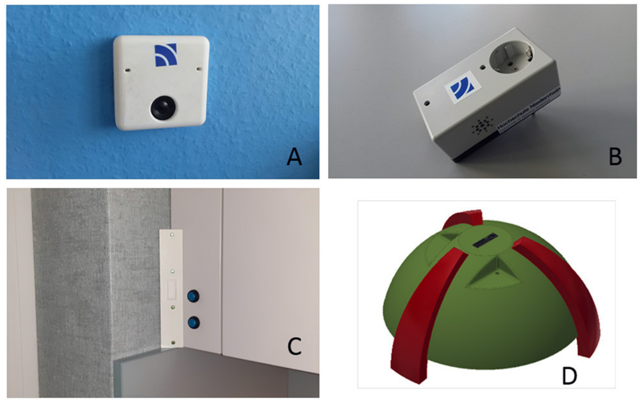


FIGURE 2 | Implementation of client devices in an outlet socket **(A)**, as a switchable power socket **(B)**, in a furniture corner **(C)**, and as a stand-alone device **(D)**.

actor by speech input from all rooms or locations in or nearby a building.

Therefore, we developed client systems with the capability of speech input and output that can be installed in each room. Our intention was to integrate these client devices in the existing electrical infrastructure as far as possible, e.g., as substitute of switches in existing outlet sockets. For the case that it is not possible to integrate them inside a room, we developed stand-alone client devices that do not need external energy supply. This leads to the requirement of an extremely low energy consumption, so that we investigated this aspect in a separate study. To cover also rooms or locations in the building that are rarely used or have almost no electrical infrastructure, we developed a software component for smart phones or smart watches where this component offers the same functionality as other client devices. Each client device can communicate with the in-house recognition server *via* LAN or WLAN. We look at the alternative of communicating and transferring speech *via* the DECT-ULE (Digital Enhanced Cordless Telephony- Ultra Low Energy) standard (ETSI, 2019) at a later point in this paper. Each client includes a dialogue control module, so that the device does not only accept the input of a single spoken command and plays back an acoustic reaction but can also manage a longer speech dialogue with the user. The client can send control commands to the local smart home server *via* LAN/WLAN. Furthermore, the direct switching *via* a relay is included in case of an integrated device as substitute for an existing switch. Cost and energy efficiency are the two main requirements for the design of the client systems. The target of low costs is also the main requirement for the choice of the local speech recognition server. We investigated different possibilities for realizing the speech recognition server depending on the demand of the application-specific recognition task. We applied our own realizations of phoneme-based GMM-HMM or DNN-HMM based recognition schemes when the recognition of smaller vocabularies containing up to a few hundred of words is needed. For cases in which the recognition task demands a larger vocabulary, we applied a Kaldi based recognition scheme (Povey et al., 2011) on a low-cost computer.

3. CLIENT

The goal of our investigation is the development of very compact client devices that fulfill the requirements of low cost and low energy consumption and that can be integrated in the existing electrical infrastructure of a building. After presenting the hardware setup, we will focus on the implemented software modules, especially the algorithm for detecting and recognizing the wake-up word. Furthermore, we will present the results of two studies to minimize the energy consumption of a stand-alone client and to communicate with a recognition server *via* DECT-ULE.

3.1. Hardware

Some of the devices we have developed are shown in **Figure 2**. **Figure 2A** shows the version of the client that can be integrated in an outlet socket. It can be taken as substitute for an existing switch. **Figure 3** is a block diagram containing all components of this version.

The small computer PI-Zero is used as a basis for this client and a separate small device with two microphones (Seeed, 2021) is applied to record the audio signal. A class-D amplifier is connected to the analog output of the sound device to allow audio output through the loudspeaker. Furthermore, a relay has been integrated that realizes the switching in case the client serves as substitute for an existing switch. The WLAN interface of the PI-Zero is used for communication with the recognition server. **Figure 2B** shows the integration of the client in a switchable

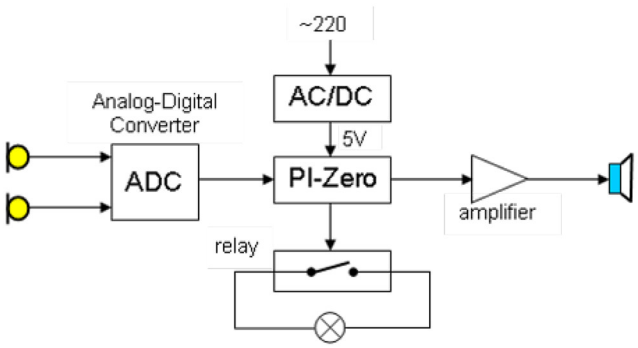


FIGURE 3 | Components of a client device.

power socket. The hardware setup is similar to the one shown in **Figure 3** but it contains two loudspeakers at the sides of the housing. **Figure 2C** shows the housing for an array of four microphones (Seeed, 2021) that is placed at the corner of kitchen furniture. A PI computer device is placed in the cavity behind the cover element in this furniture corner. **Figure 2D** shows a stand-alone version of the client, which is again based on a PI computer. The speech is recorded by an array of four microphones. The system is provided with energy by a rechargeable power bank. Later in this chapter, we will present results of a study where we investigated the application of a special microphone in combination with specially designed algorithms to setup a client system with extremely low energy consumption.

3.2. Dialogue Control

Most client devices run on Linux. The main software component is responsible for controlling the dialogue between user and client. The dialogue is modeled as a finite state machine. The states and their chronology are described by a text file, so that the dialogue can be easily defined and modified. An action is assigned to each state such as speech input, speech output, or sending a command sequence to a hardware device in the building or to the smart home server. This setup allows the individual configuration of each client, so that the dialogue can be defined depending on the room or the location of the client. For this application, we do not need a text-to-speech component in our client systems. Speech output is realized with pre-recorded audio files. The activation of the dialogue is an important feature of the client. We investigated the usage of sensors in our first versions to detect an approaching person. These sensors use ultrasound or infrared as basis for the detection. The integration and application of such sensors is useful in rooms or at locations where it is very likely that a person approaches the place with the intention to control a hardware device. In general, the recognition of a spoken keyword is the typical way of activating the client. Often, the keyword is called the "wake-up" word. The recognition of the wake-up word could be realized through the server system. However, this would usually lead to a high data traffic from the client to the server due to the permanent speech transmission or at least the transmission of the speech signal in all segments where a Voice Activity Detector (VAD) on the client assumes the presence of speech. Alternatively, certain stages of the recognition process could be distributed between client and server. In our investigation, we realized the recognition of the wake-up word in the client system, which requires an algorithm that can be implemented despite the limited computational performance.

3.3. Keyword Recognition

The performance of keyword recognition can be measured by the false acceptance rate (FAR) and the false rejection rate (FRR). In our application, where we want to apply the keyword detection for the activation of a home automation system, we prioritized lowering the FAR. Hence, we have to avoid any command recognition after an erroneous keyword detection because this could lead to the uncontrolled activation of devices at home. We developed an algorithm for the detection and the recognition of the keyword that consists of two stages as shown in **Figure 4**.

