Multimed Tools Appl
DOI 10.1007/s11042-016-3650-5

A model to measure QoE for virtual personal assistant

Umair Saad[1] · Usama Afzal[1] · Ahmad El-Issawi[2] ·
Mohamad Eid[1]

Received: 12 September 2015 / Revised: 11 April 2016 / Accepted: 30 May 2016

Abstract Until now the virtual assistants (like Siri, Google Now and Cortana) have primarily been confined to voice input and output only. Is there a justification for voice only confinement or can we enhance the user experience by adding a visual output? We hypothesized that providing a higher level of visual/auditory immersion would enhance the quality of user experience. In order to test this hypothesis, we first developed 4 variants of virtual assistant, each with a different audio/visual level of immersion. Developed virtual assistant systems were the following; audio only, audio and 2D visual display, audio and 3D visual display and audio and immersive 3D visual display. We developed a plan for usability testing of all 4 variants. The usability testing was conducted with 30 subjects against eight (8) dependent variables included presence, involvement, attention, reliability, dependency, easiness, satisfaction and expectations. Each subject rated these dependent variables based on a scale of 1–5, 5 being the highest value. The raw data collected from usability testing was then analyzed through several tools in order to determine the factors contributing towards the quality of experience for each of the 4 variants. The significant factors were then used develop a model that measures the quality of user experience. It was found that each variant had a different set of significant variables. Hence, in order to rate each system there is a need to develop a scale that is dependent upon the unique set of variables for the respective variant. Furthermore, it was found that variant 4 scored the highest rate for Quality of Experience (QoE). Lastly several other qualitative conclusions were also drawn from this research that will guide future work in the field of virtual assistants.

Keywords Virtual digital assistant · Quality of experience · Immersive environment · Multimodal interaction · Words and phrases: multimedia systems and tools · Usability study

✉ Mohamad Eid
mohamad.eid@nyu.edu

[1] Division of Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[2] Lebanese University, Beirut, Lebanon

Published online: 16 June 2016

1 Introduction

Virtual Personal Assistant (VPA) is a software agent that provides professional administrative, technical, or social assistance to a human user [25]. Apple's Siri (http://en.wikipedia.org/wiki/Siri) started the trend for VPA by providing a speech interface for questions/answers in 2011. Awareness was created through millions of customers that indeed VPA could speak, be funny, answer questions and perform simple tasks (such as making a phone call, send a text message and reminding about specific events). Since then, several companies including Google Now (http://en.wikipedia.org/wiki/Google_Now), AT&T [15], Nuance (www.nuance.com), Cortana (https://en.wikipedia.org/wiki/Cortana_(software)), etc. have entered the market of VPAs.

VPA can be found in various applications, serving different purposes. Personal assistant, virtual guide, tutor or mentor, are examples of roles these agents can endow. The ultimate goal in VPA is to make the virtual assistant communicate with a human in the same way humans communicate with each other, i.e. by relying on multimodal interaction (audio, visual, haptic, gesture, gaze, etc.).

With the dramatic rise of computation power, increased availability 3D and immersive displays, and artificial intelligence advances, multimodal and intelligent VPA is becoming a possibility. However, user experience is one essential component in guiding the design and development of interaction paradigms for VPA. It has been widely accepted that multimedia research should focus on Quality of Experience (QoE) as the primary metric for evaluating the user experience. QoE is defined as the characteristics of the sensations, perceptions, and opinions of human subjects as they interact with a multimedia system [1].

The main goal of this study is to assess the usability of VPA with various levels of multimedia immersion, ranging from audio-only, to audio-visual in 2D, to audio-visual in 3D (Pepper's ghost 3D visual display), and immersive 3D interactions (immersive virtual reality display). Representative users were asked to complete a set of typical tasks and measures were taken of effectiveness, efficiency and satisfaction.

Human personal assistants communicate using multiple modalities (such as auditory, visual, gesture, gaze, haptics, etc.) in a natural setting. Hence, it seems more natural to use more than one modality for human-like interaction with VPA. In this study, we seek to:

- Objective 1: Determine whether each variant would have different dependent variables contributing to the QoE,
- Objective 2: Quantifying and developing a model for rating QoE for each variant,
- Objective 3: Understand the impact of efficiency and accuracy of the implemented system on the QoE.

2 Related work

2.1 Modeling quality of experience

Researchers have explored modeling the QoE in a multimedia system. One of the earliest works on modeling QoE defines it as a measure of the impact of content on a specific user in a specific context [7]. QoE can be measured through subjective assessment or estimated through a model based on parameters of content, specific user, and specific context.

Nowadays there are three approaches for measuring QoE: subjective, performance-based, physiological measures, and hybrid approaches [34]. Each method enables the collection of specific type of information regarding the user's experience. For instance, subjective measures evaluate the user's satisfaction, fatigue, intuitiveness, preferences, etc., and they are collected typically via surveys. Performance measures evaluate the user's behavior when performing a specific task (such as task completion time, accuracy, error rate, etc.). Finally, the physiological measures evaluate nonvoluntary responses of the human body during and immediately after the test session.

Parameters such as presence, satisfaction, and expectations are extremely difficult to measure with performance-based or physiological-based approaches [37]. Therefore, this study adopts the subjective approach to establish a conceptual framework of QoE for interactive multimedia environments. Wu et al. [37] proposed a framework that divided user experience into cognitive perception and behavioral consequences based on subjective measures. Utilizing the same framework, we define QoE as a composition of 8 dependent variables: presence, involvement, attention, reliability, dependency, easiness, satisfaction and expectations.

2.2 From voice-only to immersive VPA

There is a growing international interest, both in academia and in industry, in developing VPA applications aimed at improving the user performance and the overall quality of user experience. Early implementations of VPA systems provided audio-only modality for interaction. Subsequent researches investigated the use of audio-visual VPA system. Lately, 3D and immersive displays are considered in an effort to provide higher level of immersion.

While dictation technology has not advanced to the state of reliability as depicted in popular science fiction movies, it has been extremely useful in number of contexts [18]. Most of the existing voice recognition systems like Siri and Google Now are mobile applications that use natural language user interfaces to answer questions, make recommendations, and perform actions by delegating requests to a set of web services [21]. They are claimed to be very smart; they adapt to the users individual preferences over time in order to personalize results, and perform tasks such as finding recommendations for nearby restaurants or getting directions [28].

Visual interaction is one of the essential components to develop a VPA into something that can be compared to a human assistant. It is question worth asking if we should even explore the visual aspect in order to improve the contemporary voice assistants. Nonetheless, there are several other components of the contemporary VPAs that are known to be the limitation. Firstly, developing an optimal tone and pitch to deliver that desired impact of the voice is a hot research topic [2]. Secondly, the accurate detection of input commands given in various accents limits the dependency and usage of a VPA system [8]. No one likes to repeat the commands multiple times in order to execute a desired outcome.

