Voice Versus Keyboard and Mouse for Text Creation on Arabic User Interfaces

Khalid Majrashi
Department of Information Technology, Institute of Public Administration, Saudi Arabia

Abstract: *Voice User Interfaces (VUIs) are increasingly popular owing to improvements in automatic speech recognition. However, the understanding of user interaction with VUIs, particularly Arabic VUIs, remains limited. Hence, this research compared user performance, learnability, and satisfaction when using voice and keyboard-and-mouse input modalities for text creation on Arabic user interfaces. A Voice-enabled Email Interface (VEI) and a Traditional Email Interface (TEI) were developed. Forty participants attempted pre-prepared and self-generated message creation tasks using voice on the VEI, and the keyboard-and-mouse modal on the TEI. The results showed that participants were faster (by 1.76 to 2.67 minutes) in pre-prepared message creation using voice than using the keyboard and mouse. Participants were also faster (by 1.72 to 2.49 minutes) in self-generated message creation using voice than using the keyboard and mouse. Although the learning curves were more efficient with the VEI, more participants were satisfied with the TEI. With the VEI, participants reported problems, such as misrecognitions and misspellings, but were satisfied about the visibility of possible executable commands and about the overall accuracy of voice recognition.*

Keywords: *Voice, speech, recognition, input modal, user interface, user performance, Arabic, text entry, keyboard, mouse.*

Received March 23, 2020; accepted June 16, 2021
https://doi.org/10.34028/iajit/19/1/15

1. Introduction

Since the early stages of computing technology, researchers have worked on advancing Automatic Speech Recognition (ASR), aiming to allow using voice as an input modal for computers. Owing to improvements made in the recent decades in the areas of ASR and Natural Language Processing (NLP), the voice-based input modal has become a reality. Therefore, the development of Voice User Interfaces (VUIs), which are powered by ASR, has been expanding. Smart speakers or autonomous screen less voice gadgets, such as Alexa by Amazon, Google Home, Cortana by Microsoft, and HomePod by Apple are all examples of VUIs. In addition, many of the modern computing devices, such as smartphones, tables, laptops/PCs, smart TVs, and smartwatches, are voice-activated. A voice recognition feature is also embedded in several software applications, such as email and word processing.

Generally, problems with the technology, such as voice recognition errors, and usability issues, such as the invisibility of the limits and the capabilities of the system, remain major obstacles to widespread acceptance of specific types of VUIs. In relation to text entry, previous studies have provided no definite evidence regarding the usability of voice for text entry. In addition, many of these user studies on text creation using voice were conducted on old-generation VUIs (e.g., [15]), and studies on text entry using VUIs, powered by modern ASR, are scarce. Further, user studies on using voice for text entry on Arabic VUIs are still limited. That is, studies related to Arabic VUIs have focused more on the technological side (e.g., [1-3, 10, 11, 14, 19, 34, 38, 40]), with some studies addressed the human side but to a limited extent [4]. Therefore, we conducted this study to empirically investigate the usability of voice input for text entry on an Arabic VUI that is powered by a modern ASR.

We developed two Arabic user interfaces: a Voice-Enabled email Interface (VEI) and a Traditional Email Interface (TEI). We conducted an experiment to demonstrate the usability of using voice input compared with the traditional input devices (keyboard and mouse). In the experiment, voice input was used for completing tasks involving email-message creation, which involved using voice for executing commands and creating text on the VEI as against using the keyboard and mouse for completing the same tasks on the TEI.

The remainder of this paper proceeds as follows. First, we present a review of the related literature. Next, we describe the methods used in this study, including the demographics of our participants, the experimental interfaces, tasks, apparatus and measures, experimental design, and the investigation procedure. Then, we present and discuss our results and link the findings with specific implications for the design of Arabic VUIs for text creation.

2. Literature Review

Many terms are used to refer to technology that allows people to interact by using their voice, including VUI, speech user interface, conversational agent, and intelligent or virtual personal assistant [7, 9, 13, 15, 18, 25, 26, 27, 29, 30, 31, 32, 39]. In this study, we used VUI to emphasize our focus on spoken word interactions. The term VUI is used to refer to a screenless device or a software that is powered by ASR, when the interaction is completely, primarily, or partially voice based.

A popular research theme relating to human-computer interaction and VUI is the comparison of different modalities (e.g., keyboard, mouse, gesture, and digital pen) with the speech modal [6]. Studies on modality comparison have yielded inconsistent results and have reported negative and positive effects of using speech input on usability [5, 6, 15, 16, 21, 22, 23, 28]. For example, Karat *et al*. [15] investigated the user performance and satisfaction when completing text creation tasks on three old-generation speech recognition systems. They also tested the user performance on similar text creation tasks using the keyboard-and-mouse modality. They found that users are generally slower in creating text when using the speech modal than when using the keyboard-and-mouse modal. However, Murata and Takahashi [28] observed certain benefits on text entry efficiency when using speech compared with the keyboard among the elderly, particularly those who were not accustomed to keyboards. The present study extends and re-assesses prior studies on text entry by investigating the user performance and perceptions related to a modern VUI powered by an advanced ASR, in an organization context.

The type of tasks, that is, transcription (simple text entry) and composition (crafting a text), can influence the user performance when using the voice input modal for text creation. Karat *et al.* [15] found that users took a longer time on composition tasks than on transcription tasks when using speech and keyboard modalities, because composition tasks require a higher cognitive load. Shneiderman [35] argued that users can find it difficult to speak and think at the same time, particularly when crafting a message. He outlined that "cognitive resources for problem solving and recall are limited when speech input/output shares the short-term and working memory." Danis *et al.* [8] discussed the differences in writing by using speech as opposed to a keyboard and stated:

"Thought for many people is very closely linked to language. In keyboarding, users can continue to hone their words while their fingers output an earlier version. In dictation, users may experience more interference between outputting their initial thought and elaborating on it."