To extract characteristic features, the speech signal is sampled at a rate of 16 kHz. The short-term DFT spectra are calculated for frames containing 400 samples (=25 ms) every 10 ms after applying a pre-emphasis filtering and weighting the 400 samples of each segment with a Hamming window. The 400 filtered and weighted samples are transformed by means of a DFT with a length of 512. The short-term logarithmic energy logE is calculated by taking the logarithm of the sum of the squared DFT magnitude coefficients in the range between 200 Hz and 7 kHz. Furthermore, 12 cepstral coefficients C1–C12 are determined by transforming the logarithmic MEL spectrum with a DCT. Thirty-six MEL filters have been defined in the range from 200 Hz to about 7 kHz to calculate the MEL spectrum from the magnitude DFT coefficients. The Delta coefficients (ΔlogE, ΔC1, ..., ΔC12) and the second derivative of the energy contour $\Delta\Delta$logE are calculated according to the filtering scheme defined in ETSI (2003). The vector containing the 26 components (C1, ..., C12, ΔlogE, ΔC1, ..., ΔC12, $\Delta\Delta$logE) is used as feature vector for the GMM-HMM recognizer. The energy coefficient logE is omitted due to its varying value in case of background noise.

As the keyword, we chose the personal Name "Esmeralda," which is rarely used as a given name in Germany and which rarely occurs in German conversations. The word begins and ends with a vowel. It contains a fairly long sequence of phonemes, which supports a better recognizability. We created two Hidden Markov Models that represent the keyword and which are applied for the realization of the first recognition stage. The first model was created from real recordings of the keyword and from augmented versions of these recordings with HTK (Young et al., 2006). For data augmentation, we applied a tool (Hirsch and Finster, 2005) to create versions containing noise and reverberation, as these occur in real scenarios when recording in hands-free mode. The second keyword HMM is built as a concatenation of the corresponding triphone HMMs from a phoneme-based recognizer that has been trained on several hundred hours of German speech with HTK. The sequence of feature vectors and the keyword HMMs are taken to set up a GMM-HMM recognizer as first stage of the recognition algorithm. A set of 25 monophone HMMs and some pause and noise models are included as so-called filler models (Rose and Paul, 1990). The intention is the

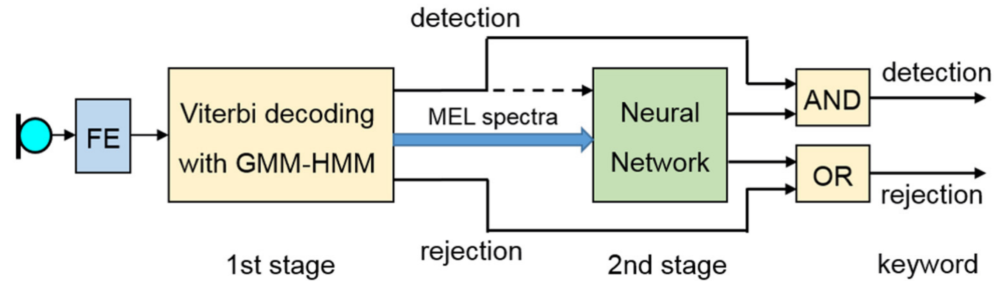


FIGURE 4 | Two-stage keyword recognition (FE, feature extraction).

modeling of speech not containing the keyword as sequence of filler and/or pause HMMs with a higher probability as with one of the keyword HMMs. Only the spoken keyword should lead to a higher probability when modeling it with one of the keyword HMMs. We implemented a fast notification at the end of a spoken keyword, so that further processing is not delayed. In case the GMM-HMM recognition stage indicates the detection of a keyword, we try to verify this assumption by a second stage. It is known that the modeling of speech by means of filler models works quite well, but the expected FAR of this approach is too high for the intended application. Therefore, we apply a neural network as the main component of the second stage. As input, we use the sequence of logarithmic MEL spectra within the speech segment that should contain the keyword according to the recognition result of the first stage. To get a fixed number of input coefficients for the neural network, we reduce the number of Mel spectra to 50 by iteratively calculating the mean of two consecutive logarithmic MEL spectra with the lowest spectral City block distance. **Figure 5** presents the average spectrum that has been calculated over 850 utterances of the keyword used to create one of the keyword HMMs. We observe a very characteristic spectral pattern where the spectral characteristics of each individual phoneme become clearly visible.

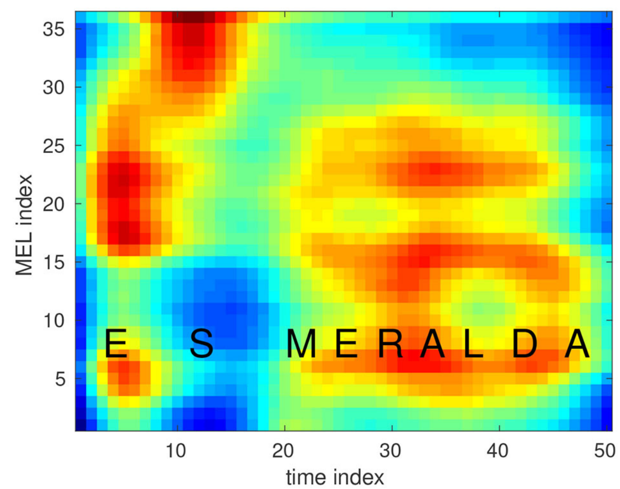


FIGURE 5 | Average MEL spectrogram of the keyword.

The input for the neural network consists of 1,800 spectral amplitudes from 50 spectra with 36 MEL coefficients. We apply a mean and variance normalization to each spectral pattern by calculating the mean and the variance over all 1,800 spectral parameters of each individual pattern. We apply a fully connected multi-layer perceptron, consisting of three layers. The first layer consists of 200 nodes, the second of 50 nodes and the output of two nodes for the two cases of a keyword and a non-keyword. To train the weights of the neural network, we needed spectral patterns for spoken keywords as well as for segments where the keyword was erroneously detected by the first stage. About 850 spectral patterns of the spoken keyword could be determined from the utterances that have been used for training the keyword HMM. To get spectrograms of speech segments where the keyword was not spoken, we applied the detection algorithm of the first stage to German speech data from different databases (Burger and Schiel, 1998; Radeck-Arneth et al., 2015). Several thousands of segments were erroneously detected. Thus, we had about 850 examples of the keyword spectrogram and several thousand examples of the non-keyword spectrogram available.

We applied the tools of Chollet (2015) to estimate the weights of the neural network. We achieved a FAR of less than one keyword per hour of speech as the result of these simulation experiments (Hirsch et al., 2020). Then, we implemented and ran the algorithm for the keyword recognition on some of the client devices for several weeks in a laboratory and in a living room. False detections are observed in cocktail party situations where a lot people are talking in the background or in situations with an active TV or radio. We stored the speech segments and the corresponding MEL spectrograms when a keyword was correctly or erroneously detected. Using these additional data for retraining the network, we could show that the recognition performance can be steadily increased by including more and more recorded data from real life scenarios (Hirsch et al., 2020).

3.4. Energy Efficient Client

So far, we applied PI computers as the basic component for the client systems. We performed a study to find out whether a client can be operated over a period of several months with extremely low energy resources. A setup with two standard batteries of type AAA was taken as energy source. Each battery offers a voltage of 1.5 V and has an energy capacity of about 1.8 Wh. PI-based systems can be operated for about an hour given the energy resource of 3.6 Wh from two batteries and assuming an energy consumption of about 3–5 W by the PI. The PI-Zero consumes about 1 W, so that it can be operated for a few hours. Obviously, other solutions are needed at this point. This study aims at the development of a concept including hardware components and algorithmic approaches with their implementation as software to reach the goal of an extremely low energy consumption.