Advancements in 3D multimedia paved the way towards more realistic visual representation of the VPA using 3D displays. Examples of existing 3D displays include the Voxi Box [14], Musion Eyeliner [26], Peppers Ghost [23], 3D Fog Projection [39], inFORM [9] and the 2.5D Shape Display [20]. These 3D displays have raised questions whether such displays can be used to create immersive and realistic VPA systems. The ultimate goal is the digital recreation of real-world presence.

During the last decade, virtual humans became very popular in gaming, entertainment, and social media [17]. Research has demonstrated that enhancing virtual reality with digital characters will fundamentally allow us to interact with computers in a human way and on a personalized basis [31]. For example, the authors in [30] examined the roles that virtual humans can play in six areas: performance, physiology, learning, connection, and security. Results showed that virtual humans enhance the quality of experience when multimodal interactions are incorporated. Jun et al. [16] highlighted the significance of not only appearance and behavior but also nonverbal communication and affective components towards the quality of user experience. The relationship between rendering/display fidelity of virtual human models and emotional reactions is investigated in [32].

The evaluation of VPA systems is important to guarantee user satisfaction [38]. Multimodal interfaces combine visual and auditory cues to enrich interactions but they raise new challenges concerning the usability and acceptability of such interfaces. As part of the European Project FASiL (Flexible and Adaptive Spoken Language and Multimodal Interfaces), a VPA system was developed to understand which factors affect the user experience and the acceptance of multimodal services [5]. Results show that a conversational and multimodal approach was very well accepted and supported by the users. Furthermore, the quality and speed of the system feedback as well as the recognition accuracy of the spoken components are key factors to a better user experience. A recent study evaluated VPA to provide the elderly with a wide range of online services such as weather information and social networking [29]. Results demonstrated the need for several input/output modalities, distribution of modalities across different devices (PCs, Tablets, etc.), and adhering to international standards and avoiding closed solutions. A subsequent work evaluated the use of smart phones for VPA in immersive virtual reality [12].

3 Implementation details

3.1 Objectives and experimental variants

The purpose of this research is to evaluate if adding a visual component to the voice-only virtual assistants would actually enhance the user experience. We hypothesized that providing a higher level of visual/auditory immersion would enhance the quality of user experience. In order to test this hypothesis, first stage involved the development of four different variants (sometimes referred to as treatments [35]) of virtual assistant, each with a different audio/visual level of immersion. Developed VPA systems were the following; audio only, audio and 2D visual display, audio and 3D visual display, and audio immersive 3D visual display. The detailed explanation of each of the aforementioned variants is given below:

Variant 1: voice-only display The first version of the prototype was an audio only version. We used an existing AI and voice recognition system as the foundation for the functionality of the virtual assistant. The primary objective was to test the interaction modality, not the development of a better AI. The development of the first version was actually the development of a benchmark with which we compared our next versions that involved different output techniques. The audio only version acted similar to Siri or Google-Now where it took an audio input and gave an audio output. The user was able to perform a certain task, using Dragon NaturallySpeaking.

Variant 2: audio and 2D visual display The second version of the prototype included adding a 2D visual display to the audio only virtual assistant. At this stage, we gave the virtual assistant some visual identity using the 2D avatar developed by SitePal [27]. The user was able to interact with a virtual assistant that had some kind of physical manifestation. The 2D display of the virtual assistant was provided in the users laptop or computer screen.

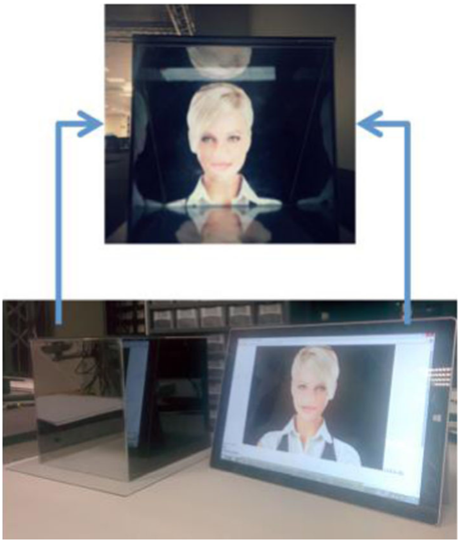
Variant 3: audio and 3D visual display The third version of the prototype brought about a paradigm shift from 2D visual display to a 3D visual display. At this stage we developed a visual representation of the virtual assistant and projected it using a 3D projection technique called peppers ghost [24]. A snapshot of the 3D virtual assistant for variant 3 is shown in Fig. 1.

Variant 4: audio and immersive 3D display The fourth version of the prototype further moved the virtual assistant from 3D (with peppers ghost) to an immersive 3D display with the Oculus Rift. The virtual assistant was projected to the users eyes through the Oculus Rift virtual reality glasses. Both the third and the fourth variants of the system enabled us to dig deeper into understanding the questions posed in our hypothesis by using different mediums of displays for the virtual assistant.

3.2 Overview of system implementation

Figure 2 shows the system implementation architecture overview. The following subsections describe representative pieces of the system implementation. The architecture consists of three overalls components: the input systems (Gmail, Calendar, Dragon NaturallySpeaking), the intermediate processing system (Processing, Web server, Web front-end, SitePal API, Microsoft Windows, Oculus HMD), and the output system (one of the four variants).

Fig. 1 3D virtual assistant for Variant 3 configuration



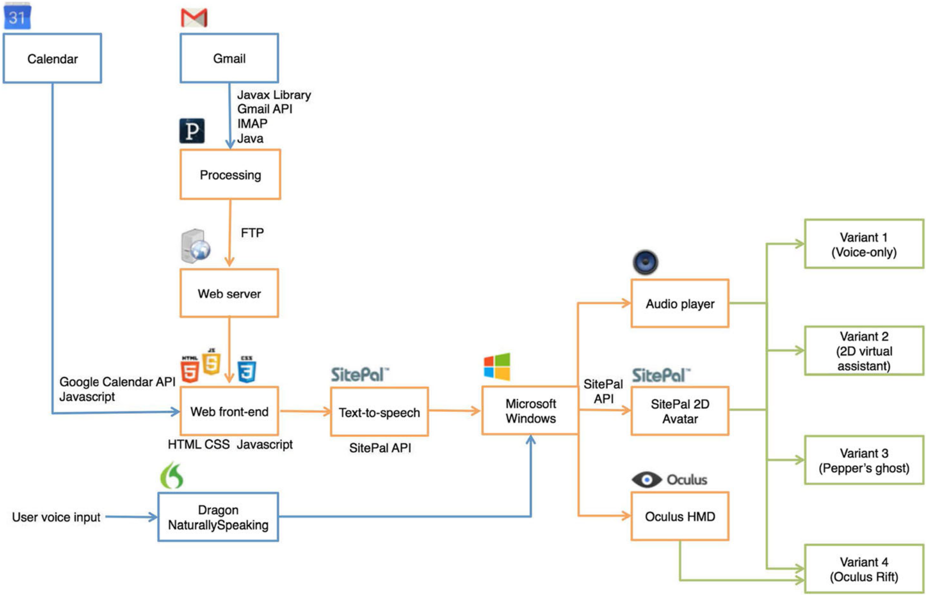


Fig. 2 Implementation architecture

3.2.1 Retrieving emails

After establishing a connection, the user's inbox is checked for unread messages, and unread messages are stored in a file named "emailsubjects.txt". The file is transferred from the local machine (that is running the code on Processing IDE) to the server where the rest of the VPA system is running on, using the FTP protocol. Now the updated version of the file "emailsubjects.txt" has been uploaded to the server, and is made available to the web server running the entire front-end of the VPA system, as well as part of the backend. This entire process is continually repeated in order to keep updated the list of unread messages in a user's inbox.