For dealing with this human limitation, Shneiderman [35] recommended that users dictate the full, or a part, of the text, and then review or proofread it. In our research, the effect of task type on user performance is taken into consideration. We investigated the usability of the voice modal on an Arabic VUI with tasks that required different levels of cognitive load.

The learnability of VUIs has also been investigated, and studies have targeted various aspects of learnability. An early study examined error correction strategies over time on an old VUI and found that user performance improved with experience [16]. Corbett and Weber [7] aimed to improve the discoverability and learnability of voice commands of a mobile VUI application. Their research identified challenges with voice interactions and explored methods (e.g., learn "as-you-go") for improving the learning experience. Another study investigated ways to improve learnability in VUIs through adaptive discovery tools [12]. In this study, we extended the literature on learnability of VUIs by comparing performance during initial and extended use of an Arabic VUI.

Early studies on VUIs identified limitations and challenges that hinder users from adopting, or interacting with, VUIs efficiently and effectively. Shneiderman [35] argued, "Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks." Karat *et al*. [15] also found that subjects felt unproductive when using VUIs owing to several reasons, including speech recognition errors, command language problems, difficulties in error correction, and difficulty in talking and thinking simultaneously. Karl *et al.* [17] also found that users find it difficult to memorize commands when interacting with a VUI and have some concerns relating to recognition errors, the interference of background noise, poor feedback, and slow response time. Some recent studies have considered obstacles and difficulties associated with VUIs. For example, Myers *et al.* [29] studied the main obstacle types and the tactics that users employ to overcome these when interacting with VUIs. They found that NLP error was the obstacle most encountered by users. Other identified obstacles include unfamiliar intents, failed feedback, and system error. They also highlighted that users relied more on "guessing" and "exploration" to overcome obstacles rather than on using "visual aids" or "recalling knowledge" [29]. Other recent studies have also reported the difficulty of using and learning VUIs because of the insufficient visibility of the VUI (e.g., its limits and capabilities) and NLP errors [7, 12, 20, 24, 33, 36]. Hence, in this study, we investigated the positive and negative experience of users with a VUI, to complement the user performance results in our study.

3. Method

3.1. Participants

Forty employees (aged 25-49 years) were recruited from a major educational institution located in Riyadh, Saudi Arabia. Half of the participants were female. Participants who had a Bachelor, Master, and PhD as their highest degree were 30, five, and five respectively. Their study major ranged from information technology to public administration and business. Most of the participants were Saudi. All participants were native Arabic-language speakers. Table 1 shows the participants' demographics.

Table 1. Participants' demographics.

	Item	Frequency	Percentage
Age	25-29	9	22.5
	30-34	12	30
	35-39	9	22.5
	40-44	5	12.5
	45-49	5	12.5
Gender	Male	20	50
	Female	20	50
Education Level	Bachelor	30	75
	Master	5	12.5
	PhD	5	12.5
Study Major	Information Technology	15	37.5
	Public Administration	13	32.5
	Business	12	30
Nationality	Saudi	38	95
	Non-Saudi	2	5
Native Language	Arabic	40	100
	Other	0	0

All participants were staff who use email applications daily. About 55% of the participants had used a VUI previously. However, none of them had interacted previously with email applications using their voice. Participants had different levels of typing skills, and in this study, 10 of them were classified as fast, 15 as moderate, and 15 as slow based on their ability to type more than 40 words per minute, 30-40 words per minute, and less than 30 words per minute, respectively.

3.2. Experimental Interfaces

We developed two interactive new-email message interfaces in Arabic: VEI (Figure 1) and TEI (Figure 2). The interfaces were designed with the right-to-left direction because Arabic is a language that is read from right to left. The VEI uses Google's Speech API that supports speech recognition for Arabic.

To create voice commands that satisfy user needs when creating an email message, we analyzed a collection of formal email messages. We requested 12 staff working in the research context (in an organization in Saudi Arabia) to each share 10 recent formal email messages they had sent, thus receiving a total of 120 messages. We informally analyzed these messages to identify their basic elements and structure. Then, we created voice commands based on the analysis results. Therefore, the voice commands are linked to the needs of users for creating formal email messages in our research context. For example, we created commands for inserting the opening, greeting, and closing content automatically, which is similar to those in the formal messages. The commands were displayed in a panel in the right side of the VEI (Figure 1).

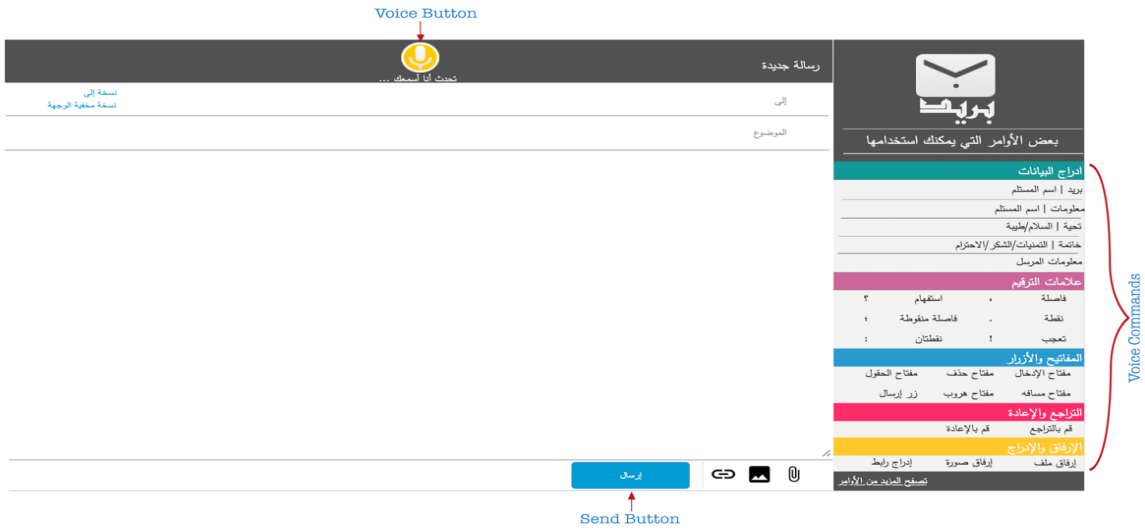

Figure 1. The voice-enabled email interface (VEI).