A lot of microcontroller units (MCU) exist that have been optimized with respect to energy consumption. We looked at a MCU with an ARM Cortex-M4 processor (Microchip, 2020a) as an example for such a device. The MCU can be operated at a voltage of less than 2 V. It can run in different modes including a special power mode where the current is dependent on the clock frequency at which the MCU runs. The value for the current is specified as 65 μA per Mhz by the manufacturer. The highest clock frequency of the MCU device is 120 MHz. Even if we would be able to realize the permanent listening for detecting the wake-up word with an efficient algorithm that could run at a clock frequency of just a few MHz, the energy consumption would be too high for operating the client over a longer period of, e.g., a few months. We started thinking about a separate analog circuit that allows the wake-up of the MCU only if the sound level exceeds a certain threshold. During our investigations of already existing solutions, we came up with the special microphone VM3011 (Vesper, 2021). This MEMS microphone and its predecessor model VM1010 have different operating modes. The VM3011 can deliver digital sample data to an MCU in its normal operating mode due to an integrated ADC. Besides this, the microphone can operate at a so-called zero-power listening mode, in which only the sound level is determined. After exceeding a certain sound level threshold the microphone sets a digital output to wake up the MCU. Then, the MCU can initiate the mode switching at the microphone. Optionally, a filtering can be enabled during the determination of the sound level. Thus, the level estimation can be focused on the frequency range of speech approximately. The sound level detection and the mode switching can be executed in the short period of a few milliseconds, so that only a short segment at the speech onset is lost. Furthermore, the VM3011 contains the feature of adapting the sound level threshold to the sound scenario in its environment. Thus, the threshold will be automatically increased in the presence of stationary background noise.

The remarkable feature of the microphone is a current of just 10 μA when running in the zero power listening mode. This allows the permanent operation over an extremely long period. Based on this microphone, we developed a processing scheme consisting of several processing stages with increasing power consumption. The processing aims at the detection and recognition of a wake-up word and the recognition of a command word or command phrase in a succeeding phase, e.g., as input for the smart home control.

As first stage, we apply a VAD algorithm to detect the beginning of speech (Hirsch and Ehrlicher, 1995). The speech signal is sampled at 10 kHz. This algorithm is based on a rough spectral analysis in 15 subbands in the frequency range from about 300 Hz to 5 kHz. The spectral analysis is performed on frames of 128 samples by applying a DFT of a length 32 on the sum of the four accumulated subframes with length 32. The energy of the background noise is estimated in each subband every 12.8 ms by looking at the smoothed energy contour. This estimation is used to define and adapt an energy threshold. In case of exceeding this threshold, the subband is considered as active. If a predefined number of active subbands is detected, the corresponding frame is considered as speech frame. The beginning of speech is indicated when speech has been detected in several consecutive frames over a period of about 100 ms. Based on the count of the needed processor cycles, we can determine a clock frequency of less than 1 MHz to run the very efficient algorithm on the MCU. We take into account the usage of the CMSLIB library (Lorenser, 2016) that has been developed for this type of ARM Cortex-M4 processor including a floating point unit. The library contains software modules for different signal processing algorithms. Detailed information is available about the number of processor cycles to realize, for example, a DFT with this library. The first processing stage can be realized with an extremely low energy consumption of the MCU. In comparison, the MEMS microphone needs a much higher current of about 700 μA in its normal operating mode.

The speech samples are buffered during the period of about 100 ms in which the beginning of speech is detected. Mel spectral features are extracted from the buffered and the succeeding samples as second processing step. Twenty-four logarithmic MEL spectral values are determined in the frequency range from about 200 Hz to 5 kHz. A DFT is applied to frames of 256 samples. The analysis window is shifted by 12.8 ms, so that we receive about 78 MEL spectra per second. We estimated again the computational resources by counting the operations that are needed for the realization of this second processing stage. The feature extraction can be implemented on the MCU at a clock frequency of $<$2 MHz.

As third processing stage, the feature vectors are fed into a neural network to perform either the recognition of the wake-up word or the recognition of a command phrase (Hwang et al., 2015; Sainath and Parada, 2015). The networks needed for the two tasks only differ in the number of nodes at the output layer. The structure of the network is shown in **Figure 6**.

The MEL spectra are fed into a LSTM layer with its recurrent structure to analyze and evaluate the sequence of MEL spectra. The neural network has three nodes at its fully-connected output layer for the recognition of the wake-up word. Due to a softmax scaling in the last layer, the output can be treated as the three probabilities that the sequence of MEL spectra contain the wake-up word, a speech pause or a non-keyword. The second network for the recognition of the command phrases has as many nodes at its output as the number of different commands plus one node for the speech pause and one node for the class of non-keywords. We examined the exemplary recognition of 20 German words. The 22 output values of the network for the spoken word "korrigieren" are shown in **Figure 7**. The algorithmic approach as described before has been realized with Matlab including the detection of the speech begin, the feature extraction and the recognition with the neural network. The training of the network

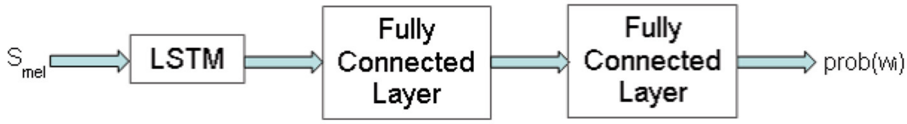


FIGURE 6 | Structure of neural network.

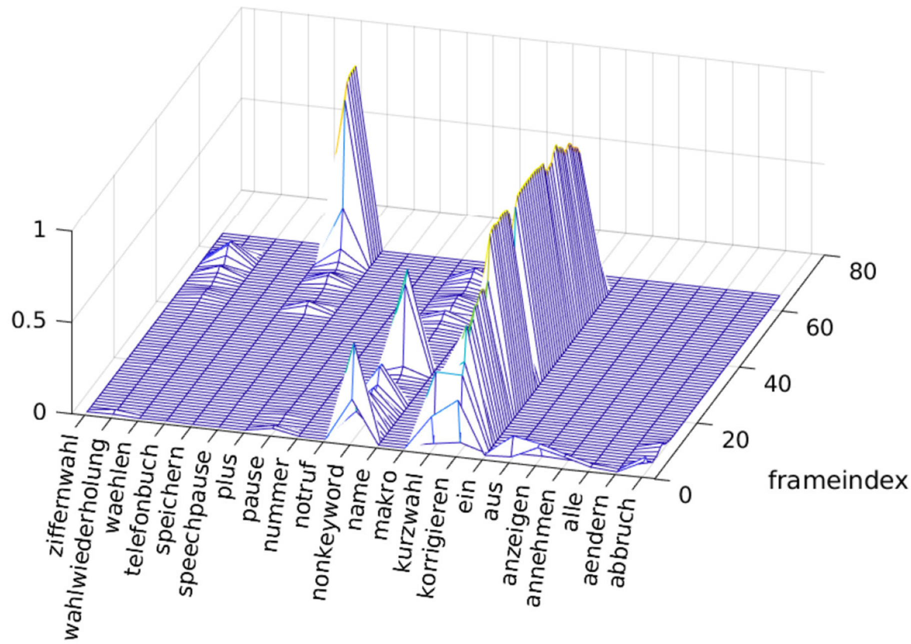


FIGURE 7 | Twenty-two output values of the neural network for the spoken word "korrigieren".

was performed with about 250 utterances of each command word and several thousand utterances of different non-keywords in Matlab.

