Multimodal Interaction and Believability: How can we design and evaluate the next generation of IPA?

Jiajia Li
Communication University of China
Beijing, China
angellijiajia1988@gmail.com

Yeuk Yu Lee
Carnegie Mellon University
Pittsburgh, USA
yeukyul@andrew.cmu.edu

Believability of Intelligent personal assistants (IPA) has proven to be an important building block of successful human-agent interaction. Yet, only a handful of studies have focused on proposing and validating possible approaches to enhance such believability. We hypothesize that IPAs that are capable of handling multimodal interaction, such as facial expression and hand gestures, can appear more believable to human users. This paper illustrates reasons why such interaction can improve believability, in turn enhancing user's interpersonal rapport with agent. It also discusses design of a study to evaluate believability in human-agent interactions.

Believability, Multimodal Interaction, Social Interaction, Intelligent Personal Assistant.

1. INTRODUCTION

When inventing new technologies, designers are keen to create human-like agents that mimic human's intelligence, appearance, and movement, since these features have often been found to evoke user's social reaction (Reeves and Nass, 1997). This has motivated designers and engineers to further incorporate social capabilities into technology and interfaces, especially for building rapport (Tickle-Degnen and Rosenthal, 1990; Huang, Morency and Gratch, 2011; Zhao, Papangelis and Cassell, 2014), likability (Bartneck et al, 2009) and trustworthiness (Nass, Isbister and Lee, 2000) in human-agent interaction. The most notable form of these artificial intelligent agents is an Intelligent Personal Assistant (IPA). IPA interacts with user mostly through speech on mobile platforms, for purposes like acquiring information, scheduling daily appointment, responding to factual questions, and starting applications (Canbek and Mutlu, 2016). However, IPAs are mostly implemented on mobile platforms, which may limit socio-emotional interactions with human users. In reality, such interactions among people depend on both verbal behavior and nonverbal behavior. Nonverbal behaviors help to establish feeling of connection and intimacy between interacting human (Zhao, Papangelis and Cassell, 2014; Gratch et al, 2007; Cassell, 2000). However, in the domain of Artificial Intelligence, researchers often evaluate the intelligence level and human-like level of IPAs by

their ability of reasoning, self-learning, and problem solving; yet, other modalities, like movement, facial expression, and gesture are often ignored. We therefore hypothesize that multimodal interaction can contribute significantly to a successful human-agent interaction and help promote an agent's believability.

2. DEFINITION OF BELIEVABLE AGENTS

To begin our discussion of believable agents, it is necessary to define what believability means. In this context, a believable agent does not simply mean an agent that possesses honest or reliable character, but one that gives an illusion of life, thus relieving the audience's suspension of disbelief (Bates, 1994). Such an illusion of being lively does not simply refer to a human-like appearance - other traits from visual and mental aspects are also necessary in creating such an illusion. These traits, usually conveyed through multimodal interaction, are what we believe to be crucial in enhancing believability of agents.

3. DISCUSSION OF EXISTING AGENTS

Currently, most commercial IPAs interact with human users using verbal interaction. Most notable examples of these IPAs are Apple's Siri and Amazon's Alexa. Figure 1 sketches a brief picture of these IPAs' interactive model, where verbal information that is gathered by an IPA during

Figure 1: Current interactive model of existing IPAs

conversations with a user is leveraged to analyze the user's intent. Combining the current interaction with user's past history of interactions, the decision making sub-system of IPAs distinguish between task-oriented and social-oriented content. Finally, the system generates verbal or textual response, sometimes with multimedia output, like image and video, to user. These IPAs can interact with a user through only verbal channel, commonly referred to as Voice User Interface (VUI).