Some examples of the created voice commands are as follows:

- Email ID of (the name of the recipient/individual): to insert the email address of a specific recipient.
- The information of (the name of the recipient): to insert the recipient's title, name, and job title.
- The greeting of (type of greetings): to insert the preferred greeting.

- The closing of (the type of closing): to insert the preferred closing information.
- The sender's information: to insert the required information of the sender.
- The Send button: to send the message.

We also created commands for punctuation marks (e.g., comma and question mark), keyboard keys (e.g., enter and space) and other functions, such as undo and redo. Participants were also able to browse more commands than those presented in the command panel, using a link at the end of the panel.

Figure 2. The traditional email interface (TEI).

3.3. Tasks

The participants were asked to complete tasks involving email-message creation on each interface: a message that the researcher had prepared and a self-generated message. Therefore, they had to complete a total of four message creation tasks: a pre-prepared task on the VEI, a pre-prepared task on the TEI, a self-generated task on the VEI, and a self-generated task on the TEI. For the self-generated task, the participants were simply asked to craft a formal email message to be sent to one of their colleagues. For the pre-prepared task, users were asked to send the following email message:

To: Alahmadya@test.edu.sa

Subject: A proposal for developing the work at the PMO

Message:

His Excellency Dr. Ahmed bin Sami Alahmdy, Director of Innovation and Business Development Department,

Peace and Allah's mercy and blessings be upon you.

The attachment is a proposal for the development of services and processes in the Project Management Office. The proposal includes benchmarking of the PMO services and processes against those of other local and international businesses considered the best in the industry, a list of opportunities for improving the PMO in our organization, and an action plan for change.

Therefore, we would be grateful if you would kindly review the proposal and let us know your comments, if any, before we proceed to the next step.

Best regards,
Mohammed bin Ahmed Al-Aqeel
Development Advisor

3.4. Apparatus and Measures

For the experiment, we used a desktop computer (EliteOne 800) with the following specifications: Windows 8.1 Enterprise edition, processor: Intel® Core™ i7-4770S, CPU 3.10GHz, 16 GB RAM, 64-bit operating system, resolution: 1920×1080×60 hertz, screen size: 23.8. The participants were requested to use the desktop's built-in microphone and an HP Bilingual Arabic and English Wireless Keyboard and HP wireless mouse.

We used a set of measures including user performance (task completion and execution time), learnability, and satisfaction. The measures were used to assess the usability of using voice input compared with the traditional input devices (keyboard and mouse) for text creation on Arabic user interfaces.

Task completion refers to whether the task was successfully completed by the user. The prototype system automatically recorded the task completion once the participants clicked the "send" button using the voice command with the VEI and the mouse or the keyboard with the TEI. In addition, the experimenter took notes about task completion while observing the user to ensure that the participant completed the tasks successfully according to our predefined completion criteria.

Task execution time means the time taken to achieve the task. The prototype system automatically recorded the time taken from the moment the "new-message" interface was displayed until the participants clicked the "send" button using the voice command with the VEI and the mouse or the keyboard with the TEI.

In our research context, learnability refers to the extent to which the user performance improves with the experience. It was measured by comparing execution time on the first use of the system with that

on repeated uses to determine whether the user performance has improved on repeated uses.

Satisfaction refers to the user's satisfaction about using the voice on the VEI and the keyboard and mouse on the TEI. The satisfaction was measured using a 7-point Likert scale.

We also analyzed the video recordings of the user interactions with the VEI. This is to determine the voice recognition accuracy for the pre-prepared message task and the self-generated message task. We calculated the percentage of words that the prototype system recognized correctly.

3.5. Experimental Design

The experiment was focused on comparing user performance when creating email messages using voice on the VEI and the keyboard and mouse on the TEI. Therefore, the independent variable is the combination of the type of the user interface and the input modality. The dependent variables are primarily the task completion and execution time.

We used a within-subject design for our experiment. To ensure that the order in which the interfaces (VEI and TEI) were used and their associated input modalities, and the order of tasks (pre-prepared message and self-generated message), did not affect the experiment results, the participants were divided into four groups (A, B, C, and D) that each had five males and five females. Each group performed the tasks on the VEI and the TEI in a different order (see Table 2). The tasks were controlled, where the pre-prepared message task given to all participants to be sent using the VEI and the TEI was identical. The participants were also asked to use the same self-generated message on both the interfaces. Each participant performed each task three times to allow the learnability to be measured. Therefore, the experiment had 40 participants×4 tasks×3 operations, resulting in 480 task completions. The participants received the tasks instructions in Arabic.

Table 2. Experimental design.

Group	Task 1	Task 2	Task 3	Task 4
A	VEI (Pre-prepared)	TEI (Pre-prepared)	VEI (Self-generated)	TEI (Self-generated)
B	VEI (Self-generated)	TEI (Self-generated)	VEI (Pre-prepared)	TEI (Pre-prepared)
C	TEI (Pre-prepared)	VEI (Pre-prepared)	TEI (Self-generated)	VEI (Self-generated)
D	TEI (Self-generated)	VEI (Self-generated)	TEI (Pre-prepared)	VEI (Pre-prepared)

3.6. Procedures

At the beginning of the session, the participants received information about the research project and were asked if they wanted to participate. Then, if they agreed to participate, they signed a written consent form. Next, they completed a questionnaire on their background information, internet experience, previous experience with VUIs, and frequency of using email programs at the workplace. Then, we informed them that they would be required to create pre-prepared and self-generated messages using a VEI and a TEI. They were informed that the goal was to investigate how quickly they could create a message using each interface. The participants were instructed to use voice as a primary input method when using the VEI. However, we did not restrict them to using voice only to complete the tasks on the VEI. We informed them that in some cases, where they felt they could not achieve specific goals using voice, they could use the keyboard and the mouse (e.g., for moving the cursor, selecting words, or correcting errors). They were told to activate the voice input feature in the VEI by pressing a voice button at the top of the interface, so that they could start feeling that they were speaking to the system.