3.2.2 Reading emails

After writing the newest set of unread emails to the file "emailsubjects.txt" and uploading it to the server, the front-end system must read the emails. In order to do this, the file "emailsubjects.txt" is read and the VPA responds as appropriate to the contents of this file. The function sayText() from the SitePal API is utilized to instruct the virtual assistant to speaking out the text. Now the virtual assistant has speaks out loud the number of new emails if any. Next a loop continues to execute and accordingly reads out metadata, such as the subject and sender, of the user's new emails.

3.2.3 SitePal application programming interface (API) implementation

The SitePal API is used to output the avatar of the front-end of the virtual assistant via a built-in text-to-speech engine [27]. In synchronization with Javascript, the SitePal API drives the visual and audio output of the VPA to the browser window. A number of API function calls are used throughout the front-end system – such as sayText() and saySilent(), as demonstrated in the previous section.

3.2.4 Google calendar authorization

Similarly to reading a user's emails and notifying when a new email is received, it is important for the VPA system to have access to the user's calendar information. When there is an upcoming calendar event, the VPA system notifies the user. In order to obtain access to the user's calendar, it is necessary to obtain authorization from both Google and from the user.

3.2.5 Querying google calendar

Once the Google Calendar API is loaded and initialized, the VPA system is able to query for information from the user's calendar. The query is called through request.execute() function that defines what to do with the response. This is defined in function(resp). In the case of calendar events, the response iterates through upcoming events and uses the sayText() function from the SitePal API to speak out information about the user's events. For example, sayText(resp.items [15].summary,3,1,3); will read out the title, of the third calendar event in the response. Note that because the output of SitePal text-to-speech engine is queued, the function to read upcoming events would simply add the information of each calendar event to the queue for the text-to-speech engine to dictate.

4 Performance evaluation

Thirty (30) subjects, primarily students of age 18–23 years from New York University Abu Dhabi, participated in this study. Out of these subjects, 15 were male and 15 were female students. The purpose of the experiment was to determine the variation in the QoE across all 4 variants of the VPA. Since the experiment involved human subjects, both the facilitators and the supervisor of the project went through the Institutional Review Board (IRB) approval process.

4.1 Experimental procedure

Each subject was given a short entrance briefing on the nature and purpose of the experiment. A training session is given to every subject before the experiment session started. The training session included a demonstration of using each of the 4 variants, how to interact with both the input and output system of each variant, and recalibrate the input system with the subject's voice. The subject was then asked to sign a consent form and provided an entrance questionnaire. The four variants are presented to the subject in random order in order to minimize learnability effects. The steps of the usability testing are provided in Table 1.

The experiment began with each subject sending an email and setting up 4 different calendar event (occurring 10 min, 15 min, 20 min and 25 min respectively in the future). The four different variants were presented to the subject in random order. The VPA provided the notification for the first calendar event through the first variant presented to the subject. Similarly, each subsequent variant of the VPA provided the notification for the subsequent calendar event. Note that the calendar events were intentionally scheduled with 5 min difference since the user might not spend more than 3 min on each output variant.

While interacting with each variant, the subject receives a new email. This email is sent by one of the facilitators as soon as the subject is about to use the respective variant. It takes about

Table 1 Steps of usability testing

Step #	Explanation
1	Entrance briefing provided by the facilitator
2	Consent form signed by subject
3	Demonstration given by facilitator
4	Training audio system
5	Completing 4 variants session in random order
6	Completing exit questionnaire
7	Debriefing subject for further feedback

30 s before the email notification is provided by the text-to-voice system. While conducting usability testing, the false detections by the input system for each variant were also measured. Figure 3a shows a subject about to send an email with the input system of the virtual assistant. Figure 3b–f show the subject doing some of the other tasks during the experiment.

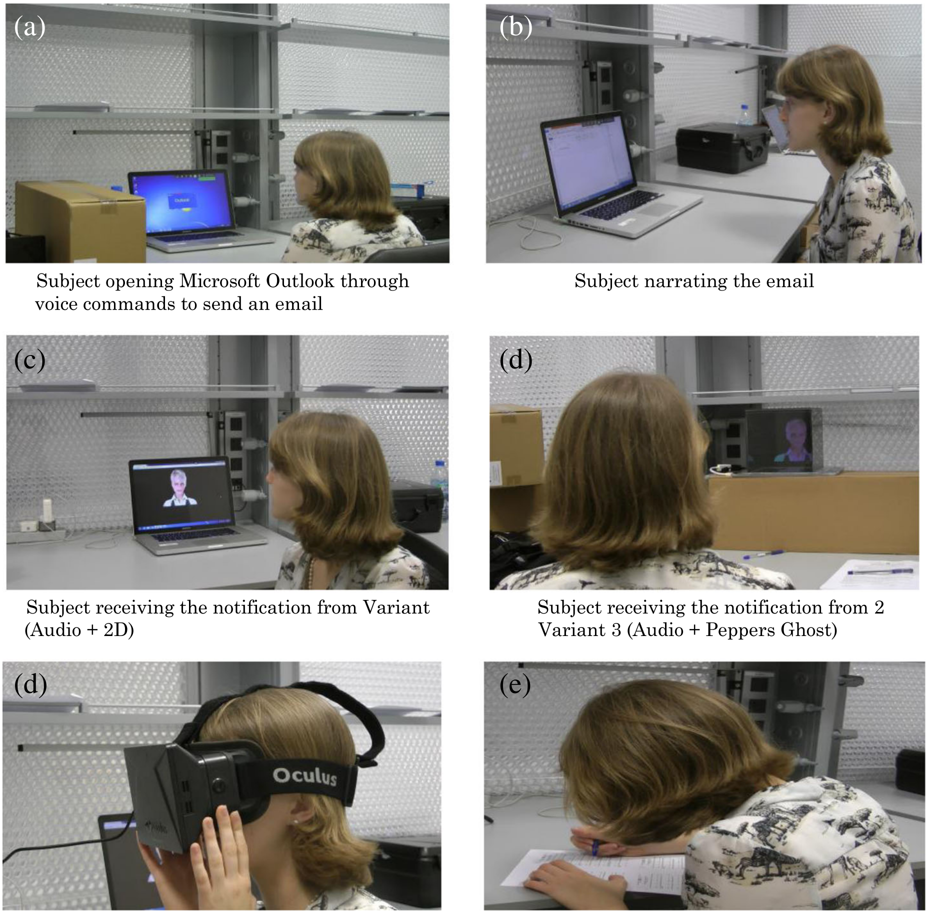
4.2 Exploratory factor analysis (EFA)

The results of the exit questionnaire from all the 30 subjects were the most essential component of our experiment. As mentioned earlier, each subject rated each of the 4 variants on a 5-point Likert scale (see Fig. 4). The quality of user experience can be classified into two parts: the level of user engagement and immersiveness [33]. The performance of each variant was split into 8 dependent parameters as shown in Table 2, based on the studies presented in [13, 33]. Note that these 8 dependent parameters are dependent variables whereas the interaction variant is the independent variable.