The disadvantages of using only VUIs in IPAs are obvious in three aspects: (1) Conversational Content: IPAs are incapable of conveying emotion, varying their social response, and reacting to user's emotional state; (2) Performance: Responses from IPAs demonstrate poor variety of social behavior, preventing users to engage in a long-time interaction with them; (3) Function: IPAs are not good at comprehending user's intent and can only carry out limited conversation spanning one or a few turns. Such conversations mostly take the form where question from a user triggers a response from the agent, and known facts from prior interactions are disregarded in subsequent IPA responses.

These shortcomings of VUI demonstrate that verbal interaction alone is insufficient for successful human-to-agent interaction. IPAs often lack necessary input to determine user's current emotional state and need. However, multimodal interaction can help to solve these problems by allowing IPAs to receive multimodal information inputs from user and deliver multimodal behavior outputs. Compared to one modality input, like verbal mode, multimodal input can help IPAs to infer additional behavioral information of user, while multimodal output can evoke user's social response. Based on information obtained from facial detection,

movement analysis, speech utterance, and tone, we can build a more complete interactive structure in IPA.

4. AN NEW INTERACTIVE MODEL OF IPA

Jonathan Gratch and Stacy Marsella (Gratch and Marsella, 2013) presented a possible computational model of social emotions which emphasizes that social appraisal is associated with working memory indicating environment (e.g. beliefs about the world), social environment (e.g. beliefs about other) and self (e.g. desire). By integrating multi-modal input and output approach to this structure, we propose a new interactive model of IPA, as shown in figure 2.

This newly proposed model incorporates multimodal input and output, in addition to this existing computational model. The IPA is assumed to have an embodied appearance to display multimodal response. In this model, the information obtained from users transfers to intent analysis level, at which IPA interprets users' intention and emotion. Meanwhile, IPA decides which kind of emotion behavior it should perform based on its pre-determined personality and understanding of user, such as user model and interaction history. Finally, IPA generates adaptive verbal behavior and nonverbal behavior as a feedback. This new architecture incorporates multimodal interaction in emotion analysis, which allows IPA to operate on more information detected from users to give a consistent and adaptive feedback.

The multimodal feedback of the IPAs should be governed by its own personality and interactive rules as outlined by its designer. This set of predefined rules is important since it not only assigns human-like traits to the agent and places responses in social

Figure 2: Proposed future interactive model of IPAs

context, but also ensures consistency of the multimodal output from agent. As Katherine Isbister and Clifford Nass mentioned, “Consistency in others allow people to predict what will happen when they engage with them,” which echoes with the social rules people commonly abide to (Isbister and Nass, 2000). By developing a social bond with the agent, users can be expected to have emotions involved in subsequent interactions. As Joseph Bates states, “Emotion is one of the primary means to achieve this believability” (Bates, 1994). By reciprocating social behaviors that are typical associated with emotions, we can expect agents to appear more believable to users, thus leading to a more successful human-agent interaction.

5. STUDY DESIGN

To prove the importance of multimodal interaction, as proposed in the earlier sections, a recommended study design is outlined as follows:

5.1 Interface Design

Prior to the study, we will recommend researchers to develop three versions of a chatbot: one with textual interaction and a textual user interface, one with vocal expression and no physical user interface, and one with both verbal and nonverbal expression, such as speech, movement, facial expression, on an embodied appearance displayed on a multimedia platform, such as TV screen. Researchers should use the same speech selection and sentence construction system across these three versions. The only difference should be the interactive approaches that these chatbots employ. Researchers should, however, be aware of the

Uncanny Valley Phenomenon (Mori, MacDorman and Kageki, 2012) when designing the chatbots. Embedding human-like behavior in cartoon-like character can effectively avoid Uncanny Valley Phenomenon, which is a common approach for animation creators, such as Disney and Pixar, to develop characters.

5.2 User Study Design

Three versions of chatbot should have access to the same database of scenarios and conversational sentences. Study participants are randomly assigned to three groups to interact with a different chatbot. In the experiment, researchers can arrange tasks which participants need to accomplish with the help of the agent. Researchers can then collect data during the course of interaction between user and agent for post-hoc analyses.