At the beginning of the experiment, each participant was given five minutes to explore the two interfaces. They then started completing the tasks. When each task started, the participant was shown the details regarding the task on the top of the screen. The pre-prepared message was given to the subjects on a sheet of paper. The participant was required to press a button to start the task. Upon pressing the button, either one of the interfaces was displayed. The task ended once the participant activated the send button using the voice command, or the mouse, or the keyboard.

For the self-generated message task, participants were asked to create a message when interacting with the first interface they had to use and to remember the message because they needed to create the same message when interacting with the second interface. This approach was preferred to auto-recording the message initially generated for the first interface and then displaying it to the participants when they attempted the same task on the second interface. In the approach that we used, the participants had to still rely on think-typing (a composition task) when interacting with the second interface, whereas in the alternative approach, they would have relied on copy-typing (a transcription task), which would have affected the measurements for the self-generated task.

At the end of each task, participants completed a brief questionnaire on their experience regarding completing each task. Once they had completed all the tasks, they filled another brief questionnaire about their experience with the interfaces and the input modalities, and their suggestions on how to improve the VEI. The experimenter took notes about the problems they faced when interacting with the VEI during the sessions.

4. Results and Discussion

4.1. Overall Voice Recognition Accuracy

The average accuracy of voice recognition by the VEI was 90% for the pre-prepared message task and 85% for the self-generated task. This difference could be because when completing the former, the participants had to use a fixed number of words, whereas when completing the self-generated task, they generated their own words, leading to a larger number of expressions, and in turn, increasing the probability of the system being unable to recognize more words.

4.2. User Performance and Learnability

All participants were able to complete their tasks using the VEI and the TEI. In the first operation, the participants' average execution time was 4.23 minutes when creating the pre-prepared message using the VEI and 5.99 minutes when using the TEI. We conducted a paired t-test to compare the task execution time for the pre-prepared message on both interfaces. The results of this test showed a significant difference in the execution time between the VEI and the TEI, with $t(39) = -7.187$ and $p<0.001$. For the self-generated message tasks, participants recorded faster times with the VEI (average: 2.91 minutes) than with the TEI (average: 4.63 minutes). The paired t-test results indicated a significant difference in the execution time between the two interfaces for the self-generated message, with $t(39) = -9.815$ and $p<0.001$.

In the second operation, participants had an average execution time of 3.68 minutes when sending the pre-prepared message using the VEI and 5.90 minutes when using the TEI. The results of the paired t-test showed a significant difference in the execution time between the VEI and TEI, with $t(39)=-8.736$ and $p< 0.001$. For the self-generated message tasks, participants recorded faster times with the VEI (average: 2.52 minutes) than with the TEI (average: 4.55 minutes). The results of the paired t-test indicated a significant difference in the execution time between the two interfaces for the self-generated message, with $t(39)= -11.323$ and $p<0.001$.

In the third operation, participants had an average execution time equal to 3.06 minutes when creating the message using the VEI and 5.73 minutes when using the TEI. The results of the paired t-test showed a significant difference in the execution time between the VEI and TEI, with $t(39)= -10.178$ and $p<0.001$. For the self-generated message, participants recorded faster times with the VEI (average: 1.85 minutes) than with the TEI (average: 4.34 minutes). The result of the paired t-test indicated a significant difference in the execution time between the two interfaces, with $t(39)= -13.458$ and $p<0.001$.

As an overall, participants were faster in message creation using voice than using the keyboard and mouse. For the pre-prepared message, the differences in execution time between the two interfaces were 1.76 minutes (first operation), 2.22 minutes (second operation), and 2.67 minutes (third operation). For the self-generated message, the differences in execution time were 1.72 minutes (first operation), 2.03 minutes (second operation), and 2.49 minutes (third operation).

We analyzed the learning curves for the interfaces based on the initial and the repeated operations. Figures 3 and 4 present the learning curves of the two interfaces for the pre-prepared and self-generated message tasks respectively. It can be observed from Figure 3 that the execution time shows a declining trend from the first to the second and third operations for all interfaces. The variance in execution time between the first and third operation is −1.18 minutes for the VEI and −0.26 minutes for the TEI. The difference in variance indicated that compared with the VEI, the TEI had a lower declining trend by 0.92 minutes.

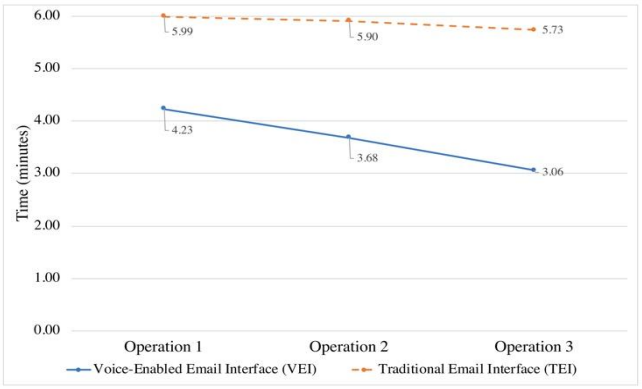

Figure 3. Learning curve of participants for VEI and TEI for pre-prepared message task.

Similarly, Figure 4 shows a declining trend from the first to the second and third operations for the interfaces. The variance in the execution time between the first and third operation is −1.06 minutes for the VEI and −0.29 minutes for the TEI. The difference in variance indicated that compared with the VEI, the TEI had a lower declining trend by 0.77 minutes.

