Perceived Usefulness of Multimodal Voice Assistant Technology

Rianna R. Baeza and Anil R. Kumar
San Jose State University

Smart speaker devices are appealing to consumers, but the perceived usefulness of the multimodal voice experience is not fully understood. The purpose of this study was to evaluate the extent to which cognitive load, the relevance of visual information, and personality influence the perceived usefulness of multimodal voice assistant technology in a within-subjects repeated measures design. A multimodal voice prototype was created to answer the question, "What are some extreme weather conditions?" Nine variants, including 3, 5 and 7 system responses with relevant, irrelevant or no information presented on a screen were included. Three tasks were embedded within each condition (Stroop task, sort M&Ms and no task). Perceived usefulness score, recall, personality score, and fluctuations in galvanic skin response (GSR) values were the subjective and objective measures. The findings suggest that when there's a smaller number of responses/words for the participant to attend to, and subsequently recall, in addition to relevant visual feedback to aid in that recall, they perceive the voice assistant experience to be more useful, while task conducted exhibits marginal significance in determining PU. Scores of conscientiousness, openness to experience, agreeableness, and neuroticism were successful in predicting some variation in the PU responses, while GSR data was not. It is highly recommended that UX designers of the multimodal interface create succinct voice responses with relevant visual feedback to accompany it, and to keep the main use cases of these products in mind to increase the experience's PU and subsequent behavioral intention to use the product.

BACKGROUND

An estimated one in five U.S homes own a smart speaker with an integrated voice assistant (Voicebot.ai, 2018). Because people have been communicating through voice for centuries, the introduction of the voice interface was initially thought to invoke the most natural human interaction with technology. However, voice user interface designers are constantly limited by the capabilities of its software and are told to design with care due to the limited memory capacity of human users (Bigot et al., 2013). Recently these voice interface systems have become multimodal, sharing the same functionality as the standalone speaker with the addition of visual feedback through a screen. This permitted designers to provide relevant visual information to voice interactions and potentially aid in the limited memory capacity of users. (Voicebot.ai, 2018). Interestingly, the overall usage data of the devices, with and without screens, shows 30% of consumers don't use their smart speaker at least once a week (Voicebot.ai, 2018). According to the Technology Acceptance Model (TAM), perceived usefulness (PU) is highly influential in a person's behavioral intention to use a product (Davis et al, 1989). Additionally, content quality significantly affects perceived usefulness of virtual assistant devices, but it is unclear if quality of visual information yields the same significance (Yang & Lee, 2018). Literature in this area of research emphasizes the use of recall of voice system responses to better understand the working memory limitations of voice interface designs, that may be influential in determining the PU of a product experience (Bigot et al., 2013). To physiologically measure cognitive load, galvanic skin response (GSR) systems are commonly used and found to be valid (Khawaji et al., 2015) Also important to note, there is little understanding of what, if any, individual personality variables influence the perceived usefulness of voice assistant technology, as some have been found to account for significant variation in PU scales of other technology products, such as smartphones (Devaraj et al., 2008) There is an emphasis on measuring the quality of experience, of new technology, by using a combination of objective, subjective and physiological measures to understand the overall usability (Weiss et al., 2015). The optimal methods to test the feature specific usability of new voice user interface designs are outlined and easily accessible for researchers (Weiss, Wechsung & Kuhnel, 2015). However, usability is only one factor that influences our perception and intent to use a product experience. With only 70% of these devices being used at least once a week, the perceived usefulness of these voice experiences needs to be explored in more depth.

Rationale and Objectives

The purpose of this study was to understand and evaluate the extent to which cognitive load, personality and the relevance of visual information influence the perceived usefulness of multimodal voice assistant experiences. This study design strives to understand how these variables can influence the PU, and overall BITU a multimodal voice experience. Subsequently, findings will inform designers, researchers and engineers creating the voice interface about how cognitive load, the visual information presented on a multimodal device, and individual personality differences impact the perceived usefulness of voice assistant technology. The hypotheses for this study were as follows.

- H1 High cognitive load conditions will negatively influence PU
- H2 Irrelevant visual feedback will negatively influence the PU
- H3 Relevant visual feedback presented on the screen will positively influence PU
- H4 Percentage of recall will be higher for the lower cognitive load conditions

H5 High scores of neuroticism on FFM inventory will negatively influence PU
H6 High scores of agreeableness on the FFM inventory will positively influence PU

METHOD

Participants

This study was approved by the San Jose State University Institutional Review Board (IRB). A total of 16 participants (8 females and 8 males) with an average age of 20.17 years old and variety of experience with voice assistants participated in this study. Participants with significant hearing impairments and impairments in vision that impact color perception, including color blindness were excluded in this study.

