It's not What It Speaks, but It's How It Speaks: A Study into Smartphone Voice-User Interfaces (VUI)

Jaeyeol Jeong and Dong-Hee Shin(✉)

Department of Interaction Science, Sungkyunkwan University, Seoul, Korea
{jael0, dshin}@skku.edu

Abstract. Since voice-user interfaces (VUI) are becoming an attractive tool for more intuitive user interactions, this study proposes a between-subject experiment in which variations in voice characteristics (i.e., voice gender and manner) of VUI are examined as key determinants of user perceptions. This study predicts that the voice gender (male vs. female) and manner (calm vs. exuberant) are likely to have significant effects on psychological and behavior outcomes, including credibility and trustworthiness of information delivered via VUI.

Keywords: Voice user interface · Voice gender · Voice manner · Smart device · Credibility · Trust

1 Introduction

As smart devices (e.g., smartphones, smartwatches, smart TVs) have become ubiquitous, an increasing emphasis is being placed on voice-user interfaces (VUI), such as Apple Siri and Google Voice Search, as an effective solution to providing greater interactivity and more positive user experience. VUI enables users to interact with computers through the voice recognition and command. It offers more convenient, safe, and user-friendly interfaces than the conventional touch-based interface, allowing seemingly hands-and eyes-free interactions via the device.

Voice is an integral component of human-human interaction because it reflects and conveys emotions, intentions, and manners of the speaker. Judgments are often made simply based on how credible and believable the speaker's voice sounds [1]. But, does the voice also matter in human-computer interactions? Does it influence how users evaluate the credibility of the information delivered by digital media and its source? The Computers Are Social Actors (CASA) paradigm suggests that human-human and human-computer interactions share similar characteristics, such that users mindlessly apply the same social rules and expectations when using computers with social cues as they would in human-human interactions [2]. Given that voice is a humanlike attribute that functions as a strong social cue, the CASA paradigm is also likely to be applicable to investigating the role of voice characteristics in inducing greater credibility of information delivered via VUI.

Ample research on human-computer interaction has demonstrated that variations in voice characteristics lead to different user attitude and perception. For example,

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2015, LNCS 9170, pp. 284–291, 2015.
DOI: 10.1007/978-3-319-20916-6_27

individuals are found to apply gender stereotypes to computers and machines. Product descriptions provided by male voice are perceived as more credible than those with female voice [3]. Women with masculine voice are perceived to be more rational and persuasive than those with feminine voice [4]. In addition, masculine voice is rated as more competent than feminine voice, regardless of the actual gender [5].

Given that such stereotypical vocal cues are found to affect the persuasiveness, credibility, and competence of the delivered messages [6], they are also likely to have notable effects on the ways in which users evaluate and perceive information conveyed via VUI. Therefore, this study intends to explore whether variations in voice gender (male vs. female) and voice manner (dynamic vs. calm) of VUI contribute to affecting psychological and behavioral outcomes.

2 Literature Review and Relative Works

Voice Interface. Popularity of voice interfaces on smartphones have been growing recently. As for most users typing on a phone is more laborious than typing on a full-size keyboard, voice interfaces serve a real practical purpose. Voice recognition technology is nothing new – having been around since the 1950 s when introduced by Bell Labs. However, it wasn't until the 1990 s when IVR systems (interactive voice response) systems became more widespread that the technology really began to solidify. A certain number of studies about audio technology in smart device have been conducted. This includes speech recognition [7], sound recognition [7, 8], speech synthesis [9] or dialogue [10–12].

Voice interfaces have gained public recognition since the release of Siri on iPhone. Siri's popularity is an exciting development for speech technologies. Many voice interfaces with functionalities similar to Siri's are also available on Android smartphones, including Google's Voice Search, SpeakToIt assistant, Vlingo Assistant, Jeannie, and Eva. Google Android Search interfaces that focus exclusively on search functions include Dragon and Bing search. An example of Apple's Siri, which can process a user's speech in natural language, reply the user within a reasonable period of time and perform routine tasks. iPhone users can make reservations at specific restaurants, buy movie tickets or call a taxi by dictating instructions in natural language to Siri. Users can also pose simple queries such as "What is the weather tomorrow?"

Why Voice makes Sense as a User Interface. As consumers grow more comfortable with voice recognition technology for everyday tasks, developers are putting speech into a variety of devices including remote controls, cars and wearable technology. Voice as a user interface makes sense because it is something that comes naturally to humans and is used on a daily basis. According to research conducted by Matthias Mehl, the average number of words spoken by an adult per day is 16,000, while the average number of words typed per day is 3,000 to 4,000 [13]. A key advantage is the fact that voice is not only hands-free, but eyes-free, making it suitable for use in a variety of environments, as well as for the visually-disabled. And unlike most written words, voice can communicate mood, gender, identity (such as recognizing a voice), emphasis and even personality.