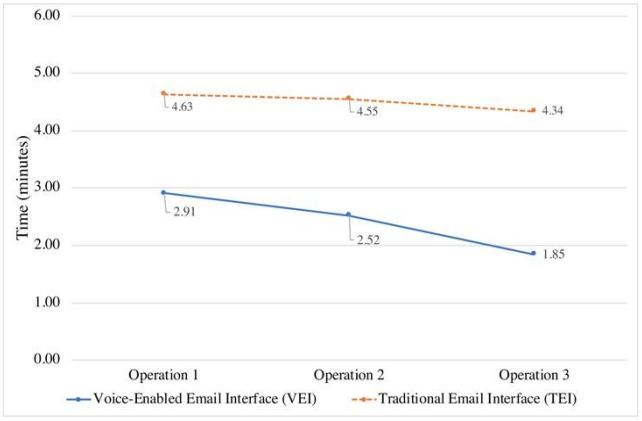

Figure 4. Learning curve of participants for VEI and TEI for self-generated message task.

We conducted repeated measure analysis of variance tests, with a significance level of 0.05, to compare the execution time on each interface across the three operations. For each interface, the results showed a significant difference in the execution time between the three operations, demonstrating that the participants performed significantly faster in the second and third operation compared with the first operation and faster in the third operation compared with the second operation; see Table 3. These results indicate that the participants' performance improved significantly on repeated use of the two interfaces.

We also conducted two-way repeated measure analysis of variance tests, with a significance level of 0.05, to compare the means of the execution time between the VEI and the TEI for each task, across the three operations. The results showed that there was an interaction effect of the interface type and the operation level on the execution time for the pre-prepared message task, with $F(1, 39)=99.765$ and $p<0.001$, and for the self-generated message task, with $F(1, 39)=95.488$ and $p<0.001$. Since the results showed that the interface type interacted significantly with the operation level, we conducted an analysis of simple effects and simple comparisons. For the pre-prepared task, we found a significant difference between the VEI and the TEI in the first operation, with $F(1, 78)=33.916$ and $p<0.001$, in the second operation, with $F(1, 78)=51.627$ and $p<0.001$, and in the third operation, with $F(1, 78)=71.069$ and $p<0.001$, reflecting that the average execution time for the VEI was lower than that for the TEI; see Table 4 for the pairwise comparisons.

Table 3. Results of the one-way repeated measure ANOVA tests for comparing execution time.

				Pairwise Comparisons		
Task	**Interface**	***F***	***P***	**Operation**		***P***
Pre-prepared	VEI	200.370	0.000	1	2	0.000
				1	3	0.000
				2	3	0.000
	TEI	245.370	0.000	1	2	0.000
				1	3	0.000
				2	3	0.000
Self-generated	VEI	190.148	0.000	1	2	0.000
				1	3	0.000
				2	3	0.000
	TEI	150.184	0.000	1	2	0.000
				1	3	0.000
				2	3	0.000

For the self-generated task, there was a significant difference between the VEI and the TEI in the first operation, with $F(1, 78)=30.613$ and $p<0.001$, in the second operation, with $F(1, 78)=41.753$ and $p<0.001$, and in the third operation, with $F(1, 78)=62.849$ and $p<0.001$; see Table 5 for the pairwise comparisons.

These results reflect that the VEI had considerably lower average execution time than the TEI and thus demonstrate that the level of performance improvement across the three menu operations is dependent on the interface type. Therefore, our interpretation of the results is that the participants had more efficient learning curves on the VEI than on the TEI.

4.3. Questionnaire and Observation Data

The participants were asked to rate their satisfaction regarding using voice on the VEI and the keyboard and mouse on the TEI on a 7-point Likert scale ranging from "1: very dissatisfied" to "7: very satisfied." In all, 36 subjects responded they were "satisfied" or "slightly satisfied" using the TEI and four responded with "neutral". As for the VEI, 28 responded "satisfied" or "slightly satisfied", 10 responded "dissatisfied" or "slightly dissatisfied", and two, "neutral." They also answered a similar question about productivity-34 and 30 participants felt "very productive" or "productive" on using the TEI and the VEI, respectively. Thus, more of them felt satisfied and productive with the TEI than with the VEI. These results are not consistent with the findings on the execution time, given that they were faster in message creation on the VEI than on the TEI. This inconsistency could be because users generally are more accustomed to the keyboard and mouse, as well as because of the problems (see Table 6) that these participants faced when interacting with the VEI.

Table 7 shows the users' positive comments about the VEI, and Table 8 presents a list of improvements they suggested for the VEI.

Table 4. Results of the pairwise comparisons for pre-prepared task.

						95% Confidence Interval for Difference	
Operation	**(I) Interface**	**(J) Interface**	**Mean difference (I − J)**	***Std. Error***	***Sig.*** [a]	**Lower Bound**	**Upper Bound**
1	VEI	TEI	−105.336*	18.087	0.000	−141.354	−69.327
	TEI	VEI	105.336*	18.087	0.000	69.327	141.354
2	VEI	TEI	−133.074*	18.521	0.000	−169.964	−96.203
	TEI	VEI	133.074*	18.521	0.000	96.203	169.964
3	VEI	TEI	−160.390*	19.022	0.000	−198.260	−122.521
	TEI	VEI	160.390*	19.022	0.000	122.521	198.260

[a] Adjustment for multiple comparisons Least Significant Difference (equivalent to no adjustments). *The mean difference is significant at the.05 level.

Table 5. Results of the pairwise comparisons for self-generated task.