EXPERIMENTAL DESIGN

A within subject's design was implemented, including three levels of independent variables (See Table 1). The conditions were randomly presented for each participant before each session.

Table 1. Design with Levels of Independent Variables

Visual information and relevance	Number of Responses		
	3	5	7
Voice Only	A (NT) Summer	C (ST) Fall	B (SMM) Winter
Irrelevant Visual Info	B (ST) Fall	A (SMM) Winter	C (None) Summer
Relevant Visual Info	C (SMM) Winter	B (None) Summer	A (ST) Fall

NT = no Task, ST = Stroop task, and SMM = Sort M&M's

Control Variables

The subject matter of the interaction was weather conditions for all treatments since obtaining weather information is the most daily-used skill across all voice assistant devices (Voicebot.ai, 2018). Extreme weather options were used as the content of responses from Summer, Fall and Winter seasons to ensure variety across the nine conditions. Spring season weather conditions was only provided in the training condition. Summer, Fall and Winter weather conditions was presented once in the 3, 5 and 7 option conditions and fixed in that treatment for all participants. The participant utterance to interact with the voice assistant, also called invocation phrase, was the same for all conditions. This statement was, "Can you give me some extreme weather conditions?" Distance from the device was also be controlled. All participants were measured at 4 feet distance away from the multimodal prototype.

Independent Variables

Number of responses. The voice assistant was designed to provide 3, 5, or 7 responses after each participant utterance of the invocation phrase, which is commonly used in a variety of voice interface cognitive load studies (Miller et al., 2013). There was a two-word maximum in each response presented, with no more than three syllables in each word. No suffix was included in the voice designs. A suffix in a voice interface design is generally used to instruct the participant when it is their turn to speak. This was not be included due to the finding that the recency effect is significantly decreased in short term memory recall when a suffix phrase is included at the end of a voice interaction (Bigot et al., 2013).

Visual information relevance. The visual information presented on the screen was either absent, irrelevant, or relevant to the voice assistant responses. An absent screen was presented as a black background that did not provide any visual information to the participant. The irrelevant visual feedback condition was provided as written text of a news article title which typically is the main background of voice assistant multimodal devices, especially Amazon Echo Show. Relevant visual information conditions included icons depicting the weather conditions for each option presented. Figure 1 provides a representation of the screen designs for multimodal visual display

Figure 1. Screen Designs for Multimodal Visual Display

Assigned task. The top contexts in which individuals use a smart assistant on their phone are (1) driving, (2) doing household chores and (3) cooking (Voicebot.ai, 2018). Although these may not be the same common tasks individuals perform when interacting with smart speakers, it is important to have the participant engage in a task to better grasp an organic use experience. Task assignment was fixed within each condition. Three task levels were implemented in this design. A muscle memory task, which was designed to not completely load the participants working memory, was assigned in three conditions. This task required the participant to separate M&Ms by color. An active task was also assigned in three conditions, requiring users to conduct a Stroop test, which has been widely used as a cognitive load measure to invoke the impedance of the processing of a stimulus attribute (Stroop, 1935). The remaining three conditions did not require the participant to perform a task.

Dependent Variables

Personality inventory. The FFM, big-five inventory, contains 50 questions on a Likert scale, which was given to each participant to complete. A score for openness to experience,

conscientiousness, extroversion, agreeableness and neuroticism from the validated model was calculated based on the instructions given in the psychometric test directions.

Recall. After the voice assistant gives 3, 5, or 7 responses to their invocation phrase, the participants were asked to recall every response that they remember out loud two seconds after the voice assistant is done speaking. Each word would count as one point.

Perceived usefulness scale. Three validated scaled inventory questions, presented in the initial proposal of the technology acceptance model in Davis et al. (1989) to measure perceived usefulness, was used with key words replaced to make it more relevant to the voice assistant experience. The following are the three questions that were asked on a Likert scale 1-7. With this set up, the highest score that could be obtained was 21.

Galvanic skin response (GSR). As a physiological measure to understand the effect of the levels of cognitive load and task assignments in the study design, a GSR system (Procomp Infiniti system with Physiology Suite) was used to analyze the potential fluctuation of sweat secretion during each task. Non-invasive disposable electrodes (nonirritant material) that adhere to the skin were used to measure the continuous variations in the electrical characteristics of the skin. A skin conductance value in micro-mhos μm, synonymous to μS, was continuously recorded at 5 second intervals throughout the session. A baseline was taken before the participant takes the personality test. A μS value was recorded after the recall was completed for each of the 9 conditions.

