Speaking and Listening: Mismatched Human-like Conversation Qualities Undermine Social Perception and Trust in AI-based Voice Assistants

Research-in-Progress

Peng Hu **Kun Wang**

Jingwen Liu

Abstract

As a relatively new IT artifact, voice assistants are growing in abilities to interact with users in a more natural way with the fast advancement of Artificial Intelligence technology, speaking out human-like voice and understanding voice information. Although prior studies have revealed that voice human-likeness of software agents could improve users' evaluation and experience, it remains unknown that how the inseparable voice (speaking) and understanding (listening) conversation qualities of voice assistants conjointly impact users' social perception and trust. Based on cognitive consistency theory, we propose that it exists a congruency effect between these two communication qualities, users may perceive lower social presence and more unlikely to trust when they experienced mismatched human-like conversation qualities. An online survey will be conducted to collect data, and polynomial modeling and response surface methodology will be used to test our congruency hypotheses. Potential implications for theory and practice are also discussed.

Keywords: voice assistants, congruency effect, cognitive consistency, social presence

Introduction

In recent years, the rapid development and advancement of Artificial Intelligence (AI) technology enables many intelligent products to step into people's daily life from laboratory. As a representative, AI-based Voice Assistants (VA) have gained good market performance since Amazon's introduction of the Echo smart speaker and the Alexa voice assistant in 2014. Other than Amazon, many international IT companies are competing fiercely in price and function in the VA marketplaces, such as Google Assistant, Apple Siri, Microsoft Cortana, and Samsung Bixby. In China, BAT (i.e., Baidu, Alibaba, and Tencent) also joined in this market in succession. In terms of VA's growing popularity, the capability of human-like conversation supported by natural language processing technology could be considered as a paramount factor. Specifically, users can perform simple tasks through voice command over VA, such as making a call, sending a message, or searching online for specific information (Saad et al. 2017). Meanwhile, VA can give users feedback by natural language too, that is to say, VA possesses the human-like conversation qualities of speaking and listening in some sense. Naturally, VA has the potential of building a social connection with its users via some social tasks, for

* Corresponding author

instance, making jokes or play games with users. In such a case, people may feel a sense of social presence, the feeling of "being with another" (Biocca et al. 2003). This socialness experience may further evolve into relations similar to interpersonal relationships, which is hardly produced between users and other traditional IT products. In summary, aside from the benefit of ease of use, natural language interaction also provides the potential of establishing a social linkage between VA and its users, pushing people's daily life to a way towards intelligence.

Given the trend that many information systems are increasingly embedded with more human-like features due to the advancement of information and communication technology, plenty of studies has investigated the impact of these human-like factors on users' attitudes and behaviors. In terms of information systems embedded with speech interface, prior studies have examined the influence of voice human-likeness, the extent to which a VA's voice (i.e., speaking quality) similar to that of a humankind, of varied virtual agents or physical robots on individual's evaluations and responses (Cowan et al. 2015; Edwards et al. 2018; Lee 2010; Qiu and Benbasat 2009), and observed the positive effects of these design choice in many cases, ascribing to the increased social attractiveness. However, the output voice of these information systems is only one-half of the elements underpinning voice interaction. As with a human who wants to communicate with others in spoken language, s/he must not only able to voice out what s/he wants to convey but also capable of understanding what others said. Likewise, for an information system which can converse with its users in a way that similar to human-human interaction, understanding human-likeness, conceptualized as the extent to which a VA's understanding of spoken language (i.e., listening quality) similar to that of humankind, also is an indispensable requirement on it besides voice human-likeness. In this sense, many previous studied information systems may be inappropriate to be considered as voice interaction systems because most of them only occupy the capability of human-like voice with the absence of spoken language understanding. Thanks to the improvement of speech recognition and natural language processing, many IT products, such as VA in this study, have to some extent satisfied the two preconditions supporting voice interaction above mentioned, in spite of a substantive disparity apart from mankind.