						95% Confidence Interval for Difference	
Operation	**(I) Interface**	**(J) Interface**	**Mean difference (I − J)**	***Std. Error***	***Sig.*** [a]	**Lower Bound**	**Upper Bound**
1	VEI	TEI	−103.432*	18.694	0.000	−140.649	−66.215
	TEI	VEI	103.432*	18.694	0.000	66.215	140.649
2	VEI	TEI	−121.732*	18.839	0.000	−159.238	−84.226
	TEI	VEI	121.732*	18.839	0.000	84.226	159.238
3	VEI	TEI	−149.431*	18.849	0.000	−186.957	−111.905
	TEI	VEI	149.431*	18.849	0.000	111.905	186.957

[a] Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments). *The mean difference is significant at the.05 level.

Table 6. Problems Encountered by Participants in VEI Interactions.

Problem	Number of participants
Misrecognition of a single spoken word or a series of spoken words	40
Misspelling of Arabic words	40
Greater difficulty in editing text using voice than in using keyboard & mouse	33
Insertion of an utterance intended as command as words in the text instead	12
Problems with command language, meaning that the word used for the commands does not match the user-spoken words when trying to execute the commands	10
Error correction using voice leading to more errors	4
Talking interfering with thinking and remembering sometimes	4
Misrecognition of an utterance intended as dictation as a command	3
Attempt to execute a command that is not supported	3
Lack of visual or audio feedback when trying to execute unsupported commands	2
No scope for adding or customizing commands and actions in the VUI	2

Table 7. Positive comments about the VEI.

Comment	Number of participants
Good visibility of possible voice commands on the interface, which helped me to learn the voice commands that can be used.	30
The accuracy of the voice recognition is high.	29
Good to have multiple input modals for text creation. The voice can be used as the primary input modal and keyboard-and-mouse modal to help with review and edit.	28
Presenting commands as categories in the right side of the interface and using different colors for each category heading helped me to remember the positions of the commands and read them quickly when I needed to use a specific command but could not remember the exact word(s) for it.	10
Good visual and audio feedback when pressing the voice button to activate the voice input.	5

Table 8. Improvements proposed by participants for the VEI.

Comment	Number of participants
Resolve the Arabic spelling mistakes.	20
A confirmation message should be shown after the send button is activated using the voice command. This is to avoid cases where an utterance intended as dictation is understood by the system as the send command. Users should be able to choose not to show the confirmation message when they become familiar with the VUI.	7
Users should be able to display and hide the command panel.	6
Provide a single optional voice command to insert the recipient's email ID and his or her information (recipient's title, name, and job title), a default preferred greeting, a default preferred closing, and the sender's information.	7
Allow users to add and customize commands and to place newly added commands in the command panel.	2
Allow users to rearrange the order of commands in the command panel.	2

We discuss our qualitative data within three themes: ASR quality, input modals, and error correction; voice commands; and visual display design.

- **ASR quality, input modals, and error correction.**

Twenty-nine participants reported that the accuracy of voice recognition is high. However, misrecognition of spoken words occurred several times. Table 6 shows that the misrecognitions of a single spoken word or a series of spoken words was the problem most encountered by all participants. ASR also generated incorrect spellings for the Arabic words, particularly words that are commonly misspelled by native Arabic speakers, such as using "ha" (ـه) instead of "tāʾ marbūṭa" (ـة), or using "hamzat waṣl" (ٱ) instead of "hamzat qaṭ" (أ or إ). The misspelling of Arabic words occurred for all participants; however, only half of them recognized the misspelled words. They spent considerable time in correcting misrecognized and misspelled words. Therefore, further advances in Arabic ASR are needed to improve recognition accuracy and spelling quality.

The collected data showed that the participants were able to correct several errors using voice, but they mostly used the mouse for positioning and selecting the word(s) and then dictated the word(s) again. Further, 33 participants reported that they found it more difficult to edit text using voice than using the keyboard-and-mouse modal. Our observations showed that the correction of errors using voice failed in many cases, particularly for misspelled words, and the participants switched to using the mouse and keyboard for correction. This result confirms the user behavior related to modality switching reported in certain related early studies (e.g., Karat *et al*. [15], Karat *et al*. [16]). Other studies have also revealed that multimodal correction is preferred by VUI users [6, 37]. In addition, some participants in the present study reported that talking sometimes interferes with thinking and remembering. This finding matches an early argument by Shneiderman [35] mentioned previously in this paper and highlights the need for a keyboard and a mouse when crafting text that requires a high cognitive load. Overall, the keyboard-and-mouse modal continues to be important for text entry on VUIs because these devices increase the efficiency and effectiveness of refinement and can be utilized in tasks requiring high cognitive loads.

- **Voice commands.**

Our qualitative data showed that sometimes utterances intended as commands were inserted in the text. This was primarily due to the misrecognition or misspelling of words, which led to a mismatch between the spoken words and the command. In addition, although the

command panels display the words that can be used for executing the commands, some participants used other words for executing specific commands, resulting in their spoken words being inserted in the text rather than being used to execute the commands. This finding means that the user expectations about the words to be used for voice commands do not match the provided commands. Therefore, an investigation of the user preference of command words should be undertaken before designing a VUI. In addition, designers should use an appropriate method in the VUI design to enable users to learn command languages easily.

The results also showed that sometimes utterances intended as dictation are recognized as commands. This finding indicates that the words used for commands should be chosen carefully, meaning that words or phrases that are rarely used in text writing should be selected. For example, the command to send the email can be "click send button" or "send button" instead of "send," and similarly, "insert comma" can be used instead of "comma." Users also wished to be able to add and customize commands and actions, as well as to link multiple actions to a single command. Therefore, designers should bear in mind such needs when designing VUIs for text entry.

- **Visual display design.**

The participants had positive and negative comments on the visual display design (or the graphical user interface) and they suggested some changes to improve the design. In relation to the command panel, they were satisfied that it was provided. They also found the presentation of commands in separate categories and the use of different colors for each category heading helpful in remembering the positions of the commands and in locating these quickly when they needed to check the command words. They also suggested that they should be able to rearrange commands in the panel and to hide and display the panel as required. Hence, designers should provide users the ability to modify the command panel based on their needs or preferences.