Figure 7 shows the output values of the neural network for all speech frames after detecting the speech begin with the VAD approach described before. There are several nodes at the beginning with a noticeable value at their output. After analyzing the sequence of approximately the first 20 spectra, the output value at the node of the spoken word takes a fairly high value. Besides this node, only the non-keyword node takes some low values from time to time, but the correct recognition will be fairly easy for this example. In **Figure 7**, the criterion for detecting the end of the spoken word becomes visible. We looked at the probability at the node of the speech pause. If this probability takes on a high value in a number of consecutive frames, we take this as indication of the end of a spoken command. The final decision can be made, for example, by calculating the average probability at each node for the last 10 or 15 speech frames. Moreover, the application of a further "attention" layer is thinkable. So far, we did not deeply investigate the optimization of the recognition with respect to the feature extraction and the structure of the neural network. Clearly, our intention was the investigation whether the recognition can be realized with minimal energy consumption. We chose the small structure of the neural network with only three layers with respect to the realization on the MCU. The number of hidden units is set to 100 for the LSTM layer, the number of nodes to 200 for the first fully connected layer. Assuming a number of 22 output nodes for the recognition of 20 command words, the number of multiplications can be estimated to be at about 75,000 to realize the matrix multiplications of the three layers. Taking into account the further effort for realizing the activation functions and the calculation of the whole network at a frame rate of 78 Hz, the number of multiplications can be estimated to be at about 6.25 million per second. We derived a factor of about 8–10 from Lorenser (2016) to estimate the number of MCU cycles based on the number of multiplications. In Lorenser (2016), the number of processor cycles is listed for different signal processing tasks for which the number of multiplications is known. Thus, we can estimate that we should be able to realize the neural network with the MCU running at a clock frequency of about 60 MHz. Assuming the usage of a processor with a maximum clock frequency between 80 and 120 MHz, the computational performance would be sufficient to apply a neural network slightly more complex. There are software development frameworks for this type of MCUs (Microchip, 2020b) to create executable software modules for the integration of neural networks which were trained with tools like Tensorflow (Abadi et al., 2015).

Finally, we looked at the total energy consumption when using the microphone with the zero power listening mode in combination with an energy efficient MCU and applying the algorithmic approach as described before. The microphone would need 0.175 Wh of energy for operating in its listening mode at a voltage of 2 V for 1 year. Assuming 50 voice activations per day with the MCU running at 100 MHz for 10 s, we can estimate 0.81 Wh of energy to operate the recognition device for one year with a total current of 8 mA at a voltage of 2 V in its active phases. We have to consider that the client will also consume energy for other interfaces, e.g., for wireless communication, but it seems to be possible that such a device can be operated with minimal energy resources for a period of several months.

3.5. Communication *via* DECT-ULE

Communication between a client and a server system *via* WLAN can be problematic, for example, when different wireless networks are active in the same frequency band, e.g., at 2.4 MHz. This can delay and distort the communication process between clients and the recognition or the smart home server. We investigated the alternative use of the DECT-ULE standard for the realization of communication in the frequency band at about 1.9 GHz. DECT is used for cordless telephony in the range of up to 50 m at a fairly high transmission power of 250 mW. The ULE extension has been introduced with the goal of reducing energy consumption. Furthermore, it allows the transmission of data in addition to the speech signal. Therefore, it is suitable in the field of smart home control. With these features, it is well-suitable for communication between client and recognition server within a building. We developed a circuit based on the DHAN-M module (DSPGroup, 2020) into which the DECT-ULE protocol has been implemented. This module contains interfaces for audio input and output. Moreover, an MCU is part of our circuit as control unit of the client. The layout of the circuit has been designed to fit in an usual outlet socket. We could successfully prove the speech and data transmission between the client and a DECT base station. As the result of this short study, we consider the use of the DECT-ULE standard as an interesting alternative for communication between clients and the recognition server within a building.

4. RECOGNITION SERVER

Our goal is the realization of speech recognition in a local server system. Users can be sure that no speech data leave their homes and that nobody from outside can get access to the microphones of the clients as long as their local networks fulfill the appropriate security guidelines. In consequence, we have to aim at low cost server systems that customers are willing to pay for. This requirement is partly contrary to the need of a high computing power. We looked at three systems differing with regard to the complexity of the recognition task in the individual application. In most applications in the field of smart home control, only small vocabulary containing a few dozens or a few hundreds of words is needed. This already includes a larger variety at uttering a certain command phrase, so that the user is not forced to utter a command with only one fixed sequence of words. A fixed grammar is applied for these recognition tasks.

As the first system, we apply a PI computer which contains an ARM processor with four kernels running at a clock frequency of 1.5 GHz as recognition server. We developed a triphone-based GMM-HMM recognition scheme within the context of earlier research (Hirsch, 2008). This recognizer enables the recognition of up to a few hundred words on a PI device in real time. The triphone models have been trained on several hundreds of hours of German speech with HTK (Young et al., 2006). To increase the robustness of the recognition, data augmentation is applied to create further versions of the speech data including the acoustic effects of recording speech in hands-free mode in noisy and reverberant environments. The acoustic features consist of 12 MEL cepstral coefficients and the logarithmic energy plus the Delta and Delta-Delta coefficients. We can run this server module also on a client, so that we can set up a stand-alone device if necessary. But in case of using a PI-Zero as basis for the client, the recognition is limited to about a dozen words which could be sufficient for a simple command recognition. The dialogue software of the client is designed to freely choose different recognition servers for the individual recognition tasks within a single dialogue. Thus, we do not need communication between the client and a server at a different location in the building in dialogue situations in which only the recognition of a few words like "yes" and "no" is necessary.

We derived a second recognition system from the first one by substituting the GMM for a deep neural network (DNN). The setup of this system is shown in **Figure 8**. Spectral analysis is performed by means of a DFT at the rate of 100 frames per second. We do not determine a set of MEL cepstral coefficients as in case of the GMM-HMM recognizer. Instead, a set of 138 logarithmic compressed DFT coefficients is used as feature vector. The set of 138 coefficients consists of the DFT coefficients within the range from 250 Hz to 3 kHz. Furthermore, the mean of two DFT coefficients is calculated within the frequency range from 3 to 4 kHz and the mean of three values within the range from 4 to 7 kHz. DFT components above 7 kHz are not used. 11 consecutive vectors are taken as input for the neural net that consists of a CNN layer, a few LSTM layers and a few fully-connected layers. The number of output nodes corresponds to the number of tied triphone HMM states, which are 4,715 in our current implementation. By applying a softmax scaling at the output layer we try to estimate the emission probabilities of all tied triphone states as output of the neural network, so that we can calculate the probabilities of the HMMs as it is done with the GMM-HMM recognizer. The training of the neural net is done with Keras (Chollet, 2015) and Tensorflow (Abadi et al., 2015) by taking the same speech data as used for the training of the GMM-HMM recognizer. The mapping of feature vectors to tied triphone states was achieved by Viterbi alignment with the GMM-HMM recognizer.

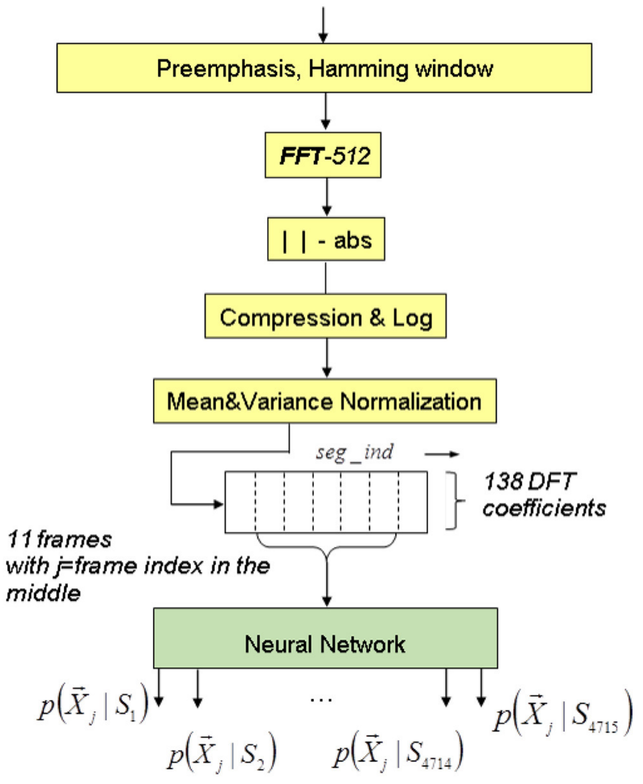


FIGURE 8 | Feature extraction and determination of emission probabilities with the DNN-HMM recognizer.