Voice Interface Types. Voice interfaces help users perform functions such as sending an email or a text message, playing a song from a music library, accessing calendar, performing a web search, or checking weather forecast. The interfaces differ in the functions that they support, modality of communication, text-to-speech and recognition components, amount of initiative taken by the system, and dialogue handling methods. Virtual assistant voice interfaces take a role of a personal assistant. Siri, SpeakToIt assistant, and Eva/Evan apps speak back to a user. They address the user by name and exhibit emotions both in a choice of a spoken response and in a facial expression. SpeakToIt assistant shows happiness when it is turned on by saying "Good to see you again", apologizes when they do not understand or is unable to handle a command "I am sorry I'm not able to do that just yet but I will be soon". When iPhone loses network connection, Siri (which relies on the connection) responds with error messages, such as "There is something wrong", "I cannot answer you now". These responses sound very cute and human-like and give an impression that the app has a personality, however more helpful responses "Network is down, try again later" or "Try turning on your wireless connection", would reveal the problem to a user or suggest a solution.
Quality of TTS makes a big difference for the perceived quality of a voice app. Neither of the mechanical voices used by the free versions of Android systems compares to affective Siri's voice. SpeakToIt and Eva assistants also have a graphical persona. A character in SpeakToIt assistant app has a customizable appearance and displays subtle facial expressions during communication. However, it is not clear how much this adds to the system's functionality.

Voice Search Interfaces. Other types of apps, such as Dragon and Bing specialize on search only. These interfaces are not attempting to be 'can-do-all' assistants. Instead, they have a focused set of functions relevant to search. Bing has a pre-set list of search types: images, videos, maps, local, etc. We found this helpful because it suggests to a user which functionalities are supported by the system. Dragon, on the other hand, provides a speaking-only interface equivalent to 'how may I help you'. A user can guess by trial and error what the system capabilities are.

User Evaluation of Speech Output. User evaluation of speech output comprises three factors; the users' personal preferences; the operational context; and overall system functionality. These factors are not independent of each other and will integrate to inform the user in their evaluation of speech output. Personal preferences for speech are constructed through the application of the social representations held by the listener. Social representations are created by, for example, norms and stereotypes, and are employed to facilitate communication with in social groups. Simply based on the 'sound' of speech, social representations are formed in relation to, for example, geographical origin, gender and age [14].
A large amount of research has focused on the social representations generated by speech as 'human' output, but few studies consider those generated by speech as system output. The limited number of studies that have investigated social reactions to synthetic speech have discovered that synthetic speech, is often reacted to in similar ways to natural(human) speech [15]. For example, personality and gender are awarded to synthetic speech outputs and gender stereotypes may arise similarly for both natural and synthetic speech [4–14]. Given that speech output is evaluated on the basis of user

preferences informed by social representations, it appears appropriate to consider the smartphone users' social representations of speech outputs to determine the spoken characteristics that provoke both positive and negative user evaluations.

Speech Gender. Even when a speaker is not present, speech is extremely useful in identifying the gender of a speaker and provokes associated social representations [15]. Additionally, Reeves and Nass [4] argue that everyone assigns gender to both natural and synthetic speech outputs. They found that even when the content of the speech output was identical, male speech outputs that praised the user were better received than female speech outputs; male speech output was evaluated as more friendly than female speech outputs; and male speech outputs were evaluated as better information providers on computers. Here, it would be interesting to investigate the current social representations of gender that are invoked in relation to speech as smart phone output and the impact that gender may have on smartphone user evaluations.

3 Research Question

In human-human interaction, gender bias affects credibility of agent. For example, male news casters are rated as more credible and competent than female casters [18]. Research in human-computer interaction has also demonstrated that users apply similar gender stereotypes when interacting with computers [19]. By extension, this gender bias is also likely to be discovered when exposed to VUI with either male or female voices and affect user experience with VUI and perception of information conveyed via VUI if the CASA paradigm is indeed applicable to the smartphone VUI context.

In addition to voice gender, voice manner is another critical vocal characteristic that contribute to shaping user perceptions. Voice manner is largely determined by variations in pitch, such that voice with calm manner is defined as having low pitch and monotone sounds while voice with dynamic manner is defined as having high pitch and wide range of varied tones. Researchers have demonstrated that voice with dynamic manner is generally perceived to be more enjoyable and useful [20]. Together, voice gender and manner are likely to serve as critical components of VUI and influence user experience and perception. The following research question is aimed at examining this possibility.

RQ: Does voice gender (male vs. female) and voice manner (dynamic vs. calm) of VUI affect credibility and trust in messages delivered via VUI and enjoyment of interacting with VUI?