However, it remains unknown how these two human-like conversation attributes (i.e., human-like voice and human-like understanding) conjointly affect users' beliefs and evaluations on it. To this end, building on cognitive consistency theory and social presence theory, this study examines the congruency effect of voice human-likeness and understanding human-likeness of VA on users social presence and the resulting trust belief. Prior research on trust in technology suggested that trust acts as a pivotal indicator of IT acceptance (Pavlou 2003; Benbasat et al. 2005), post-adoption usage (Paravastu et al. 2014), delegation (Stout et al. 2014). Especially, some VA already supports voice shopping (e.g., Amazon Echo, Google Home), shopping online through voice interaction with VA (Maarek 2018), as such trust in VA is a crucial step to achieve profit maximum for corporates involved in VA. As a result, this study treats trust in VA as the focal outcome. To sum up, this study aims to address the following research questions:

RQ#1. Does the mismatch between human-likeness of voice and understanding impair users' social presence on VA?

RQ#2. How users' social presence on VA relates to the magnitude of matched human-like voice and human-like understanding?

Theory and Hypotheses

Social Presence

Initially, social presence was used to represents the subjective experiences of other human beings in a computer-mediated environment (Yoo and Alavi 2001). Recently, this concept has been used to appraising social perceptions of IT artifacts itself such as websites (Lu et al. 2016), virtual agents (Ben Mimoun et al. 2017), and social robots (Edwards et al. 2019). Prior studies revealed that social cues such as natural language, human-like voice, interactivity, induce social responses such as self-disclosure, politeness, and trust belief from users (Nass and Moon 2000). The underpinning

theoretical framework that explains these social responses is ‘‘Computers as Social Actors’’ paradigm, which proposes that individuals treat and respond to computers as they do to humans, despite being aware of the object they are interacting with are machines and not humans. Related works that accepted this paradigm have designed user interfaces with human-like qualities such as human-like appearance and representation that reacts to users by verbal or non-verbal interaction, bringing about a feeling of social presence for users. Social presence is of great importance in designing social agents, such as VA in this study, because one of the ultimate goals of these social actors is to provide users with strong feelings of socialness (Breazeal 2003; Lee et al. 2006).

Interacting with VA is more natural than traditional touch-based IT products, because of its human-like voice and human-like understanding of voice information users convey. Human-like voice and human-like understanding constitute the two crucial social cues eliciting social presence perception of users. However, based on perceptual mismatch theory, this study argues that the effects of these two human-like qualities on social presence are neither independent of nor simply interacted with each other. Instead, it is the match/congruency between the two human-like features that really matter, for which the rationales are discussed in the following.

Cognitive Consistency Theory

Since many virtual agents involve speech interface (e.g., smartphone, online shopping assistant), literature has examined the voice properties of these agents as potential factors to drive user’s attitude change. For example, a study showed that the perceived gender of an agent’s voice already suffices to incur gender stereotypes, thus fostering or impairing persuasion success (Powers and Kiesler, 2006). Other scholars argued that only adding a voice to a silent software agent could be sufficient to enhance its perceived trustworthiness, even with a simple text-to-speech voice (Qiu and Bensabat 2005). In other words, the speaking ability of an agent may be enough to induce some social responses.

Another literature stream examined the combined effects of multiple human-like cues on individuals’ social perceptions. For instance, Gong and Nass (2007) uncovered the consistency effect between voice and visual cues of computer agents, human voice was evaluated as more trustworthy when paired with a human-like appearance, while mechanical voice may be more preferable when combined with a machine-like face. Recently, MacDorman and Chattopadhyay (2016) also observed this consistency effect, with the finding that digital agents with mismatched visual and audio attributes tend to trigger eerie feelings or even disgust emotions. In terms of potential explanations for this consistency effect, cognitive psychology argue that inconsistent cues manifested by a single object could elicit confused categorical perception and categorization difficulty, individuals’ perceived mismatching between incongruent cues may furtherly engender cognitive dissonance, which is an undesirable psychological state (Carr et al. 2017; Kätsyri et al. 2015; MacDorman and Chattopadhyay 2016). On the other hand, scholars in evolutionary psychology postulate that stimulus consistency serves as an indicator of physical fitness and reproductive health (Kätsyri et al. 2015).