TABLE 1 | Word error rates (%).

Task	GMM-HMM	DNN-HMM
Single-word	2.8	0.2
Digit-sequence	4.5	3.1
Short-sentence	27.1	15.4

The computational performance of a PI is not sufficient to realize the calculation of the emission probabilities in real time. Therefore, we take the Jetson Nano device from NVidia (NVidia, 2020) as server component. This device contains an ARM processor plus a GPU containing 128 Cuda kernels. Thus, it is well-suited for performing the matrix multiplications in order to calculate the output of the neural net. The cost of this board is higher than that of a PI, but it is affordable for usage in the field of smart home applications. The computational performance of the board is high enough to run several instances of the neural network computation in parallel. Some word error rates are listed in **Table 1** to indicate and to compare the recognition performance of the GMM-HMM and the DNN-HMM recognizers regarding three different tasks.

These error rates were achieved by training the GMMs, the HMMs, and the neural network on about 2,000 h of clean German speech data from different databases (e.g., Burger and Schiel, 1998; Radeck-Arneth et al., 2015; Pratap et al., 2020). The first task, marked as *single-word* in the table, is the recognition of a single word from a set of 64 German command words. This self-recorded database contains about 22,000 utterances in total. The second task, marked as *digit-sequence*, contains the recognition of sequences of German digits. The database consists of about 19,000 utterances with 78,500 digits in total. Moreover, the third task, marked as *short-sentence* in the table, is the recognition of the word sequence in short sentences from a train query task. The database contains about 3,200 utterances with a total of approximately 34,000 spoken words (Schiel and Baumann, 2013). No language model or grammar was applied in the third task. We allowed any sequence of words from a total of 364 words. Here, the main focus was on the word accuracy regarding a larger vocabulary. Most errors were due to the misrecognition of a declination or a conjugation that have no influence on the determination of the required information. All test data were separate sets not used for training. The performance was higher in all tasks applying the DNN rather than the GMM for determining the emission probabilities. As mentioned before, the error rates are presented to show the recognition schemes' basic performance that can be achieved in simulation experiments without the usual techniques to improve the performance in different application scenarios. Applying a client in a particularly noisy environment, we improve the recognition performance by training GMMs, HMMs, and neural networks on multi-condition data. Besides clean data, we create noisy data for the requested application scenario with appropriate tools for data augmentation (Hirsch and Finster, 2005). We have a large set of noise signals and a large collection of room impulse responses available to simulate the transmission of speech in a noisy room environment (e.g., Jeub et al., 2009; Avosound, 2022). To measure room impulse responses directly in the application scenario, we have a measuring set-up (Hirsch et al., 2010). Furthermore, we include additional garbage HMMs for modeling background noises or speech artifacts like breathing or hesitations.

In case the recognition of a larger vocabulary without a fixed grammar is needed, we looked at the Kaldi recognizer (Povey et al., 2011) as a third recognition scheme. Kaldi has become a tool that is often used for research in the field of DNN based recognition. Besides its application in the field of research, there are also versions available for the recognition in real time (Alumäe, 2014). We implemented the interface into our dialogue module on the client side to communicate with Kaldi as third recognition scheme. We train the system on the same speech data that we use for training the other recognizers. Thus, we enable our clients to have access to the recognition of a large vocabulary of ten thousands of words when it might be needed within a dialogue. The use of the Kaldi recognition can be initiated with or without a fixed grammar. Especially for the recognition without a fixed grammar, an additional module is needed for the interpretation of the recognition result. From the field of natural language processing, different methods for parsing the text string coming from the recognizer are known. Nowadays, neural networks are applied for the realization of speech interpretation. We implemented a first module into our clients based on the well known word2vec approach (Mikolov et al., 2013). But, its integration and use in the client is still an ongoing project. We implemented the real-time version of Kaldi into a low cost PC to fulfill our requirement of limited total cost.

We could proof that Kaldi can be run on such a system with almost no noticeable delay in comparison to running on a much more powerful system. Overall, we set up a system where the clients are able to freely choose one out of several recognition servers at a certain dialogue state depending on the demand of the individual recognition task at this state.

5. CONCLUSIONS

We looked at technical solutions to set up a speech assistant system that reduces the concerns of a lot of users with respect to data privacy. Our focus is on the local realization of all needed client and server components inside a local network, so that no communication in a public network is necessary to realize speech control of hardware devices in a building. Usually, the storage of the recorded speech signals or speech features is disabled on client and server side. But the user can enable the storing in his local system during an initial operating phase. This data can be used to retrain and adapt the recognition system to the acoustic environment. The local realization implies the selection or the development of cost-efficient client and server components. We presented our hardware and software approach to realize the client in a very compact shape. The client includes the dialogue control and the communication with hardware devices besides speech input and output. We conducted a separate study to realize an extremely energy-efficient client that does not need any external power supply.

Three different recognition servers have been presented. These can be applied depending on the demand and the complexity of the individual recognition task in a dialogue state. The client can easily access all servers. We could prove our concept by setting up several demonstrator systems in the field of smart home control.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). "TensorFlow: large-scale machine learning on heterogeneous systems," in *OSDI'16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation* (Savannah, GA).

Alumäe, T. (2014). "Full-duplex speech-to-text system for Estonian," in *Baltic HLT 2014* (Kaunas).

Avosound (2022). *Digiffects Sound Library*. Available online at: https://www.avosound.com/ (accessed May 02, 2022).

Bäckström, T., Brüggemeier, B., and Fischer, J. (2020). "Privacy in speech interfaces," in *VDE Dialog - The Technology Magazine*. Offenbach. 11–14.

Bahmaninezhad, F., Zhang, C., and Hansen, J. H. (2018). "Convolutional neural network based speaker de-identification," in *Proceedings the Speaker and Language Recognition Workshop* (Les Sables d'Olonne), 255–260. doi: 10.21437/Odyssey.2018-36

Brasser, F., Frassetto, T., Riedhammer, K., Sadeghi, A.-R., Schneider, T., and Weinert, C. (2018). "Voiceguard: secure and private speech processing," in *Proceedings of Interspeech* (Hyderabad), 1303–1307. doi: 10.21437/Interspeech.2018-2032

Burger, S., and Schiel, F. (1998). RVG 1 - a database for regional variants of contemporary german. Available online at: https://www.phonetik.uni-muenchen.de/forschung/publikationen/Burger-98-RVG1.pdf (accessed May 02, 2022).

Chollet, F. (2015). Keras. GitHub repository. Available online at: https://github.com/fchollet/keras (accessed May 02, 2022).

Chung, H., Iorga, M., Voas, J., and Lee, S. (2017). Alexa, can I trust you? *Computer*. 50, 100–104. doi: 10.1109/MC.2017.3571053

DSPGroup (2020). *DHAN-M Module Dect Ule Platform Datasheet Version 4.0*. DSPGroup. Available online at: www.dspg.com (accessed May 02, 2022).

ETSI (2003). *Speech Processing, Transmission and Quality Aspects; Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithm*. European Telecommunications Standards Institute document ES 202 050 v1.1.3 (2003-11).

ETSI (2019). *Digital Enhanced Cordless Telecommunications (dect)*; *Ultra Low Energy (ule)*; *Machine to Machine Communications; Part 2: Home Automation Network*. European Telecommunications Standards Institute document TS 102 939-2 V1.3.1 (2019-01).