4 Methodology and Analysis

A 2 (voice gender: male vs. female) × 2 (voice manner: dynamic vs. calm) factorial design experiment (Fig. 1) will be conducted. Participants will be recruited from a large private university in Seoul, Korea, and paid five dollars for their participation.

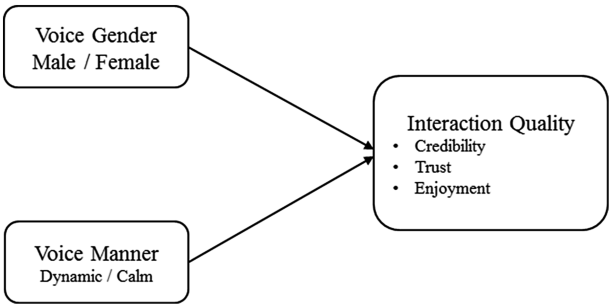

Fig. 1. Research model

4.1 Stimulus Material

The researcher will select 10 restaurants near the experiment site and create verbal descriptions of each restaurant. In order to select restaurants that do not elicit strong positive or negative reactions, 15 respondents will participate in a pretest that is designed to rate their familiarity and involvement in each restaurant (e.g., "this place is relevant/irrelevant," "important/unimportant," and means nothing/means a lot to me") [9]. Based on the result of this pretest, five restaurants rated as the most neutral will be selected. Next, professional voice actors will be hired to record the descriptions of the selected restaurants in the four different types (i.e., voice gender x voice manner) of VUI.

4.2 Procedure

80 undergraduate and graduate students will be recruited from a large university in Seoul, Korea. Upon arrival at the laboratory, participants will be randomly assigned to one of the four conditions and given a brief instruction on operating the VUI. Participants will then launch the VUI app and ask it to recommend a restaurant near the university. The VUI app will recommend one of the five selected restaurants. After receiving the recommendation from the VUI app, participants will complete a questionnaire with a paper and pencil measuring the following variables.

4.3 Dependent Variables

Participants will respond to the questionnaire by marking on 7-point Likert scales (1 = "strongly disagree/not at all," 7 = "strongly agree/very much so"). *Credibility* will be measured with 10 items adopted from [19]: believable, trustworthy, convincing, credible, reasonable, unquestionable, conclusive, authentic, honest, and likable. *Trust* will be assessed by two items adopted from [20]: "I have to be cautious about recommendation made by the smartphone because it is somewhat questionable (reversed)" and "The smartphone voice was warm and caring." *Enjoyment* will be measured with two items adopted from [19]: "Using the smartphone with VUI was exciting" and "I enjoyed using the smartphone with VUI" [20].

Participants will also respond to the manipulation check items that are designed to confirm that the independent variables are successfully manipulated: "The smartphone had male/female voice" and "The smartphone had dynamic/calm voice."

5 Discussions and Conclusion

What makes a user perceive system as intelligent? Subjective characteristics play an important role in user perception: such as witty and varied responses, personalized addressing to a user, quality of TTS and affective intonation. Objective characteristics that affect a user's perception, besides quality of speech recognition, include the ability of the system to communicate to a user what the system capabilities are, effective error detection and handling. Voice interfaces that take initiative can be perceived as more intelligent, however taking initiative also means taking a risk of annoying a user with an unnecessary question or request. The new popularity of voice interfaces on smartphones is an exciting opportunity that can drive advances in dialogue research.

The advantages of the voice-based interaction design have been highlighted as the following:

- It provides a simple, usable and interesting user interface and satisfies the need for more freedom in a human computer interaction environment.
- It provides people new experience and great pleasure which traditional interaction could not offer.
- It makes the interaction between human and computer more natural. It has been illustrated in science fiction movies that this technology can improve people's lives if it is applied rightly.
- It is widely used in various application areas since it gives the user a new experience of feeling.

In this paper we have study the trends in Voice User Interface and experimental design. Smart devices become more autonomous they also have become increasingly present in human world. More research can follow the upcoming trends. We also concluded that the biggest challenge will be to keep the user engaged and continually using the device. Voice is the most natural and easiest fit to keep consumers engaged and will become a necessity as most wearable devices will be in increasingly smaller form factors. For example, a number of smartphones such as Moto X are widely incorporating this technology via Google Now to deliver a "touchless" experience. A lot of the assistants have very robotic voices, but a few(e.g. Evi, assistant) have very nice/real voices. A few have their own voices (e.g. VoiceBrief), and SVOX has custom voices for android.