Materials

Software. To create a voice and graphical user interface to adequately mimic a multimodal experience, Adobe XD's Amazon Alexa toolkit was used. A variety of choices including voices of different genders and ages was available to use with this software. All options provided by this software are dissimilar to the widely available voices output by Alexa, Google Assistant, Siri and others. Because the standard voices of these assistants are female and sound adult in age, the synthetic voice option chosen was English US, Female, named Joanna. The screen variant designs were also created using the Adobe XD software.

Hardware. A Windows PC laptop will be used to analyze GSR fluctuations via Procomp Infiniti system hardware.
A 9.7-inch iPad will be used to project the prototype of the 9 conditions created in Adobe XD, outputting the artificial voice. The iPad also served as the device that provides the visual feedback conditions, because its size is more similar to those of the current models of multimodal voice assistant devices.

Other materials. To invoke a muscle memory task that is similar to natural use cases where a user's hands are occupied, M&Ms were provided for a separation by color task. In addition, a printed version of a Stroop task, originally presented in Stroop (1935), was provided to the participant to use during an active task condition. A printed version of the big five personality inventory with 50 scale questions was also provided to the participants.

RESULTS

Kolmogorov-Smirnov test for normality conducted for the response variables (percentage of recall, GSR, % of GSR change from baseline, and PU scores) did not indicate any assumptions of normality. A Levene's test for equal variances of the response variables assumption was met for recall, PU scores, GSR scores, and % GSR change. The significance level was set at 0.05.

Descriptive Statistics

Table 2 presents a summary of the responses obtained for the dependent variables based on the different combinations of independent variables. A review of the table indicates that participants recalled the most voice assistant responses (M = 85.4%, SD = 0.17), when there were 3 weather condition responses presented, relevant visual information on the screen, and a sorting M&Ms task being conducted followed by 3 weather condition responses presented, voice only, and no task (M=.844, SD = 0.18). The lowest recall was obtained for the 7 weather response condition with irrelevant information with no task assigned. The table also indicates that for the Sum fo PU inventory, the highest and lowest scores were obtained for the same combination as above. In other words, 3-response condition with relevant corresponding visual weather information on the screen, and a low cognitive load task of sorting M&Ms being conducted (M = 15.375, SD= 3.73)) and 7 weather response condition with irrelevant information with no task assigned (M = 9.44, SD = 4.21). For the GSR measure, the highest mean GSR score was obtained in the 7-response condition, where no visual information was presented on the screen and the sorting M&Ms task was being conducted (M =2.685, SD = 1.97).

Table 2. Descriptive Statistics of IV and DVs

Number of Responses	Visual Information	Task	Percentage of Recall Mean (SD)	Sum of PU Mean (SD)	GSR Score Mean (SD)
3	Irrelevant	Stroop	0.708 (0.18)	13.810 (4.09)	2.602 (1.99)
	Relevant	Sort M&Ms	0.854 (0.17)	15.375 (3.73)	2.631 (2.15)
	Voice Only	No Task	0.844 (0.18)	14.625 (3.59)	2.525 (1.69)
5	Irrelevant	Sort M&Ms	0.469 (0.11)	12.750 (3.58)	2.569 (1.59)
	Relevant	No Task	0.519 (0.15)	13.060 (4.25)	2.457 (1.56)
	Voice Only	Stroop	0.394 (0.12)	11.063 (2.74)	2.384 (1.48)
7	Irrelevant	No Task	0.322 (0.11)	9.440 (4.21)	2.381 (1.50)
	Relevant	Stroop	0.339 (0.12)	10.813 (3.60)	2.643 (2.36)
	Voice Only	Sort M&Ms	0.411 (0.14)	11.630 (4.70)	2.685 (1.97)

Recall

The average percentage of recall was 80.21% for 3 responses, 46.04% for 5 responses, and 35.74% for 7 responses (See Figure 2). Results from ANOVA indicate the number of responses (F(2,114),p = 0.000), visual information presented on the screen (F(2,114),p = 0.019), and the task that is being conducted significantly influence the recall (F(2,114),p = 0.009). Interaction effects between variables did not yield any significance. Post Hoc Tukey analysis for variations in the mean recall was performed for number of responses, visual information, and task. Grouping data obtained from the post hoc analysis using Tukey procedures for number of responses indicated that each response (3, 5 and 7) were significantly different for percentage of recall. Visual information and task grouping data did not indicate any significant differences