The participants expressed dissatisfaction about the lack of visual and audio feedback when trying to execute unsupported commands. Hence, VUI designers should take this finding into consideration. The participants also suggested that confirmation messages should be provided to prevent possible critical unwanted actions (e.g., sending the message before completing it) that can be caused by an unintended execution of commands, such as when utterances intended as dictation are recognized as commands. They also recommended that the design should allow them to permanently hide the confirmation message after they become familiar with the interface.

5. Conclusions

We conducted an experiment to measure the usability of using voice input modality for text entry as compared with using the keyboard and mouse. Participants in this experiment attempted text creation tasks on a VEI using voice and on a TEI using the keyboard and mouse. The analysis results of this study show that, overall, those using voice completed tasks faster than those using the keyboard-and-mouse modal. Further, the participants learned to use the VEI more efficiently than they did the TEI. Nevertheless, a larger number of participants reported that they were satisfied with the TEI than with the VEI. They reported experiencing several problems with the VEI, such as the misrecognition of single or multiple words, the misspelling of Arabic words, and the difficulty of editing text using voice.

However, they commented positively about the VEI features, such as the visibility of possible executable commands. We linked our findings to implications for designing Arabic VUIs that allow more efficiency in text creation. Moreover, many of the findings and implications, which are general and not linked to the Arabic language or culture, can be applied to non-Arabic VUIs as well. Overall, the experiment results provide evidence that although the voice input is useful for text creation on a VEI, the keyboard and mouse should be provided as complementary input devices.

The average accuracy of voice recognition by the VEI was 90% for the pre-prepared message task and 85% for the self-generated task. This difference could be because when completing the former, the participants had to use a fixed number of words, whereas when completing the self-generated task, they generated their own words, leading to a larger number of expressions, and in turn, increasing the probability of the system being unable to recognize more words.

The tested VEI is commonly used in businesses and our research was conducted with employees in an organization setting. Therefore, the results of our study can encourage the adoption of the voice recognition technology in workplaces and embedding voice recognition feature in business applications.

Our study is limited to 40 participants; hence, future studies can extend it by analyzing data on a larger number of users. In addition, Arabic has several spoken forms across countries. Hence, further studies should also take into consideration the users of different Arabic speaking varieties. Moreover, the VEI tested in the present study does not represent all applications that might use voice for text entry, and therefore, its specific findings might not be generalizable. Hence, further studies are required on other VUIs that use voice for text entry, to complement this study's findings.

References

[1] Al-Anzi F. and AbuZeina D., "The Effect of Diacritization on Arabic Speech Recogntion," *in Proceedings IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies*, Aqaba, pp. 1-5, 2017.

[2] Alsharhan E. and Ramsay A., "Investigating The Effects of Gender, Dialect, and Training Size on The Performance of Arabic Speech Recognition," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 975-998, 2020.

[3] Alsharhan E., Ramsay A., and Ahmed H., "Evaluating the Effect of Using Different Transcription Schemes in Building A Speech Recognition System for Arabic," *International Journal of Speech Technology*, 2020.

[4] Amrouche A., Falek L., and Teffahi H., "Design and Implementation of a Diacritic Arabic Text-To-Speech System," *The International Arab Journal of Information Technology*, vol. 14, no. 4, pp. 488-494, 2017.

[5] Begany G., Sa N., and Yuan X., "Factors Affecting User Perception of A Spoken Language Vs. Textual Search Interface: A Content Analysis," *Interacting with Computers*, vol. 28, no. 2, pp. 170-180, 2015.

[6] Clark L., Doyle P., Garaialde D., Gilmartin E., Schlögl S., Edlund J., Aylett M., Cabral J., Munteanu C., Edwards J., and Cowan B., "The State of Speech in HCI: Trends, Themes and Challenges," *in Interacting with Computers*, vol. 31, no. 3, pp. 349-371, 2019.

[7] Corbett E. and Weber A., "What Can I Say?: Addressing User Experience Challenges of A Mobile Voice User Interface for Accessibility," *in Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, Florence, pp. 72-82, 2016.

[8] Danis C., Comerford L., Janke E., Davies K., De Vries J., Bertrand A., "Storywriter: A Speech Oriented Editor," *in Proceedings of the Conference Companion on Human Factors in Computing Systems*, Massachusetts, pp. 277-278, 1994.

[9] De Barcelos Silva A., Gomes M., Da Costaa C., da Rosa Righi R., Barbosa J., Pessin G., De Doncker G., and Federizzi G., "Intelligent Personal Assistants: A Systematic Literature Review," *Expert Systems with Applications*, vol. 147, 2020.

[10] Elmahdy M., Gruhn R., Minker W., Abdennadher S., "Cross-lingual Acoustic Modeling for Dialectal Arabic Speech Recognition," *in Proceedings of 8th Annual Conference of the International Speech Communication Association*, Makuhari, pp. 873-876, 2010.

[11] Elmahdy M., Gruhn R., and Minker W., *Novel Techniques for Dialectal Arabic Speech Recognition*, Springer Science and Business Media, 2012.

[12] Furqan A., Myers C., and Zhu J., "Learnability through Adaptive Discovery Tools in Voice User Interfaces," *in Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*, Colorado, pp. 1617-1623, 2017.

[13] Gardner-Bonneau D. and Blanchard H., *Human Factors and Voice Interactive Systems*, Springer Science and Business Media, 2007.

[14] Hassine M., Boussaid L., and Massaoud H., "Tunisian Dialect Recognition Based on Hybrid Techniques," *The International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 58-65, 2018.

[15] Karat C., Halverson C., Horn D., and Karat J., "Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems," *in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Pennsylvania, pp. 568-575, 1999.