As speech synthesis and speech recognition technologies improve, applications requiring more complex and more natural human-computer interactions are becoming feasible. A well-designed user interface is critical to the success of these applications. A carefully crafted user interface can overcome many of the limitations of current technology to produce a successful outcome from the user's point of view, even when the technology works imperfectly. With further technological improvements, the primary role of the user interface will gradually shift from a focus on adapting the user's

input to fit the limitations of the technology to facilitating interactive dialogue between human and computer by recognizing and providing appropriate conversational cues. Among the factors that must be considered in designing voice interfaces are the task requirements of the application, the capabilities and limitations of the technology, and the characteristics of the user population. This paper discussed how these factors influence voice user interface design and then describes components of voice user interfaces that can be used to facilitate efficient and effective human-computer voice-based interactions.

6 Contribution of This Study

This study intends to demonstrate whether variations in vocal characteristics of VUI influence user experience and perception. In doing so, this study will offer both theoretical and practical insights for manufacturers, designers, and communication scholars who are interested in the role of VUI in promoting adoption for emerging smart communication media such as smartphones and smartwatches.

Acknowledgment. This research was supported by the Ministry of Education, South Korea, under the Brain Korea 21 Plus Project (Grant No. 10Z20130000013).

References

1. Addington, D.W.: The effect of vocal variations on ratings of source credibility. Speech Monogr. **38**, 242–247 (1971)
2. Sundar, S.S., Nass, C.: Source orientation in human-computer interaction programmer, networker, or independent social actor. Commun. Res. **27**(6), 683–703 (2000)
3. Morishima, Y., Nass, C., Bennett, C., Lee, K.M.: Effects of 'gender' of computer-generated speech on credibility. Technical report of IEICE TL2001-16, 31(8), pp. 557-562 (2001)
4. Reeves, B., Nass, C.: How People Treat Computers, Television, and New Media Like Real People and Places. CSLI Publications and Cambridge University Press, New York (1996)
5. Ko, S.J., Judd, C.M., Blair, I.V.: What the voice reveals: within-and between-category stereotyping on the basis of voice. Pers. Soc. Psychol. Bull. **32**(6), 806–819 (2006)
6. Leigh, T.W., Summers, J.O.: An initial evaluation of industrial buyers' impressions of salespersons' nonverbal cues. J. Pers. Selling Sales Manage. **22**(1), 41–53 (2002)
7. Vacher, M., Fleury, A., Portet, F., Serignat, J.F., Noury, N.: Complete sound and speech recognition system for health smart homes: application to the recognition of activities of daily living. In: New Developments in Biomedical Engineering, pp. 645–673 (2010)
8. Rougui, J.E., Istrate, D., Souidene, W.: Audio sound event identification for distress situations and context awareness. In: Annual of the IEEE International Conference on Engineering in Medicine and Biology Society, EMBC 2009, pp. 3501–3504. IEEE (2009)
9. Lines, L., Hone, K.S.: Multiple voices, multiple choices: older adults' evaluation of speech output to support independent living. Gerontechnology **5**(2), 78–91 (2006)
10. Gödde, F., Möller, S., Engelbrecht, K.P., Kühnel, C., Schleicher, R., Naumann, A., Wolters, M.: Study of a speech-based smart home system with older users. In: International Workshop on Intelligent User Interfaces for Ambient Assisted Living, pp. 17–22 (2008)

11. Hamill, M., Young, V., Boger, J., Mihailidis, A.: Development of an automated speech recognition interface for personal emergency response systems. J. NeuroEngineering Rehabil. **6**(1), 26 (2009)
12. López-Cózar, R., Callejas, Z.: Multimodal dialogue for ambient intelligence and smart environments. In: Nakashima, H., Aghajan, H., Augusto, J.C. (eds.) Handbook of ambient intelligence and smart environments, pp. 559–579. Springer, Heidelberg (2010)
13. Stevens, C., Lees, N., Vonwiller, J., Burnham, D.: On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. Comput. Speech Lang. **19**(2), 129–146 (2005)
14. Lines, L., Hone, K.S.: Multiple voices, multiple choices: older adults' evaluation of speech output to support independent living. Gerontechnology **5**(2), 78–91 (2006)
15. Mullennix, J.W., Stern, S.E., Wilson, S.J., Dyson, C.L.: Social perception of male and female computer synthesized speech. Comput. Hum. Behav. **19**(4), 407–424 (2003)
16. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**, 457–500 (2007)
17. Roe, D.B., Wilpon, J.G. (eds.): Voice Communication Between Humans and Machines. National Academies Press, Washington, DC (1994)
18. Brann, M., Himes, K.L.: Perceived credibility of male versus female television newscasters. Commun. Res. Rep. **27**(3), 243–252 (2010)
19. Niculescu, A., Van Dijk, B., Nijholt, A., See, S.L.: The influence of voice pitch on the evaluation of a social robot receptionist. In: 2011 International Conference on User Science and Engineering (i-USEr), pp. 18–23. IEEE (2011)
20. Nass, C., Lee, K.M.: Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. J. Exp. Psychol. Appl. **7**(3), 171 (2001)