Hernández Acosta, L., and Reinhardt, D. (2020). "Smart speakers and privacy: users? perspectives," in *VDE Dialog - The Technology Magazine* (Offenbach), 8–10.

Hirsch, H.-G. (2008). "Automatic speech recognition in adverse acoustic conditions," in *Advances in Digital Speech Transmission*, eds R. Martin, U. Heute, and C. Antweiler (Hoboken, NJ: John Wiley & Sons, Inc), 461–496. doi: 10.1002/9780470727188.ch16

Hirsch, H.-G., and Ehrlicher, C. (1995). "Noise estimation techniques for robust speech recognition," in *Proceedings of ICASSP* (Detroit, MI), 153–156. doi: 10.1109/ICASSP.1995.479387

Hirsch, H.-G., and Finster, H. (2005). "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proceedings of the 9th European Conference on Speech Communication and Technology* (Lisbon). doi: 10.21437/Interspeech.2005-263

Hirsch, H.-G., Kitzig, A., and Linhard, K. (2010). *Simulation of the Hands-Free Speech Input to Speech Recognition Systems by Measuring Room Impulse Responses*. Bochum: ITG Fachtagung Sprachkommunikation.

Hirsch, H.-G., Micheel, A., and Gref, M. (2020). "Keyword detection for the activation of speech dialogue systems," in *Proceedings Elektronische Sprachsignalverarbeitung* (Magdeburg).

Hwang, K., Lee, M., and Sung, W. (2015). *Online Keyword Spotting With a Character-Level Recurrent Neural Network*. Available online at: https://arxiv.org/pdf/1512.08903.pdf (accessed May 02, 2022).

Jeub, M., Schaefer, M., and Vary, P. (2009). "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of International Conference on Digital Signal Processing (DSP)* (Santorini). doi: 10.1109/ICDSP.2009.5201259

Lau, J., Zimmerman, B., and Schaub, F. (2019). "Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers privacy attitudes of smart speaker users," in *Proceedings ACM Human-Computer Interaction 2* (Glasgow, UK), 1–31. doi: 10.1145/3274371

Lorenser, T. (2016). The DSP capabilities of arm cortex-m4 and cortex-m7 processors - DSP feature set and benchmarks. Available online at: https://community.arm.com/cfs-file/__key/communityserver-blogs-components-weblogfiles/00-00-00-21-42/7563.ARM-white-paper-_2D00_-

DSP-capabilities-of-Cortex_2D00_M4-and-Cortex_2D00_M7.pdf (accessed May 02, 2022).

Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., and Wagner, D. (2019). "Privacy attitudes of smart speaker users," in *Proceedings on Privacy Enhancing Technologies* (Stockholm), 250–271. doi: 10.2478/popets-2019-0068

Microchip (2020a). *Data Sheet of Microcontroller Unit* Atsame53j20a. Available online at: www.st.com (accessed May 02, 2022).

Microchip (2020b). *Software Development Tool STM32CUBEIDE With the Submodules STM32CUBEMX and X-Cube-Ai*. Available online at: www.st.com (accessed May 02, 2022).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Available online at: https://doi.org/10.48550/arXiv.1301.3781 (accessed May 02, 2022).

Mtbiaa, A., Petrovska-Delacrétaz, D., Boudy, J., and Hamida, A. B. (2021). "Privacy-preserving speaker verification system based on binary I-vectors," in *IET Biometrics*, eds R. Martin, U. Heute, and C. Antweiler (Hoboken, NJ: John Wiley & Sons), 233–245. doi: 10.1049/bme2.12013

Nautsch, A., Isadskiy, S., Kolberg, J., Gomez-Barrero, M., and Busch, C. (2018). "Homomorphic encryption for speaker recognition: protection of biometric templates and vendor model parameters," in *Proceedings the Speaker and Language Recognition Workshop* (Les Sables d'Olonne), 16–23. doi: 10.21437/Odyssey.2018-3

Nautsch, A., Jimenez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., et al. (2019). "Preserving privacy in speaker and speech characterization," in *Computer Speech and Language: Special Issue on Speaker and Language Characterisation*, (Amsterdam: Elsevier), 441–480. doi: 10.1016/j.csl.2019.06.001

NVidia (2020). *Nvidia Jetson Nano*. Available online at: developer.nvidia.com (accessed May 02, 2022).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society* (Big Island; Hawaii).

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). "MLS: a large-scale multilingual dataset for speech research," in *Proceedings of Interspeech*, 2757–2761. doi: 10.21437/Interspeech.2020-2826

Radeck-Arneth, S., Milde, B., Lange, A., Gouvea, E., Radomski, S., Muehlhaeuser, M., et al. (2015). "Open-source German distant speech recognition: corpus and acoustic model," in *Proceedings of the 18th International Conference TSD2015* (Shanghai). doi: 10.1007/978-3-319-24033-6_54

Rose, R., and Paul, D. (1990). "A hidden Markov model based keyword recognition system," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (Albuquerque, NM), 129–132. doi: 10.1109/ICASSP.1990.115555

Sainath, T., and Parada, C. (2015). "Convolutional neural networks for small-footprint keyword spotting," in *Proceedings of Interspeech* (Dresden), 1478–1482. doi: 10.21437/Interspeech.2015-352

Schiel, F., and Baumann, A. (2013). *Phondat 2. Bavarian Archive for Speech Signals*. Available online at: https://www.bas.uni-muenchen.de/forschung/Bas/BasPD2eng.html (accessed May 02, 2022).

Seeed (2021). *Respeaker Microphones*. Available online at: wiki.seeedstudio.com (accessed May 02, 2022).

Seiderer, A., Ritschel, H., and André, E. (2020). "Development of a privacy-by-design speech assistant providing nutrient information for German seniors," in *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good* (Antwerp). doi: 10.1145/3411170.3411227

Vesper (2021). *Data sheet of Microphone VM3011*. Available online at: vespermems.com (accessed May 02, 2022).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, X. G., et al. (2006). *The HTK Book. Version 3.4*. Available online at: https://www.academia.edu/18598764/The_HTK_book_for_HTK_version_3_4 (Accessed May 02, 2022).

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

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

Check for updates

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

Simon Schreibelmayr and Martina Mara*

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

OPEN ACCESS

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

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

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

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

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

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

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

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

INTRODUCTION

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

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

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

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

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

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

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

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

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

Human-Like Voice as Anthropomorphic Cue

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

Looking at the crucial role of human voice to exchange information and to interpret others in our social life, it is not surprising that voice emitted by a computer is considered a particularly strong anthropomorphic cue (Nass and Brave, 2005; Qiu and Benbasat, 2009; Eyssel et al., 2012; Whang and Im, 2021), along with visual cues, such as human-like embodiment or non-verbal behavior of a machine (cf. Mara and Appel, 2015a,b). Anthropomorphism describes the widespread tendency to attribute human characteristics, motivations, intentions, or emotions to non-human entities, or in short, to sense something human where there is actually nothing human (Epley et al., 2007). This can happen with things that do not use natural speech or resemble human appearance at all, such as cuddly toys or even plants. According to Theory of Anthropomorphism of Epley et al. (2007), however, readily observable human-like features increase an object's likelihood of being anthropomorphized because they facilitate the accessibility of anthropocentric knowledge structures and thus increase the chance that such knowledge will be applied to the non-human target. This is in line with Nass and colleagues' Computers Are Social Actors paradigm (CASA, Nass et al., 1994; Reeves and Nass, 1996; Nass and Brave, 2005), which posits that individuals mindlessly apply social heuristics from interpersonal interactions to their interactions with computers. According to the authors, perceiving a computer as social actor is particularly likely when it takes on a role that was typically fulfilled by a human (e.g., tutor, salesperson, and therapist), when it is interactive, or when it uses natural speech (Nass et al., 1994; Nass and Brave, 2005).