In summary, no one unified framework to explain the consistency effect till now, but we argue that, in essence, viewpoints of evolutionary psychologists could be regarded as automated cognition that evolves from conscious cognition. Nonetheless, the convergent conclusion remains that human has a strong consistency preference, and the entities that exhibit different but matched attributes (e.g., human-like voice and understanding in this study) could obtain higher acceptance and evaluations. Accordingly, the perceptual mismatch between voice human-likeness and understanding human-likeness is expected to result in categorization difficulty and uncertainty (i.e., difficult to categorize human-like VA as human or nonhuman), and then disrupt the social experience with VA. Though, the positive effect of natural language interaction with VA on social presence could be elicited in the case of matched qualities above, along with the same mechanisms as prior works examing other human-like social cues. We propose that:

Hypothesis 1. When users experience a mismatch between human-likeness of voice and understanding of their VA, their perceived social presence is lower, than when matched human-likeness of voice and understanding is perceived.

Hypothesis 2. Social presence is higher when voice and understanding of their VA are matched at a high level of human-likeness than when these two qualities of their VA are matched at a low level of human-likeness.

Social Presence and Trust

The most salient activity between users and VA is Question-Answer interaction (i.e., voice search). Unlike web-based search in which numerous search results are presented in separated pages for searchers to browse and screen, the outputs are pre-screened by VA to reduce cognitive load for users and can provide information that tailored to users. Apparently, the screening mechanism of VA entails uncertainty and risk, which possibly caused by the inability to offer correct information or by the malevolence of providing recommendations that beneficial for their producers but violating interests of their users. Hence, trust is a central construct in User-VA interaction.

Generally, trust denotes "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor" (Mayer et al. 1995). Most scholars agree that trust in virtual agents is similar to trust in human, including the three indigents of competence, benevolence, and integrity (Komiak and Benbasat 2006). In an online shopping context, social presence caused by animated virtual agents could increase consumer trust in the online store (Qiu and Benbasat 2009). Studies have also uncovered that enhancing websites' social presence via the text-to-speech voice of virtual agents may increase cognitive and emotional trust (Qiu and Benbasat 2005). In essence, VA is virtual agents that embedded in some existed IT products such as smartphone and speaker, thus it is expected that the observed social presence effect on trust in prior studies may be similar to that of VA in this study. Moreover, based on the hypotheses developments above and the observation that lots of studies in HCI have consistently found that social presence acts as a mediator between users perception on human-like qualities of artificial agents and responses to these artifacts, we furtherly postulate that social presence mediates the impact of human-like voice and understanding on users' trust in VA in our study. thereby we propose that:

Hypothesis 3. social presence is positively associated with users' trust in VA.

Hypothesis 4. social presence mediates the congruency effects of perceived human-like voice and understanding on users' trust in VA.

Figure 1 displays the research model of this study, in which some important control variables are included. Specifically, individuals with a high immersive tendency, the degree to which one tends to be immersed in an interaction activity, are more likely to perceive and experience a social agent (e.g., VA in this study) as animated with low awareness of surroundings (Kim et al. 2012). Thus immersive tendency is included are a control for social presence. Propensity to trust, a disposition that reflects an individual's general tendency of trusting others was suggested as affecting users' trust beliefs toward technology (Wang and Benbasat 2008). Need to belong is also found to impact our social interactions (Greenwood and Long 2011). Individuals with a high need to belong are more likely to trust in VA due to their innate desire to form social connections. Additionally, Trust in VA itself may also be confounded with trust in its producer such as Amazon, Google and so on. Hence, Propensity to trust, Need to belong, and Brand attitudes are embraced as control variables for the outcome trust in VA. We also control potential impacts of demographic variables (i.e., gender, age, education, and income) and prior experience with VA.