After collecting the raw data, the first step towards developing a scale to rate the QoE for each variant was to conduct the Exploratory Factor Analysis (EFA) [6]. The EFA analysis utilizes correlation metrics to study inter-correlations between dependent variables. The dimensionality of the correlation metrics can be reduced by looking for variables that correlate highly with a group of other variables, such variables with high inter-correlations could well measure one underlying variable – commonly known as the "factor". A freely available software package called FACTOR[1] is used to apply EFA analysis where recommendations made by Baglin [3] are adopted to improve EFA for ordinal data. Once the results from the EFA were obtained for each variant, it became evident what parameters (among the 8 parameters) accounted for greatest variation in the data (resulting in lowest dependability), and thus were eliminated for the subsequent analysis. The subsequent analysis involved using the significant parameters (i.e. remaining ones after eliminating less relevant ones) in order to develop the scale to rate the QoE. This scale was then used to rate the overall system.

EFA was conducted on the raw dataset to extract qualitative factors that are significant for each variant. Kaiser-Mayer-Olkin (KMO) and Bartlett's Test, Component Score Coefficient Matrix and Scree plot were three of the most significant tools that enabled us to determine the significant dependent variables for each of the four variants. It is important to mention that all three tools were used in a way that

[1] http://psico.fcep.urv.es/utilitats/factor/.



Subject opening Microsoft Outlook through voice commands to send an email

Subject narrating the email

Subject receiving the notification from Variant (Audio + 2D)

Subject receiving the notification from 2 Variant 3 (Audio + Peppers Ghost)

Subject receiving the notification from Variant 4 (Audio + Oculus)

Subject completing the post usability testing questionnaire

Fig. 3 Subject in usability testing experiment. **a** Subject opening Microsoft Outlook through voice commands to send an email. **b** Subject narrating the email. **c** Subject receiving the notification from Variant (Audio +2D). **d** Subject receiving the notification from 2 Variant 3 (Audio + Peppers Ghost). **e** Subject receiving the notification from Variant 4 (Audio + Oculus).**f** Subject completing the post usability testing questionnaire

they complemented each other. For instance, if KMO analysis concluded that four of the eight dependent variables are significant, Scree Plot analysis was conducted to validate the findings of KMO and vice versa. Once the results were validated, the significant factors were extracted and used to rate the QoE for each variant. This analysis was also important because we hypothesized that each variant will have different factors that will be significant.

○	○	○	○	○
Not at all	A little	Moderately	Very much	Extremely

Fig. 4 The 5-point Likert scale used to rate the four variants of the experiment

Table 2 8 dependent variables considered in the usability study

Dependent variable	Question #	Question
Presence	1	How strong was the sense of presence of the VPA?
Involvement	2	How involved did you feel while interacting with the VPA?
Attention	3	How focused were you wile getting the notifications?
Reliability	4	To what degree would you rely on the VPA for email and calendar notifications?
Dependency	5	To what degree would you depend on the VPA for email and calendar notifications?
Easiness	6	How easy was it to use the VPA?
Satisfaction	7	To what extent were you satisfied with the notification style and content?
Expectations	8	To what extent were your expectations met by the VPA?

4.2.1 Mean and standard deviation

As a first step during the EFA, basic statistical analysis was carried out for the usability testing results of all 4 variants. Figure 5 represents the mean and standard deviation for each dependent variable for each of the four variants respectively.

4.2.2 KMO and Bartlett's test

KMO and Bartlett's Test is an essential component of EFA. We used it to determine if our raw data is suitable for drawing conclusions by understanding the correlation between all the dependent variables. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy [6] is a statistic measure that indicates the proportion of variance in the variables that might be caused by the underlying dependent variables. High values (close to 1.0) generally indicate that a factor analysis may be useful with the raw data. If the value is less than 0.50, the results of the factor

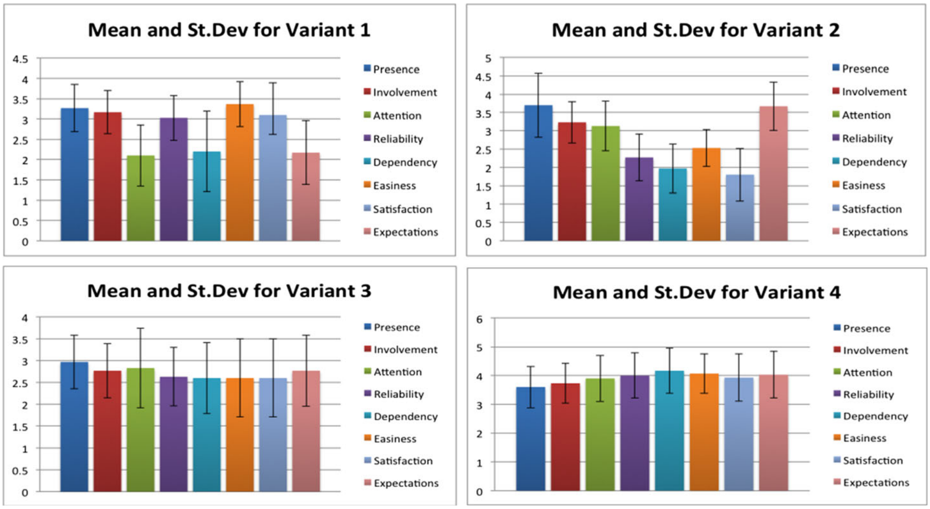


Fig. 5 Mean and Std. Deviation of the 30 subjects for each dependent variable for Variant 1, 2, 3, and 4

analysis probably is not very useful. Bartlett's test of sphericity tests the level of relationship between the dependent variables [4]. Small value (less than 0.05) of significance level indicates that factor analysis might be very useful with the raw data. The lower the significance value, the greater the level of relationship between the dependent variables.

KMO and Bartlett's Test for variant 1 and variant 2 reveal low values of Kaiser-Meyer-Olkin Measure of Sampling Adequacy while variant 3 and variant 4 reveal very high values (as shown in Table 3). This provides an important insight into the effectiveness of the selected dependent variables for the usability testing. Even before moving further into the analysis, we already know that some of the dependent variables for variant 1 and variant 2 are not significant towards the final rating of the QoE. Overall, Bartlett's test of sphericity revealed significance values that indicated high correlation between the dependent variables. The only exception to this was the significance value of variant 2. It is essential to mention that no single element of EFA can completely support or reject the suitability of raw data for EFA. In fact it is the combination of the results of multiple elements that enabled us to conclude whether the usability test results obtained are suitable for EFA or not.