5.3 Evaluation

People prefer to accept information provided by people they trust. Therefore, researchers can evaluate believability of chatbot by the amount of information that participant accepts during their interaction with the agent. Meanwhile, researchers can also model believability level using measurements like length of interaction and the quantity of conversational sentences that chatbot uses. For qualitative evaluation, researchers can use questionnaires to gather participant’s subjective opinion of respective chatbot’s believability. (Tickle-Degnen and Rosenthal, 1990; Gratch et al, 2007; Astrid al, 2010).

6. IMPORTANCE

Once the hypothesis is proven, the launch of multimodal interaction approach within existing IPAs can potentially improve users' reliance on agent, thus increasing usage and boosting commercial values of these IPAs. Besides, since multimodal interaction helps to build social relationship between user and IPA (Zhao et al, 2016), the social bond could potentially ease the tension caused by shortcomings of the IPA. This further leads to possible explorations on whether higher believability can potentially increase users' perception of IPAs' ability. These studies will be helpful in shaping characteristics of future IPAs.

7. CONCLUSION

By carefully avoiding Uncanny Valley Phenomenon, IPAs can potentially evoke users' social reactions and emotions by integrating multimodal interaction design, thus open up the possibility of maintaining long-term relationship with human users. In our future work, we are looking forward to explore design of human-like appearance and implementation of multimodal features, as well as validating these approaches to further explore possibility in fostering human-agent interaction.

8. REFERENCES

- Astrid M, Krämer N C, Gratch J, et al. It doesn't matter what you are! Explaining social effects of agents and avatars. *Computers in Human Behavior*, 2010, 26(6): 1641-1650.
- Bartneck C, Kanda T, Ishiguro H, et al. My robotic doppelgänger-A critical look at the uncanny valley, *Robot and Human Interactive Communication*, 2009. ROMAN 2009. The 18th IEEE International Symposium, 2009: 269-276.
- Bates J. The role of emotion in believable agents. *Communications of the ACM*, 1994, 37(7): 122-125.
- Canbek N G, Mutlu M E. On the track of Artificial Intelligence: Learning with Intelligent Personal Assistants. *Journal of Human Sciences*, 2016, 13(1): 592-601.
- Cassell J. Embodied conversational agents. MIT press, 2000: 29.
- Gratch J, Marsella S, Social emotions in nature and artifact. Oxford University Press, 2013.
- Gratch J, Wang N, Gerten J, et al. Creating rapport with virtual agents, *International Workshop on Intelligent Virtual Agents*. Springer Berlin Heidelberg, 2007: 125-138.
- Huang L, Morency L P, Gratch J. Virtual Rapport 2.0, *International Workshop on Intelligent Virtual Agents*. Springer Berlin Heidelberg, 2011: 68-79.
- Isbister K, and Nass C Consistency of Personality in Interactive Characters: Verbal Cues, Non-Verbal Cues, and User Characteristics *International Journal of Human- Computer Studies* 53 (2000): 251-267.
- Mori M, MacDorman K F, Kageki N. The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, 2012, 19(2): 98-100.
- Nass C, Isbister K, Lee E J. Truth is beauty: Researching embodied conversational agents. *Embodied conversational agents*, 2000: 374-402.
- Reeves B, Nass C. The media equation: how people treat computers, television, new media like real people places. *Computers and Mathematics with Applications*, 1997, 5(33): 7.
- Tickle-Degnen, L., Rosenthal, R.: The Nature of Rapport and its Nonverbal Correlates. *Psychological Inquiry* 1(4), 285-293 (1990)
- Zhao R, Papangelis A, Cassell J. Towards a dyadic computational model of rapport management for human-virtual agent interaction, *International Conference on Intelligent Virtual Agents*. Springer International Publishing, 2014: 514-527.
- Zhao R, Sinha T, Black A W, et al. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behaviour, *International Conference on Intelligent Virtual Agents*. Springer International Publishing, 2016: 218-233.