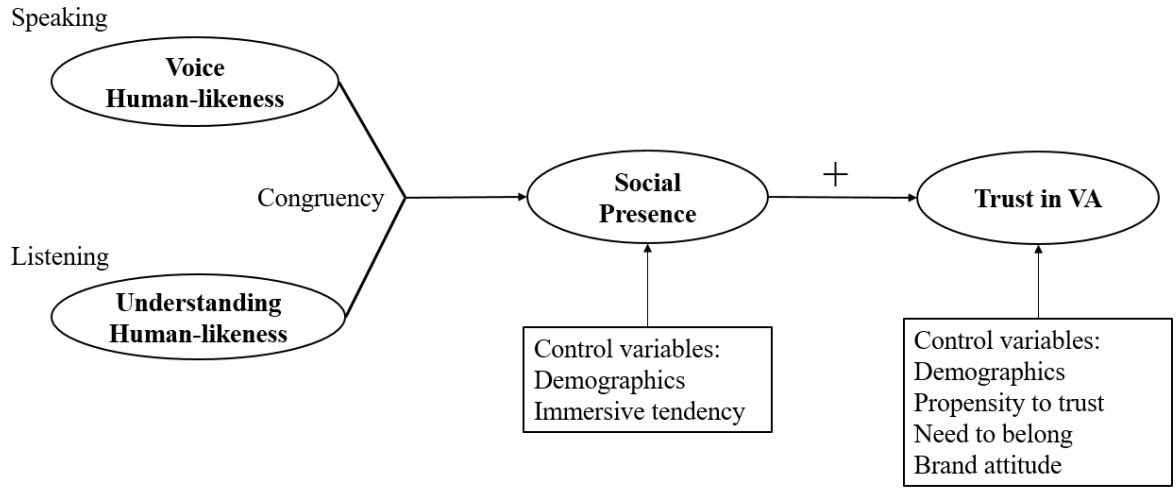

Figure 1. Research Model

Research Method

Sample and Data Collection

The IT artifact in the current study is VA, which are mainly embedded in smartphones(e.g., Apple Siri) and smart speakers(e.g., Amazon Alexa), and our research model focus on post-adoption interaction on VA. Hence, Owners of smartphones or smart speakers who have experienced voice interaction with their VA are the sample population of this study. An online survey will be conducted to test the proposed model. We choose the survey method because it can obtain personal and social beliefs, and attitudes, and it also enjoys the merit of improving the generalizability of research findings. In particular, participants will be recruited on a specialized crowdsourcing platform of survey research in China, with which diverse demographic individuals can be captured. The online survey is designed as follows. First, questions on demographic information are presented, including gender, age, income, education level, etc. Second, two questions on VA usage are provided to obtain the specific VA in use and how long VA have been used by participants. Finally, items of variables in the proposed model are measured. We plan to measure social presence and trust in VA using extant scales which hold a good psychometrika performance in prior studies. Based on the related work on speech interfaces in HCI, we will develop scale instruments for human-likeness of voice and understanding by following the guidelines provided in Moore and Benbasat (1991). Moreover, we will calculate the requirement on sample size for the research model to ensure adequate statistical power for hypothesis testing before data collection.

Data Analysis Approach

Drawing on the congruency research in the organizational behavior area, polynomial modeling and response surface methodology were selected as proper approaches to test the hypotheses concerning congruency/match in this study. Due to their considerable potential for IS researchers to tackle this specific type of hypothesis, we provide the basic ideas of the two approaches. In addition, the mediation effect testing regarding congruency effects is also different from testing a conventional mediation model, the block variable approach will be adopted to address this issue (Edwards and Cable 2009).

Unlike traditional analytical approaches that entail problems associated with linear models, difference scores, and direct measures (Edwards and Parry 1993), polynomial modeling includes a hierarchical analysis of polynomial equations that allows researchers to examine complex relationships between component measures and outcomes (Venkatesh and Goyal 2010). As with the testing procedure, firstly, component scores (i.e., voice human-likeness, understanding human-likeness in this proposal) are entered into a linear equation to test their relationship with outcome variables. Secondly, higher-

order terms along with product terms are included in the equation to test curvilinear relations among the variables. Quadratic and cubic terms can also be added to the equation to identify the existence of cubic effects (Edwards and Parry 1993).