4.2.3 Component score coefficient matrix

The rotated component matrix enabled us to determine what is represented by all the dependent variables. As mentioned earlier, this analysis was the second tool to determine which dependent variables to use for evaluating QoE. Table 4 (Left) represents the rotated component matrix of variant 1. It is evident that only the first 4 components were extracted because the remaining 4 components represented a lot of variability and it could not be concluded that they had any correlation with the first 4 components. For example, the first component is highly correlated with involvement, the second component is highly correlated with dependency, the third component is highly correlated with reliability, and the fourth component is highly correlated with satisfaction. Therefore, only 4 dependent variables scored high, namely involvement, dependency, reliability, and satisfaction. A similar approach is taken for studying variant 2 as shown in Table 4 (Right) where dependency, expectations, involvement, and reliability are the variables used to evaluate QoE. Note that the level of correlation between the same dependent variables for variant 3 and variant 4 was extremely high as suggested by the low significance values in Section 4.2.2. Hence, all the dependent variables for variant 3 and variant 4 will be used when rating the QoE.

Component score coefficient matrix of EFA analysis suggests that we can focus on involvement, dependency, reliability and the level of focus for rating the QoE while using variant 1. Using similar analysis, the first two columns of Table 5 indicate dependent variables

Table 3 KMO and Bartlett's Test values for all 4 variants

Variant name	Kaiser-Meyer-Olkin measure of sampling adequacy	Bartlett's test
Variant 1 (Audio only)	0.36	0.05
Variant 2 (Audio +2D Visual)	0.47	0.84
Variant 3 (Audio +3D Visual)	0.84	0.00
Variant 4 (Audio +3D Immersive Visual)	0.85	0.00

Table 4 Component score coefficient matrix for Variant 1 (left) and Variant 2 (right)

	Component					Component			
	1	2	3	4		1	2	3	4
Presence	0.399	−0.072	0.008	0.035	Presence	−0.154	0.528	0.005	0.028
Involvement	0.446	−1.49	−0.02	−0.144	Involvement	−0.065	−0.074	0.538	−0.157
Attention	−0.010	0.435	−0.208	0.435	Attention	−0.415	0.050	0.253	−0.180
Reliability	0.153	0.117	0.645	−0.105	Reliability	−0.081	0.058	−0.050	0.780
Dependency	−0.081	0.666	0.060	−0.128	Dependency	0.563	0.082	0.142	−0.312
Easiness	0.375	0.084	0.127	−0.175	Easiness	−0.563	−0.163	−0.607	−0.099
Satisfaction	0.008	−0.095	0.086	0.699	Satisfaction	0.317	−0.252	0.070	0.290
Expectations	0.128	0.186	−0.560	−0.226	Expectations	0.196	0.590	0.101	0.019

with high correlation for each variant respectively. The variables that are not listed in Table 5 for variant 1 and variant 2 were eliminated during the analysis for this element of EFA.

In order to support the findings of the component score coefficient matrix, the Scree Plot is utilized to augment the findings and visually represent them. It helps determine the optimal number of components from the raw data. Component coefficient matrix suggested extracting 4 components from both variant 1 and variant 2. Figure 6 (Left) represents the Scree Plot of variant 1. It shows that only 4 components have the eigenvalue of more than 1.0 hence those are the only 4 components that are correlated. All the other components will be eliminated due to high variability. A similarly Scree Plot for variant 2 is shown in Fig. 6 (Right). The Scree Plots of variant 3 and variant 4 were of no value since all the dependent variables are equally correlated.

4.2.4 Statistical analysis (ANOVA)

Analysis of variance (ANOVA) provides statistical test to determine whether there are any significant differences between the means of three or more independent (unrelated) groups [11]. Repeated one-way ANOVA tests were conducted to evaluate the contribution of each quality variable towards the QoE based on participants' responses to the questionnaire for the 4 variants. The hypothesis for each dependent variable was defined as follows: "would the dependent variable have a significant effect on the QoE". Results are shown in Table 6. For variance 1 (the audio-only variance), there was a significant effect of involvement, dependency, reliability, and satisfaction for rating QoE whereas the other 4 dependent variables showed

Table 5 QoE rating and significant dependent variables for each of the 4 variants

Variant	Considered dependent variables	QoE score
Variant 1 (Audio only)	Involvement, dependency, reliability, satisfaction	3.23
Variant 2 (Audio +2D Visual)	Dependency, expectations, involvement, reliability	3.43
Variant 3 (Audio +3D Visual)	All dependent variables	2.72
Variant 4 (Audio +3D Immersive Visual)	All dependent variables	3.93

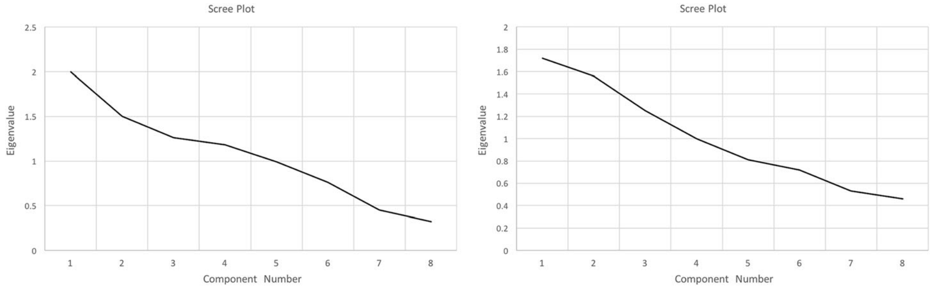


Fig. 6 Scree plot of Variant 1 and 2

no significant effect. Similar results are observed for variance 2. However, it is clear for variant 3 and variant 4 that all the dependent variables contribute significantly to rating the QoE. These results are in line with the findings in Section 4.2.3.

5 Results

5.1 QoE rating

Based on the findings of EFA, we were able to extract the significant dependent variables for each of the variant. In order to rate each variant, we took the average of the scores for all the significant variables. Based on this approach, the rating of each of the system is provided in Table 5.

5.2 Implicit analysis

In addition to EFA that was used to extract the relevant dependent variables to rate each of the variants in terms of QoE, there were several other implicit analysis that were conducted with the aid of raw data. Splitting the subjects into different groups and representing their results

Table 6 One-way ANOVA for dependent variables for each of the 4 variants

Dependent variable	Variant 1		Variant 2		Variant 3		Variant 4	
	F-value	*P*-value	F-value	*P*-value	F-value	*P*-value	F-value	*P*-value
Presence	0.17	=0.681	1.76	=0.601	81.66	<0.001	91.14	<0.001
Involvement	49.229	<0.001	44.29	<0.001	71.82	<0.001	71.24	<0.001
Attention	1.61	=0.508	2.27	=0.501	27.19	<0.001	29.33	<0.001
Reliability	25.61	<0.001	22.19	<0.001	18.33	<0.001	26.19	<0.001
Dependency	41.91	<0.001	41.39	<0.001	28.12	<0.001	22.12	<0.001
Easiness	3.91	=0.617	2.98	=0.707	22.18	<0.001	29.13	<0.001
Satisfaction	80.33	<0.001	8.33	=0.513	88.12	<0.001	98.77	<0.001
Expectations	0.11	=0.581	22.18	<0.001	27.91	<0.001	41.19	<0.001

through graphs enabled us to better comment on the more implicit factors that might have influenced the outcome of the usability testing.