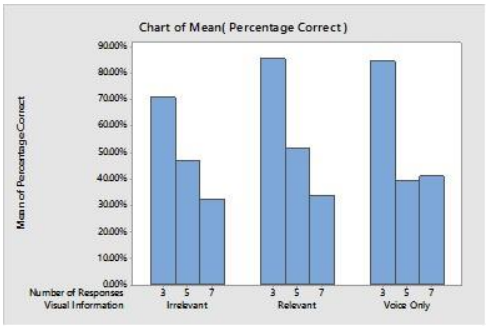

Figure 2. Mean of Percentage of Recall

Perceived Usefulness Inventory

Results from the ANOVA indicate that the number of responses were significant in l predicting the perceived usefulness of the voice assistant experience in the various conditions presented (F(2,114), p = 0.000) while the presentation of the visual information was not significant (F(2,114),p = 0.462). The highest mean PU value was obtained for 3 responses and relevant information presented on screen (Figure 3). The order that the participant received the condition, visual information relevance, and task conducted were not found significant in predicting variation in the sum of PU values.

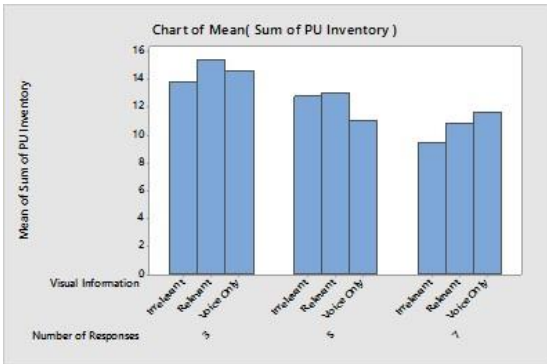

Figure 3. Sum of PU by Independent Variables

Interactions between variables were checked for multicollinearity and did not yield significant p-values. Post hoc analysis using Tukey procedures for the variations of visual information, number of responses, and task were conducted for the PU scores. Grouping data for number of responses indicates the 3, 5, and 7 response conditions yielded significantly different PU scores from 3 responses compared to 5 and 7. Groupings for visual information and task conditions were not significantly different in predicting PU scores.

Personality

A Kolmogorov-Smirnov test for normality of the scores was met for all variables, agreeableness (p = 0.152), extroversion (p = 0.155), conscientiousness (p = 0.121), neuroticism (p = 0.241), and openness to experience (p = 0.129). The sample of this study yielded a mean of 17.06 for extroversion, 28.87 for agreeableness, 23.68 for conscientiousness, 21.12 for neuroticism and 23.93 for openness to experience. A regression analysis for the five personality scores' prediction of the subsequent perceived usefulness of the voice assistant conditions was conducted. Scores for neuroticism (p = 0.020), conscientiousness (p = 0.003), openness to experience (p = 0.00), and neuroticism (p = 0.020) account for significant variation in the PU inventory scores throughout the session. Coefficient values for significant variables in this model are the following: neuroticism (+0.1640), conscientiousness (+0.2183), agreeableness (-0.1595) and openness to experience (+0.2701). Extroversion (-0.0809), was not statistically significant in predicting variation with a p-value of 0.120.

GSR

ANOVA for GSR scores after recall for each condition did not yield any significance for number of responses (p = 0.471), visual information (p = 0.936), and task (p = 0.210) %GSR (i.e. change from the baseline measurement) was used as a response variable in the ANOVA analysis. As before, Number of responses (p=0.991), visual information (p = 0.232) were found to be insignificant, while the task (p = 0.023) was significant.

DISCUSSION

Based on results from current study, it appears that PU scores across conditions show the amount of responses presented by a voice assistant is important in determining variation between participants, while the task conducted and visual information are not. The highest mean PU score was found in the 3-response condition with relevant multimodal visual information on the screen. The lowest mean PU value was for the 7-response condition with no visual information on the screen, supporting H1 and H3, but not H2. These findings suggest there is strong evidence that the voice assistant experience is perceived to be more useful when there is smaller number of responses presented along with relevant visual feedback. A potential reason is that this combination helps recall. This finding coupled with the work of Yang & Lee

(2018), which details that quality of content affects PU of the experience produced by virtual assistant devices, leads us to infer that designs with cognitive load, quality visual information and content in mind can increase PU, and subsequent behavioral intention to use.

Recall data results indicates the apparent human memory limitations exhibited when interacting with auditory and audio-visual stimuli. Number of responses, the relevance of visual information, and cognitive load of the task being conducted at the time of the interaction, accounted for the variation in the percentage of recall. Supplying evidence for H4, results imply that the percentage of recall is highly dependent on the amount of information the voice assistant gives in response to user questions, which parallel the findings of Bigot et al. (2013). The attentional requirements of the Stroop test and sorting M&Ms task being performed by participants also influenced recall. Driving, doing household chores, and cooking are the top tasks users perform when interacting with these systems, so it is imperative for designers of the voice interface to be wary of the various levels of cognitive demand these and other common tasks can present.