Coefficients in polynomial models are often difficult to interpret. Response surface methodology is deployed as an interpretive technique to show how to use these coefficients to justify the hypotheses we proposed (Edwards and Parry 1993). Response surface methodology focuses on three primary features of the surfaces generated from polynomial coefficients: stationary point, principal axes, and slopes along lines of interest. The stationary point refers to the point at which the slope of a surface is zero in all directions. Principal axes run perpendicular to each other and interact at the stationary point. For a convex surface, the upward curvature is greatest along the first principal axis and least along the second principal axis. For a concave surface, the downward curvature is least along the first principal axis and greatest along the second principal axis. The other lines of interest are the congruence line, where both the component measures are equal (X = Y), and the incongruence line, where both the component measures are equal but opposite sign (X = -Y) (Edwards and Parry 1993).

Potential Implications

For theory, although prior works have uncovered that social agents with a more human-like voice could improve users' social perceptions and evaluations, little is known about the impact of spoken language understanding of these agents on users' social responses. More importantly, drawing on related findings in prior congruency research, we propose the existence of a consistency effect between the two human-like conversation qualities on users' social perception and trust belief. This study has the potential to expands the cognitive consistency theory if congruency hypotheses can be testified. For practice, AI-based Voice Assistants are a relatively new type of IT artifacts with growing strong ability to naturally converse with users, this study has the potential to enhance our understanding on how the conversation qualities affects users' social beliefs towards these new AI-enabled IT and provide some design advice (e.g., balance the speaking and listening qualities of VA) for developers or managers in VA industry.

Acknowledgements

This work was supported by grants from the Natural Science Foundation of China (NSFC) (71810107003).

References

Benbasat, I., and Wang, W. 2005. "Trust In and Adoption of Online Recommendation Agents," *Journal of the Association for Information Systems* (6:3), pp. 72–101.

Biocca, F., Harms, C., and Burgood, J. K. 2003. "Toward a More Robust Theory and Measure of Social Presence," *Presence: Teleoperators & virtual environments* (12:5), pp. 456–481.

Breazeal, C. 2003. "Toward Sociable Robots," *Robotics and Autonomous Systems* (42:3), pp. 167–175.

Carr, E. W., Hofree, G., Sheldon, K., Saygin, A. P., and Winkielman, P. 2017. "Is That a Human? Categorization (Dis)Fluency Drives Evaluations of Agents Ambiguous on Human-Likeness," *Journal of Experimental Psychology: Human Perception and Performance* (43:4), pp. 651–666.

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," *International Journal of Human Computer Studies* (83), pp. 27–42.

Dennis, A., Stout, N., and Wells, T. 2017. "The Buck Stops There: The Impact of Perceived Accountability and Control on the Intention to Delegate to Software Agents," *AIS Transactions on Human-Computer Interaction* (6:1), pp. 1–15.

Edwards, J. R., and Parry, M. E. 1993. "On the Use of Polynomial Regression Equations as an Alternative to Difference Scores in Organizational Research," *Academy of Management Journal* (36:6), pp. 1577-1613.

Edwards, A., Edwards, C., Westerman, D., and Spence, P. R. 2019. "Initial Expectations, Interactions, and beyond with Social Robots," *Computers in Human Behavior* (90), pp. 308–314.

Edwards, A., Stoll, B., Massey, N., Edwards, C., and Lin, X. 2018. "Evaluations of an Artificial Intelligence Instructor's Voice: Social Identity Theory in Human-Robot Interactions," *Computers in Human Behavior* (90), pp. 357–362.

Edwards, J. R., and Cable, D. M. 2009. "The Value of Value Congruence," *Journal of Applied Psychology* (94:3), pp. 654–677.

Gong, L., and Nass, C. 2007. "When a Talking-Face Computer Agent Is Half-Human and Half-Humanoid: Human Identity and Consistency Preference," *Human Communication Research* (33:2), pp. 163–193.

Greenwood, D. N., and Long, C. R. 2011. "Attachment, Belongingness Needs, and Relationship Status Predict Imagined Intimacy with Media Figures," *Communication Research* (38:2), pp. 278–297.