The number of male and female participants was intentionally kept equal during the usability testing. After considering the responses of both male and female subjects, it was found that there was a slight difference in the average score assigned by both male and female to each of the VPA variants. Figure 7 shows the comparison of the average scores assigned by 15 male and 15 female participants to each of the variants. It is evident that for almost all of the variants, female subjects rated the variant slightly higher than the male subjects. Overall, there was a 32 % increase in the average assigned score from variant 1 to variant 4 for male subjects. On the other hand, there was a 43 % increase in the average score assigned by the female subjects. It is promising to see that both male and female subjects liked the idea of an immersive 3D VPA, however female subjects were slightly more receptive of the idea.

Figure 8 represents the variability in the scores assigned by each of the subjects to each of the 4 variants. It is evident that there was relatively less variability in the assigned average scores between each subject of variant 1 and variant 2. As soon as we start analyzing the scores assigned to variant 3 and variant 4, we realize a great deal of variability but this variability is almost evenly distributed throughout the participants. This demonstrated that ratings for variant 1 and variant 2 were comparable for all the participants whereas those ratings for variant 3 and variant 4 varied largely across participants. This is probably because participants had different opinions about 3D/immersive 3D technologies in general.

Figure 9 demonstrates the percentage of contribution of the eight dependent variables towards calculating the QoE for the 4 variants. For example, dependency has the largest contribution for variant 4 (Oculus 3D) whereas satisfaction contributed the least for variant 2.

Lastly, the results of the Gaussian distribution of the average score assigned to each of the 4 variants are shown in Fig. 10. It is evident, that variant 2 had the least variation in the response of subjects i.e. most of the subjects rated the QoE more or less similarly. On the other hand, variant 4 had most variability in the response from the subjects. Moreover, it is apparent that the average score (value corresponding to the peak of the distribution curve) of the first 3 variants was quite close to each other. However, as soon as we shift to variant 4, the average score is significantly higher, depicting that most of the subjects felt that immersive 3D environment has the highest potential to increase the QoE when interacting with a virtual assistant.

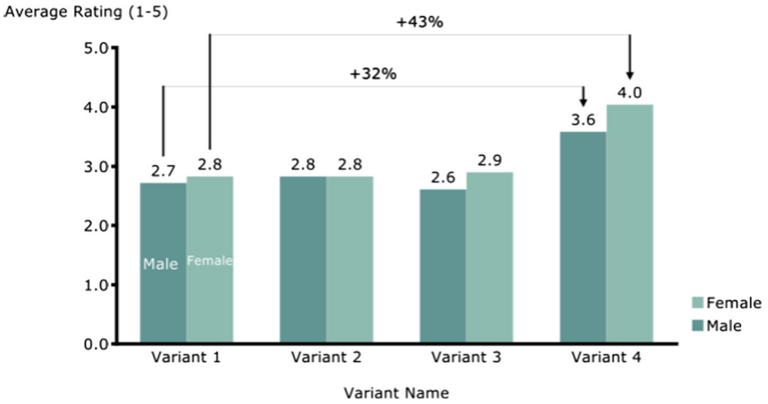


Fig. 7 Graph representing the comparison between the average score assigned by male and female subjects

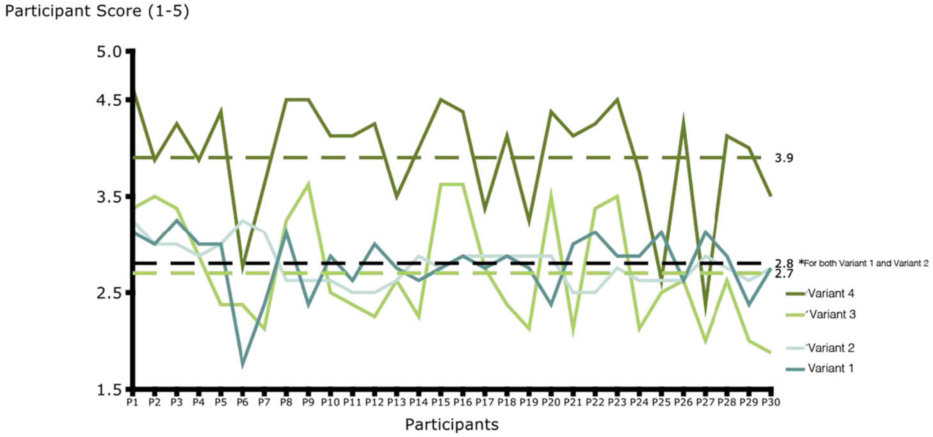


Fig. 8 Plot of the average score assigned by each subject to the respective variant

6 Discussion

We proposed 8 dependent variables for evaluating each variant of the VPA; sense of presence, extent of interactivity, extent of focused attention, degree of reliability, degree of dependency, extent of easiness, extent of satisfaction and extent of expectation being met. EFA analysis yielded only 4 relevant dependent variables for variant 1 and variant 2. For variant 3 and variant 4, a high degree of correlation was shown between all the factors and thus all the 8 dependent variables were considered.

It is also interesting to explore the interaction of each of the factors. According to Forlizzi and Battarbee [10] "emotions affect how we interact with products and the perceptions and outcomes that surround those interactions". The presence of motivation and affective elements in most of the chosen factors is not incomprehensible. The presence of emotions augments the user experience frameworks such as the one proposed by McCarthy and Wright [36], where emotions are embedded throughout the analysis of QoE.