[16] Karat J., Horn D., Halverson C., and Karat C., "Overcoming Unusability: Developing Efficient Strategies in Speech Recognition Systems," *in Proceedings of CHI'00 Extended Abstracts on Human factors in Computing Systems*, The Hague, pp. 141-142, 2000.

[17] Karl L., Pettey M., and Shneiderman B., "Speech Versus Mouse Commands for Word Processing: An Empirical Evaluation," *International Journal of Man-Machine Studies*, vol. 39, no. 4, pp. 667-687, 1993.

[18] Kepuska V. and Bohouta G., "Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa And Google Home)," *in Proceedings of IEEE 8th Annual Computing and Communication Workshop and Conference*, Las Vegas, pp. 99-103, 2018.

[19] Kirchhoff K. and Vergyri D., "Cross-Dialectal Data Sharing for Acoustic Modeling in Arabic Speech Recognition," *Speech Communication*, vol. 46, no. 1, pp. 37-51, 2005.

[20] Kirschthaler P., Porcheron M., and Fischer J., "What Can I Say? Effects of Discoverability in Vuis on Task Performance and User Experience," *in Proceedings of the 2nd Conference on Conversational User Interfaces*, Bilbao, pp. 1-9, 2020.

[21] Le Bigot L., Jamet E., Rouet J., and Amiela V., "Mode and Modal Transfer Effects on Performance and Discourse Organization with an Information Retrieval Dialogue System in Natural Language," *Computers in Human Behavior*, vol. 22, no. 3, pp. 467-500, 2006.

[22] Le Bigot L., Terrier P., Amiel V., Poulain G., Jamet É., Rouet J., "Effect of Modality on Collaboration with A Dialogue System," *International Journal of Human-Computer Studies*, vol. 65, no. 12, pp. 983-991, 2007.

[23] Limerick H., Moore J., and Coyle D., "Empirical Evidence for A Diminished Sense of Agency in Speech Interfaces," *in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul Republic of Korea, pp. 3967-3970, 2015.

[24] Luger E. and Sellen A., "Like Having A Really Bad PA: The Gulf Between User Expectation and Experience of Conversational Agents," *in Proceedings of the CHI Conference on Human Factors in Computing Systems*, San Jose, pp. 5286-5297, 2016.

[25] Maguire M., "Development of A Heuristic Evaluation Tool for Voice User Interfaces," *in Proceedings of International Conference on Human-Computer Interaction*, Orlando, pp. 212-225, 2019.

[26] McTear M., Callejas Z., and Griol D., *The Conversational Interface*, Springer International Publishing, 2016.

[27] Murad C., Munteanu C., Cowan B., and Clark L., "Revolution or Evolution? Speech Interaction and HCI Design Guidelines," *IEEE Pervasive Computing*, vol. 18, no. 2, pp. 33-45, 2019.

[28] Murata A. and Takahashi Y., "Does speech Input System Lead to Improved Performance for elderly? Discussion of Problems When Using Speech Interfaces for Elderly," *in Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, Yasmine Hammamet, pp. 108-113, 2002.

[29] Myers C., Furqan A., Nebolsky J., Caro K., and Zhu J., "Patterns for How Users Overcome Obstacles in Voice User Interfaces," *in Proceedings of the CHI Conference on Human Factors in Computing Systems*, Montreal, pp. 1-7, 2018.

[30] Nowacki C., Gordeeva A., and Lizé A., "Improving the Usability of Voice User Interfaces: A New Set of Ergonomic Criteria," *in Proceedings of International Conference on Human-Computer Interaction*, Denmark, pp. 117-133, 2020.

[31] Porcheron M., Fischer J., Reeves S., and Sharples S., "Voice Interfaces in Everyday Life," *in Proceedings of the CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada, pp. 1-12, 2018.

[32] Rheu M., Shin J., Peng W., and Huh-Yoo J., "Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design," *International Journal of Human-Computer Interaction*, vol. 37, no. 1, pp. 81-96, 2021.

[33] Sa N. and Yuan X., "Examining User Perception and Usage of Voice Search," *Data and Information Management*, vol. 5, no. 1, pp. 40-47, 2021.

[34] Satori H., Harti M., and Chenfour N., "Introduction to Arabic Speech Recognition Using CMUSphinx System," *arXiv preprint arXiv:0704.2083*, 2007.

[35] Shneiderman B., "The Limits of Speech Recognition," *Communications of the ACM*, vol. 43, no. 9, pp. 63-65, 2000.

[36] Srinivasan A., Dontcheva M., Adar E., and Walker S., "Discovering Natural Language Commands in Multimodal Interfaces," *in Proceedings of the 24th International Conference on Intelligent User Interfaces*, California, pp. 661-672, 2019.

[37] Suhm B., Myers B., and Waibel A., "Multimodal Error Correction for Speech User Interfaces," *ACM Transactions on Computer-Human Interaction*, vol. 8, no. 1, pp. 60-98, 2001.

[38] Vergyri D. and Kirchhoff K., "Automatic Diacritization of Arabic for acoustic Modeling in Speech Recognition," *in Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, Geneva Switzerland, pp. 66-73, 2004.

[39] Xu Y., Branham S., Deng X., Collins P., and Warschauer M., "Are Current Voice Interfaces Designed to Support Children's Language Development?," *in Proceedings of CHI Conference on Human Factors in Computing Systems*, Yokohama, pp. 1-12, 2021.

[40] Zaidi B., Boudraa M., Selouani S., and Yakoub M., "Control Interface of an Automatic Continuous Speech Recognition System in Standard Arabic Language," *in Proceedings of SAI Intelligent Systems Conference*, London, pp. 295-303, 2020.

Khalid Majrashi is an Associate Professor of computer science at the Department of Information Technology, Institute of Public Administration, Saudi Arabia. He received his Master's and Ph.D. degrees in Computer Science from RMIT University, Melbourne, Australia. His research interests include Human-Computer Interaction, and Human-centered AI.