GSR data gathered in this study showed fluctuations of μS values throughout the session, with gradual increases as the session time progressed. Because of this, order of condition was significant in predicting the variation in the μS values throughout the session. Although there were visible increases in the μS fluctuations during the time the participant recalled responses, ANOVA analysis does not indicate number of responses, the visual information presented on the screen, and task adequately account for this variation. GSR did not adequately measure the cognitive load of the variables in this study, although this measure has been utilized in other studies such as Khawaji et al. (2015).

The sample in this study had relatively high means in agreeableness, low means in extroversion, and average scores in conscientiousness, neuroticism and openness to experience. Agreeableness was significant in the prediction of PU scores, however it yielded a negative coefficient value, opposite of the predicted positive value. Additionally, neuroticism scores yielded a positive coefficient rather than the predicted negative influence on PU scores. There is no support for H5 or H6, dissimilar to the findings of Devaraj et al. (2008).

Limitations

The sample in this study were all college students with an average age of 20.17, which is not representative of the current consumers or users of multimodal voice interface products. A more diverse group of ages could tell us more about what other variables could be influencing the PU of these systems. Further research may benefit from more recent expansions of the TAM that outline perception of risk and trust in a product that can also influence an individual's BITU a system or experience (Venkatesh & Bala, 2008). Additionally, perceived ease of use (PEOU) is also a component of the TAM that adequately predicts BITU, which we did not measure in this study. Finally, studies of trust and cybersecurity factors are imperative in this field as 16% of people who don't own a smart assistant say it is because of privacy concerns, and none of these potential cybersecurity questions were addressed in this study (Voicebot.ai, 2018).

CONCLUSION

In this study based on the current sample size, the results indicate that people perceive multimodal systems to be more useful, and perform recall better, when the voice assistant gives less responses and presents visual information on the screen that is relevant to the question the user asked. Similarly, it is known that the attentional demands of the task when user is performing tasks while interacting with the multimodal system can influence their ability to recall what the voice assistant just said. Particularly in high cognitive load conditions when the assistant provides 7 responses with a total of 14 words. Irrelevant visual information presented on the screen as news stories, significantly decreased the systems score of perceived usefulness and the user's ability to recall the responses. With this new understanding, it is then recommended that UX designers of the multimodal interface create succinct voice responses with relevant visual feedback to accompany it, and to keep the main use cases of these products in mind to increase the experience's perceived usefulness and subsequent behavioral intention to use the product.

REFERENCES

Bigot, L., Caroux, L., Ros, C., Lacroix, A., & Botherel, V. (2013). Investigating memory constraints on recall of options in interactive voice response system messages. *Behaviour & Information Technology*,*32*(2), 106-116.

Bigot, L., Terrier, P., Jamet, E., Botherel, V., & Rouet, J. (2010). Does textual feedback hinder spoken interaction in natural language? *Ergonomics*, *53*(1), 43-55.

Bond C., Camack M.(1999). Your call is important to us, please hold. *Ergonomics in Design*, *7*, 9-15.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly*, *13*(3), 319-340.

Devaraj, S., Easley, R.F. and Crant, J.M. (2008). How does personality matter? Relating the five-factor model to technology acceptance and use, *Information Systems Research*, *19*(1), 93-105.

Khawaji, A., Zhou, J., Chen, F., & Marcus, N. (2015). Using Galvanic Skin Response (GSR) to Measure Trust and Cognitive Load in the Text-Chat Environment. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, *18*, 1989-1994.

Miller, D., Gagnon, M., Talbot, V., & Messier, C. (2013). Predictors of Successful Communication With Interactive Voice Response Systems in Older People. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*,*68*(4), 495-503.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643-662.http://dx.doi.org/10.1037/h0054651

Voicebot.AI (2018, November). Voice Assistant Consumer Adoption Report 2018. Retrieved From https://voicebot.ai/wp-content/uploads/2018/11/voice-assistant-consumer-adoption-report-2018-voicebot.pdf

Weiss, B., Wechsung, I., Kühnel, C., & Möller, S. (2015). Evaluating embodied conversational agents in multimodal interfaces. *Computational Cognitive Science*, *1*(1), 1-21.

Yang, H., & Lee, H. (2018). Understanding user behavior of virtual personal assistant devices. *Information Systems and EBusiness Management*, 1-23.