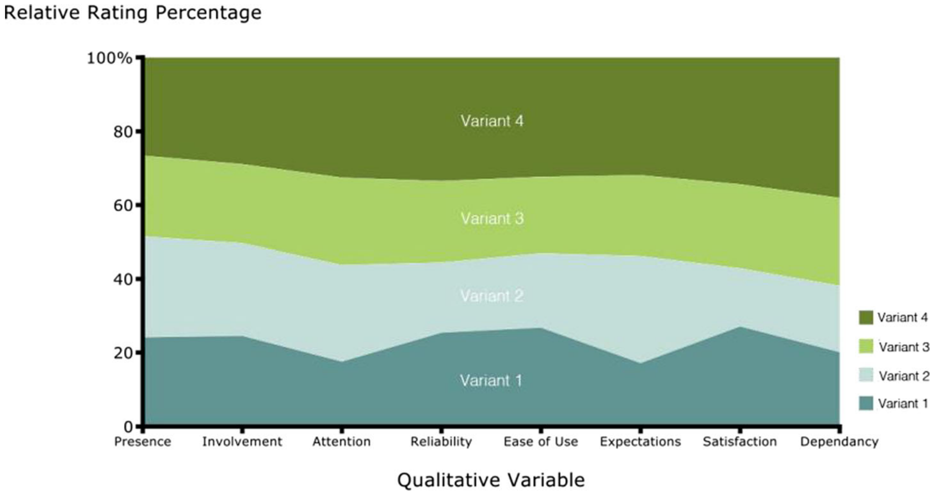


Fig. 9 Relative rating for all 4 variants for each of the 8 dependent variables

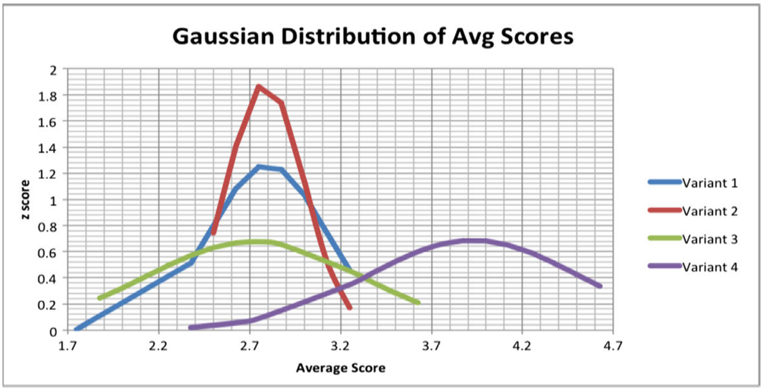


Fig. 10 Gaussian distribution of the average score assigned to each of the 4 variants

Factors like "felt involvement" and "felt attention" are directly related to the perceived utility of the system, hence to the reliability and the dependency on the system. The results of EFA and the overall average score for variant 4 also raise another important question whether personal interest, involvement and fun influenced users' appraisals of an experience. In other words, there are chances that the user actually rate variant 4 (variant with the Oculus display) higher than any other variant because they had fun using it? It was obviously a unique and novel experience for most of the subjects. They were extremely curious even before interacting with the variant. This question supports the related work that advocates for creating a positive user experience and considers emotion as an extremely essential component in system development [19, 22].

6.1 Learning outcomes from the experiment

Analysis of the usability testing data unveiled several interesting insights and met the objectives of this study. In references to the objectives presented in Section 1, here is a summary of our findings.

- Objective 1: it is clear that each variant has different set of dependent variables to contribute to QoE. For instance, variant 1 (audio only) ended up having involvement, dependency, reliability, and satisfaction as the most effective for measuring the QoE. Variants 3 and 4 required all the 8 parameters to measure QoE.
- Objective 2: quantifying and developing a model for rating QoE for each variant is successfully accomplished. Table 5 reported the QoE rating for each variant. Results demonstrated that users rated immersive personal virtual assistance (variant 4) with the most QoE value (QoE score of 3.93).
- Objective 3: the experiment demonstrated a clear relationship between efficiency/accuracy and the QoE. Results showed that there is a correlation between the number of false detections by the VPA input and the QoE. It was observed that on average there were 2–3 false detections. When the number of false detections for a

given subject increased beyond the average, we generally recalibrated the input system with the subject's voice. It was also observed that false detections were the primary cause of frustration amongst the subjects. They often expressed that the inefficiency of the input system was one of the primary reason for them not being permanently dependent upon the virtual assistant. It is important to mention that regardless of what dimension is taken towards developing future virtual assistants, efficiency and accuracy of voice detection should be the epicenter of the development phase. Eliminating background noise and accurate detections of various accents of a given language is extremely important.

An interesting outcome is that the interfacing system did affect the QoE. We observed very distinct results for all 4 variants of the VPA. This difference can be attributed to both, the perceived usability of the system and the actual usability of the system. It is important to explain this distinction. Perceived usability varied from one subject to the other. For instance, since most of the subjects had never seen a virtual assistant with a visual output system, their take on its usability and efficiency was highly influenced by the factors mentioned above. On the other hand, the actual usability of each variant was also different from the other. Despite having a similar input system and text-to-voice recognition system, the impact of each variant on the subject was very distinct primarily due to its unique nature of presence. For example, subjects felt that using the Oculus isolated them from the real environment and ensured that they focus on the avatar being projected inside. This developed a feeling of higher usability for variant 4 over variant 3 (pepper's ghost).

Finally, the excitement of virtual environment is an opportunity to developing exciting systems in the future. Since most of our subjects were students that live in a technology paradigm that is evolving at a fast pace, they are generally excited about the changes that the future will bring. This phenomenon was very obvious during our usability testing. As soon as the subjects were asked to use variant 4 (audio + Oculus), the extent of excitement on their faces was far more compared to when they were interacting with other variants. During the general feedback, most of the subjects, felt that the potential use of a virtual assistant would exponentially increase in virtual environments. Providing a 3D display for the virtual assistant might not be enough to directly enhance the user experience. As long as we are in a real environment, we still need to market and promote the use of a virtual assistant by comparing it to an actual human personal assistant. This burden of proof is a real challenge for virtual assistant developers.

6.1.1 Limitations

Although, a considerable amount of effort was invested to determine the structure and the protocol for the usability testing, we still believe there were some limitations of our work. On one hand, the chosen sample size was an absolute minimum size need to plot Gaussian distribution and test our initial hypothesis. Due to the variation in the background of the test subjects, we believe the data is still influenced by many external factors that were beyond our control. An improved version of the usability testing will involve testing the 4 variants with distinct groups of subjects.

7 Conclusion and future work

There are 3 major findings in this study. Firstly, there is a need for a scale dependent upon unique set of variables for each variant to rate its QoE. Secondly, immersive 3D visual output system emerged as the variant with highest QoE. Users felt that the need for a virtual assistant in a virtual environment increases dramatically. Thirdly, the study demonstrates a clear relationship between efficiency/accuracy and the rating of QoE. Lastly, there is need to perform the usability testing with virtual assistant variants that are able to perform a whole range of tasks - not only notifying about new emails and upcoming calendar events - to better understand the differences between their QoE rating. This paper also serves as a framework for conducting future research. The analysis tools provided in the paper can be used to assess even more dependent variables that will impact QoE. These new dependent factors can then be incorporated into the scale for an even better and accurate rating mechanism.

This work focused on the development of a scale to measure the QoE and used it to evaluate the performance for 4 variants of the VPA. The next steps include adding more functionality to the input/output system with more complex tasks, optimizing usability testing and testing with other visual feedback systems. Moreover, we also believe that our research is the beginning of further research towards developing future virtual assistants. Our insights especially provide promising data for developing virtual assistants for the virtual environments.