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

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

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

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

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

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

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

The literature, however, also features a number of studies that report non-favorable user reactions to high levels of human-likeness in machines. For example, in several experiments from the field of human-robot interaction it was found that people prefer more machine-like robot appearances over more human-like ones (Bartneck et al., 2007; Broadbent et al., 2011; Mara and Appel, 2015a,b; Mathur and Reichling, 2016; Vlachos et al., 2016; Jia et al., 2021). Works that suggest negative effects of anthropomorphic designs typically refer to the Uncanny Valley hypothesis (Mori, 1970; Mori et al., 2012), which proposes a non-linear relationship between the human-likeness of an artificial character and the valence it evokes in its observers. According to Mori's hypothesis, in a generally low range of human-likeness, pleasantness grows with increasing realism. At a point of rather high human-likeness, however, the effect reverses and the artificial entity is perceived as eerie or threatening. Only when the entity's degree of realism reaches near-perfection or perfection will pleasantness go up again, since no distinction can be made any longer between artificial and human (Mori et al., 2012; Mara et al., 2022). Various perceptual and cognitive mechanisms have been suggested to underlie uncanny experiences (cf. Diel and MacDorman, 2021). These include categorical uncertainty or prediction difficulties if features of a given entity seem to belong to different conceptual categories (e.g., a mechanoid robot head with a human-like voice, Mitchell et al., 2011; Meah and Moore, 2014).

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

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

Acceptance and Application Context

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

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

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

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

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

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

Acceptance and User Personality

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

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

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

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

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

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

MATERIALS AND METHODS

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

Voice Stimuli

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

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

Human

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

TABLE 1 | Description of the five experimental robot voices.

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

Synthetic I

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

Synthetic II

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

Metallic

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

Comic

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

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

Sample Size Justification and Participants

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

Procedure

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

Measures

Dependent Variables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

RESULTS

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

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

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

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

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

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

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

Human-Likeness and the Uncanny Valley

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

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

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

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

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

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

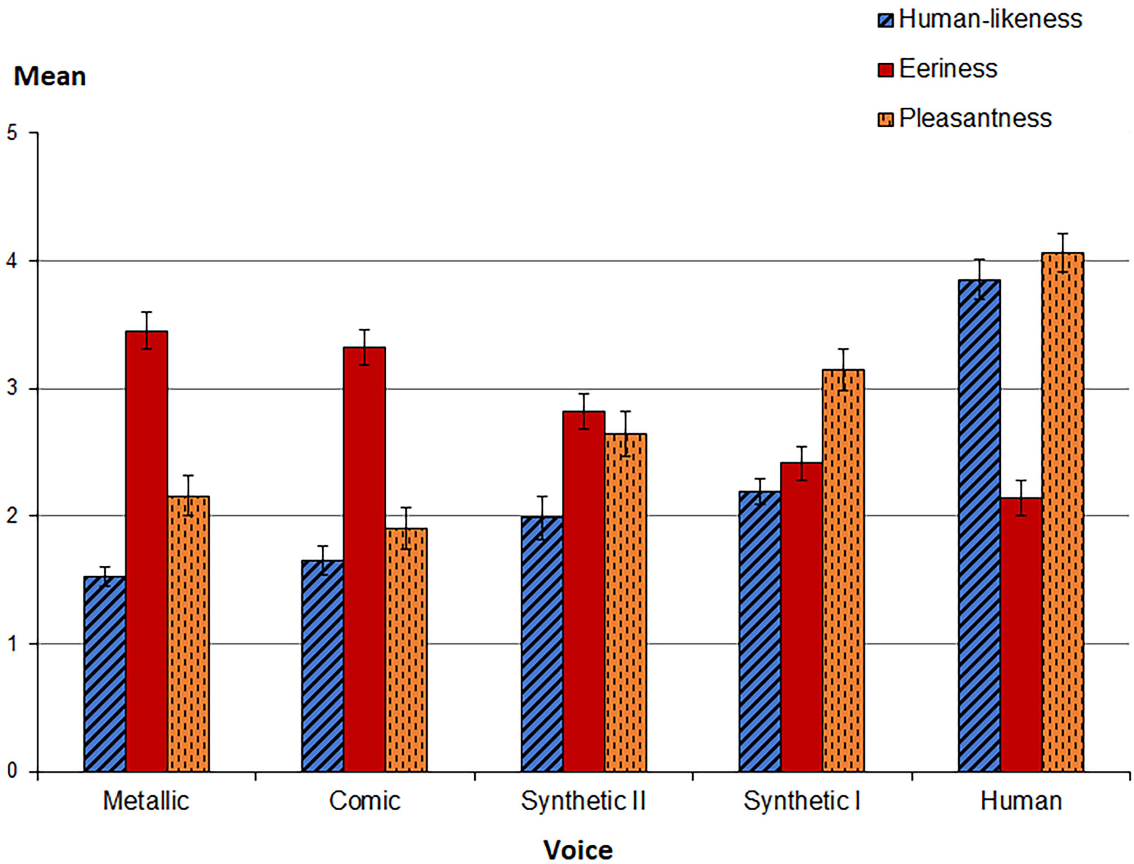


FIGURE 1 | The bar chart shows the mean values of the variables *Human-likeness*, *Eeriness,* and *Pleasantness* depending on the heard voice. The five voices are arranged from left to right in an increasing degree of Human-likeness.