Kim, K. J., Park, E., Sundar, S. S., & del Pobil, A. P. 2012. "The effects of immersive tendency and need to belong on human–robot interaction," In *Proceedings of HRI'12*. pp. 207–208.

Kätsyri, J., Förger, K., Mäkäräinen, M., and Takala, T. 2015. "A Review of Empirical Evidence on Different Uncanny Valley Hypotheses: Support for Perceptual Mismatch as One Road to the Valley of Eeriness," *Frontiers in Psychology* (6), pp. 1–16.

Komiak, and Benbasat. 2006. "The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents," *MIS Quarterly* (30:4), pp. 941–960.

Lee, E. J. 2010. "The More Humanlike, the Better? How Speech Type and Users' Cognitive Style Affect Social Responses to Computers," *Computers in Human Behavior* (26:4), pp. 665–672.

Lee, K. M., Peng, W., Jin, S. A., and Yan, C. 2006. "Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human-Robot Interaction," *Journal of Communication* (56:4), pp. 754–772.

Lu, B., Fan, W., and Zhou, M. 2016. "Social Presence, Trust, and Social Commerce Purchase Intention: An Empirical Research," *Computers in Human Behavior* (56), pp. 225–237.

Maarek, Yoelle. 2018. "Alexa and Her Shopping Journey. " *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1–1.

MacDorman, K. F., and Chattopadhyay, D. 2016. "Reducing Consistency in Human Realism Increases the Uncanny Valley Effect; Increasing Category Uncertainty Does Not," *Cognition* (146), pp. 190–205.

Mayer, R. C., Schoorman, F. D., and Davis, J. H. 1995. "An Integrative Model of Organizational Trust: Past, Present, and Future," *Academy of Management Review* (20:3), pp. 709–734.

Ben Mimoun, M. S., Poncin, I., and Garnier, M. 2017. "Animated Conversational Agents and E-Consumer Productivity: The Roles of Agents and Individual Characteristics," *Information and Management* (54:5), pp. 545–559.

Moore, G. C., and Benbasat, I. 1991. "Development of an Instrument to Measure the Perceptions of Adopting an IT Innovation," *Information Systems Research* (2:3), pp. 192–222.

Nass, C., and Moon, Y. 2000. "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues* (1), pp. 81–104.

Paravastu, N., Gefen, D., and Creason, S. 2014. "Understanding Trust in IT Artifacts - An Evaluation of the Impact of Trustworthiness and Trust on Satisfaction with Antiviral Software," *ACM SIGMIS Database* (45:4), pp. 30–50.

Pavlou, P. A. 2018. "Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model," *International Journal of Electronic Commerce* (7:3), pp. 101–134.

Qiu, L. and Benbasat, I. 2005. "Online Consumer Trust and Live Help Interfaces: The Effects of Text-to-Speech Voice and Three-Dimensional Avatars," *International Journal of Human-Computer Interaction* (19:1), pp. 75–94.

Qiu, L., and Benbasat, I. 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems," *Journal of Management Information Systems* (25:4), pp. 145–182.

Saad, U., Afzal, U., El-Issawi, A., and Eid, M. 2017. "A Model to Measure QoE for Virtual Personal Assistant," *Multimedia Tools and Applications* (76:10), pp. 12517–12537.

Stout, N., Dennis, A. and Wells, T. 2014. "The Buck Stops There: The Impact of Perceived Accountability and Control on the Intention to Delegate to Software Agents," *AIS Transactions on Human-Computer Interaction* (6:1), pp.1-15.

Venkatesh, V., and Goyal, S. 2010. "Expectation Disconfirmation and Technology Adoption: Polynomial Modeling and Response Surface Analysis," *MIS Quarterly* (34:2), pp. 281-303.

Wang, W., and Benbasat, I. 2008. "Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for E-Commerce," *Journal of Management Information Systems* (24:4), pp. 249–273.

Yoo, Y., and Alavi, M. 2007. "Media and Group Cohesion: Relative Influences on Social Presence, Task Participation, and Group Consensus," *MIS Quarterly* (25:3), p. 371–390.