Acknowledgments Permission to make digital or hardcopies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credits permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

References

1. Alben L (1996) Quality of experience: defining the criteria for effective interaction design. Interactions 3(3): 11–15
2. Avery DR, McKay PF, Wilson DC, Volpone SD, Killham EA (2011) Does voice go flat? How tenure diminishes the impact of voice. Hum Resour Manag 50(1):147–158
3. Baglin J (2014) Improving your exploratory factor analysis for ordinal data: a demonstration using FACTOR. J Pract Assess Res Eval 19(5):2
4. Bartlett’s Test, https://en.wikipedia.org/wiki/Bartlett%27s_test. Accessed 12 Sept 2015
5. Branco G, Almeida L, Beires N, Gomes R (2006) Evaluation of a multimodal virtual personal assistant, 20th International Symposium on Human Factors in Telecommunication, HFT
6. Costello AB, Osborne JW (2005) Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. J Pract Assess Res Eval 10(7)
7. Ebrahimi T (2001) Quality of experience: a new look into quality and its impact in future personal communications
8. Fogg BJ, Tseng H (1999) The elements of computer credibility. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, p 80–87
9. Follmer S, Leithinger D, Olwal A, Hogge A, Ishii H (2013) inFORM: dynamic physical affordances and constraints through shape and object actuation. InProceedings of the 26th annual ACM symposium on User interface software and technology(UIST ‘13). ACM, New York, NY, USA, p 417–426

10. Forlizzi J, Battarbee K. Understanding experience in interactive systems. In Proceedings of the 5th conference on designing interactive systems: processes, practices, methods, and techniques. ACM, p 261–268
11. Freedman DA, Pisani R, Purves R (2007) Statistics, 4th edn. W.W. Norton & Company. ISBN 978–0–393-92972-0
12. Gebhardt S, Pick S, Oster T, Hentschel B, Kuhlen T (2014) An evaluation of a smart-phone-based menu system for immersive virtual environments. In IEEE Symposium on 3D User Interfaces 2014, 3DUI 2014 - Proceedings, p 31–34
13. Hamam A, El Saddik A, Alja'am J (2014) A quality of experience model for haptic virtual environments. ACM Trans Multimed Comput Commun Appl 10(3), Article 28
14. http://www.voxiebox.com/. Accessed 11 Sept 2015
15. Johnston M et al. (2014) MVA: the multimodal virtual assistant. In Proc. SIGDIAL Conference, p 257–259
16. Jung Y, Kuijper A, Fellner D, Kipp M, Miksatko J, Gratch J, Thalmann D (2011) Believable virtual characters in human-computer dialogs. Eurographics 2011 - State of The Art Report, p 75–100
17. Kazap Z, Magnenat-Thalmann N (2007) Intelligent virtual humans with autonomy and personality: State-of the-art. IOS Press, Intelligent Decision Technologies
18. Kumar D, Sachan A (2014) Bridging the gap between disabled people and new technology in interactive web application with the help of voice. 2014 International Conference on Advances in Engineering and Technology Research (ICAETR), p 1–5
19. Laurel B (2013) Computers as theatre. Addison-Wesley
20. Leithinger D, Ishii H (2010) Relief: a scalable actuated shape display. In Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction. ACM, p 221–222
21. Lodato J (2005) Advances in voice recognition: a first-hand look at the magic of voice-recognition technology. Futurist 39:1
22. Nahl D, Bilal D (2007) Information and emotion: the emergent affective paradigm in information behavior research and theory. Information Today, Inc
23. Nickell J (2005) Secrets of the sideshows. University Press of Kentucky, Kentucky
24. O'connell I, Rock J (2011) Projection apparatus and method for pepper's ghost illusion. US Patent 7(883): 212
25. Riccardi G (2014) Towards healthcare personal agents. In Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges (RFMIR '14). ACM, New York, NY, USA, p 53–56
26. Sabrowski JC (2013) Holographic technology at home
27. SitePal. SitePal. Oddcast Inc. http://www.sitepal.com/. Accessed 10 Sept 2015
28. Song Q, Shen H (2012) Intelligent voice assistant
29. Teixeiraa A, Hämäläinenb A, Avelarb J, Almeidaa N, Némethd G, Fegyód T, Zainkód C, Csapód T, Tóthd B, Oliveiraa A, Dias MS (2013) Speech-centric multimodal interaction for easy-to-access online services – a personal life assistant for the elderly. 5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, DSAI
30. Thalmann NM, Thalmann D (2012) Virtual humans: back to the future. p 1–8. Proceedings of the 2012 Graphics Interace Conference
31. Vinayagamoorthy V, Gillies M, Steed A (2006) Building expression into virtual characters. Eurographics 2006: State of the Art Report
32. Volonte M, Babu S, Chaturvedi H, Newsome N, Ebrahimi E, Roy T, Daily SB, Fasolino T (2016) Effects of virtual human appearance fidelity on emotion contagion in affective inter-personal simulations. IEEE Trans Vis Comput Graph (99):1–1
33. Weiss B, Wechsung I, Kühnel C, Möller S (2015) Evaluating embodied conversational agents in multimodal interfaces. Rev Comput Cogn Sci 1:6
34. Whalen TE, Noel S, Stewart J (2003) Measuring the human side of virtual reality. In Proceedings of the IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS'03), p 8–12
35. Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering: an introduction. Springer
36. Wright P, McCarthy J (2005) The value of the novel in designing for experience. In Future interaction design. Springer, p 9–30
37. Wu W, Arefin A, Rivas R, Nahrstedt K, Sheppard R, Yang Z (2009) Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. In Proceedings of the 17th ACM International Conference on Multimedia, p 481–490

38. Xu B, Yu Y (2010) A personalized assistant in 3D virtual shopping environment. 2010 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), p 266–269
39. Yagi A, Imura M, Kuroda Y, Oshiro O (2011) 360-degree fog projection interactive display. In SIGGRAPH Asia 2011 Emerging Technologies. ACM, p 19

Umair Saad is a graduate from the Electrical Engineering program at New York University Abu Dhabi. He is interested in virtual reality and multimodal interactions.

Usama Afzal is a graduate from the Computer Engineering program at New York University Abu Dhabi. He is interested in virtual reality and holographic display.

Ahmad El Issawi is currently an assistant professor at the Lebanese University, Lebanon. Dr. El Issawi has a PhD in Mathematics from France in the domain of "Math Discret-Théorie des Graphes-Relations Binaires" at Universite Claude Bernard - LYON I. His current research interests include multi-modal human computer interaction, graph theory, and ambient intelligence.

Mohamad Eid received the PhD in Electrical and Computer Engineering from the University of Ottawa, Canada, in 2010. He is currently an assistant professor of practice of electrical engineering in the engineering division at New York University Abu Dhabi (NYUAD). He was previously a teaching and research associate at the University of Ottawa from June 2008 until April 2012. He is the co-author of the book: "Haptics Technologies: Bringing Touch to Multimedia", Springers 2011, the co-chair of the 3rd International IEEE Workshop on Multimedia Services and Technologies for E-health (MUST-EH 2013), technical chair for the Haptic-Audio-Visual Environment and Gaming (HAVE) workshop in 2013. He is the recipient of the best paper award of DS-RT 2008 conference and the prestigious ACM Multimedia 2009 Grand Challenge Most Entertaining Award for "HugMe: Synchronous Haptic Teleconferencing" System. He has more than 70 conference and journal papers and 4 patents. His academic interests include Multimedia haptics, affective haptics, and tangible human computer interaction for assistive living. See my Google Scholar profile: Google Scholar Profile).