Application Context and Acceptance of Vocal Human-Likeness

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

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

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

User Personality and Acceptance of Vocal Human-Likeness

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

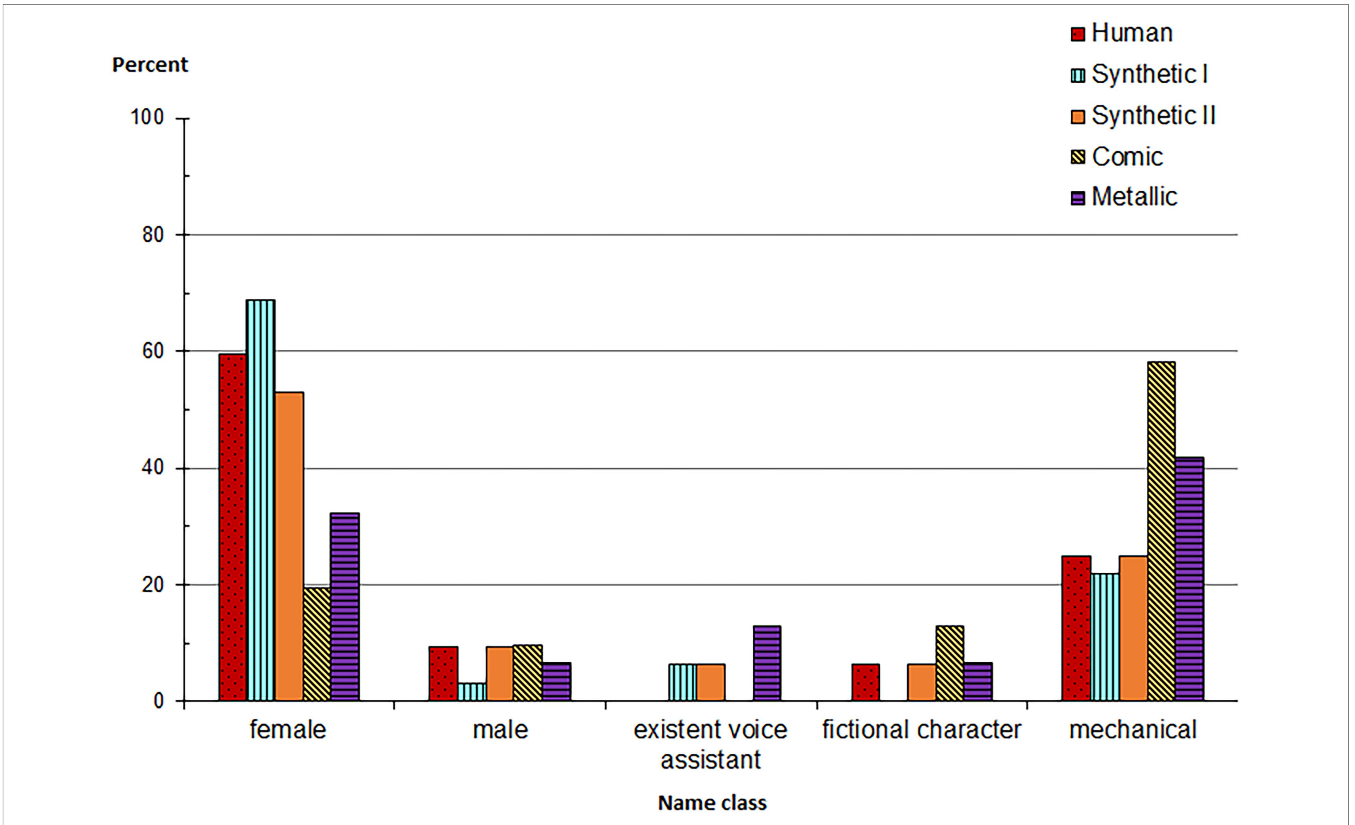


FIGURE 2 | The bar chart shows the absolute values as percentage of invented names depending on the heard voice. The names were assigned to one of the five name classes.

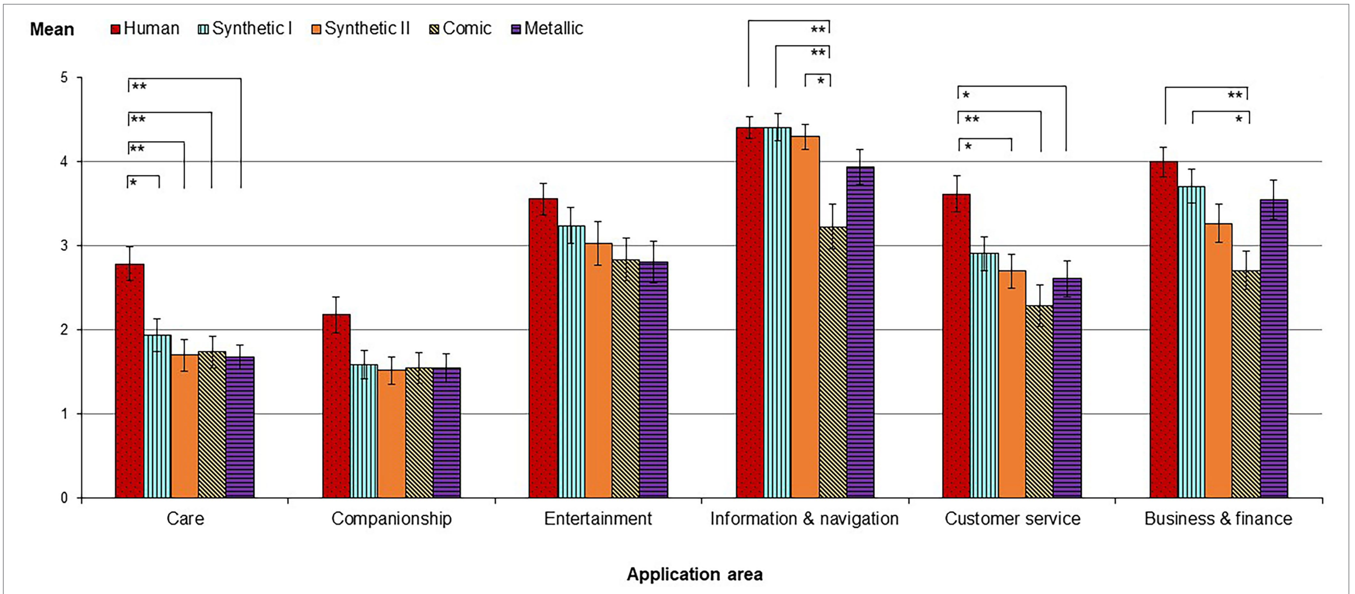


FIGURE 3 | The bar chart shows the mean values of *acceptance* of the five different voices depending on the respective context. A Kruskal–Wallis test was used for pairwise group comparisons (**$p < 0.01$; *$p < 0.05$).

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

Using the PROCESS macro (version 3.3) for SPSS by Andrew Hayes (Hayes and Cai, 2007; Baltes-Götz, 2017; Process, 2019), we conducted moderation analyses to examine whether *tolerance* *of ambiguity, neuroticism* and, for exploratory purposes, the other personality variables of the *Big Five* had a significant influence on the associations between *human-likeness* and *eeriness*. No such interactions on a significance level of $\alpha = 0.05$ were revealed (find more information on the moderation models 1–6 in **Table 8; see Appendix C**). Thus, our hypotheses *H4a* and *H4b* did not find support within this study.

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

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

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

DISCUSSION

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

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

General Discussion

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

Our non-directional Hypotheses 2a–b, stating that there would be significant group differences in pleasantness and eeriness ratings between the voices, found support in such a way that more human-like voices were experienced as significantly more pleasant and less eerie than more mechanical sounding voices. This is in agreement with prior empirical studies that also observed positive effects of anthropomorphic design features (Romportl, 2014; Baird et al., 2018; Kühne et al., 2020; Roesler et al., 2021). At the same time, it seems to contradict the Uncanny Valley hypothesis (Mori, 1970) according to which we would have expected either the quite realistic yet not perfect voices Synthetic I or II receiving the highest eeriness ratings or alternatively—assuming categorical conflicts as an important mechanism behind uncanny experiences—the real human voice (given that participants were told they were listening to a robot). What needs to be noted here is that according to Mori's popular Uncanny Valley graph, which illustrates the assumed curvilinear relationship between human-likeness of a figure and the valence of observer evaluations, a positively valenced peak (most likable, pleasant) should occur at about 70% and a negatively valenced "valley" (most eerie, uncanny) at about 85% along the human-likeness continuum. However, with a mean value of 3.8 (on a range of 1–5) in reported human-likeness perceptions, even the real human voice in our experiment was relatively far from the right end point of the human-likeness continuum, but closer to the predicted positive peak. From this perspective, by following Mori's postulations, it is not surprising that linear rather than curvilinear relationships between perceived human-likeness and eeriness (or pleasantness) were identified from our data, since the Uncanny Valley hypothesis itself predicts a rather linear increase of positive valence in a low to medium-high range of human similarity, that is, left of the positive peak.

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

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

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

Limitations and Outlook

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

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

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

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

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

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

CONCLUSION

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

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

We would like to thank the professional singer Raffaela Gmeiner, whose voice we used for the recordings of the "Human" voice condition. We are also grateful to Benedikt Leichtmann and Christine Busch for their valuable comments on an earlier version of the manuscript and to Franz Berger, who helped us creating the synthetic voice samples. We would also like to thank Wolfgang Schreibelmayr, who supported us with the acquisition of participants. The publication of this work was supported by the Open Access Publishing Fund of the Johannes Kepler University Linz.

SUPPLEMENTARY MATERIAL

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., and Robert, L. P. (2021). "A meta-analysis of human personality and robot acceptance in human-robot interaction." in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sutton, S. J., Foulkes, P., Kirk, D., and Lawson, S. (2019). "Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction." in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership