User Expectations from Dictation on Mobile Devices

Santosh Basapur[1], Shuang Xu[1], Mark Ahlenius[1], and Young Seok Lee[2]

[1] Human Interaction Research Center of Excellence,
Motorola Labs, Schaumburg, IL 60196, USA
{sbasapur,shuangxu,mark.ahlenius}@motorola.com
[2] The Grado Department of Industrial and Systems Engineering,
Virginia Tech, Blacksburg, Virginia 24061 USA
yolee10@vt.edu

Abstract. Mobile phones, with their increasing processing power and memory, are enabling a diversity of tasks. The traditional text entry method using keypad is falling short in numerous ways. Some solutions to this problem include: QWERTY keypads on phone, external keypads, virtual keypads on table tops (Seimens at CeBIT '05) and last but not the least, automatic speech recognition (ASR) technology. Speech recognition allows for dictation which facilitates text input via voice. Despite the progress, ASR systems still do not perform satisfactorily in mobile environments. This is mainly due to the complexity of capturing large vocabulary spoken by diverse speakers in various acoustic conditions. Therefore, dictation has its advantages but also comes with its own set of usability problems. The objective of this research is to uncover the various uses and benefits of using dictation on a mobile phone. This study focused on the users' needs, expectations, and their concerns regarding the new input medium. Focus groups were conducted to investigate and discuss current data entry methods, potential use and usefulness of dictation feature, users' reaction to errors from ASR during dictation, and possible error correction methods. Our findings indicate a strong requirement for dictation. All participants perceived dictation to be very useful, as long as it is easily accessible and usable. Potential applications for dictation were found in two distinct areas namely communication and personal use.

1 Introduction

Mobile phones, with their increasing processing power and memory, are enabling a diversity of tasks which demand extensive user interaction. Text input using a small keypad becomes one of the most critical usability issues for mobile interaction design. For instance, the number of text messages sent has continuously increased around the world since its introduction in the late nineties. North American users sent out approximately 12.5 billion text messages in the month of June 2006 [2]. In UK all records were shattered when 4.4 billion messages were sent in the month of December 2006 [14]. It has been shown that teenagers prefer to send text messages to their friends rather than calling them [7].

To solve the problem of text input, research has focused on many methods of interaction which include Speech, Gestures, Handwriting, and Camera-Based such as

J. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2007, LNCS 4551, pp. 217–225, 2007.

Face Tracking [24]. Some promising solutions currently in the market include: QWERTY keypads on phone, external keypads, virtual keypads on table tops (Seimens at CeBIT '05) and last but not the least, automatic speech recognition (ASR) technology. 'Dictation' using ASR allows users to dictate text to a mobile phone. Users' voice input is converted into text via a speech recognition engine, embedded in the cell phones. Despite their progress, ASR systems do not perform satisfactorily in mobile environments. The complexity of capturing large vocabulary spoken by diverse speakers in various acoustic conditions is the major issue that hinders the successful deployment of ASR.

This research intended to uncover the various uses that a customer could benefit from by using dictation on a mobile phone. This paper focuses on the users' present day requirements and expectations from dictation as a text input medium. Another goal of this study was to seek the users' reactions to common errors found in dictation and also their preferences for error correction techniques. Lastly, design guidelines need to be developed for user-centric applications that optimally use dictation as an input medium.

2 Related Work

Text input in mobile devices has been a known usability problem for some time now. Some of the solutions available in today's handheld device market are described here. Palm utilizes Graffiti [18] that requires users to learn and memorize the predefined letter strokes. Also, it requires a touch screen for the text entry. Similarly, Motorola's WisdomPen [12] supports natural handwriting recognition of Chinese and Japanese characters. Another solution is the Thumbwheel [11]. Thumbwheel is used to scroll and highlight a character in a list of characters shown on a display. The select key inputs the high-lighted character. However, Thumbwheel has been shown to be too slow for text input on a cell phone [23]. Multi-tap and predictive methods of text entry have been proposed as an alterative approach to the mobile text input problem. There have been various estimations of text typing speeds: 14.9~20.8 words per minute (wpm) has been expected for multi-tap and 17.6~40.6 wpm has been expected for predictive text on average [5, 21]. These speed rates are good but they pale when compared to the average speed of typing on a QWERTY keyboard which is 80 wpm [1]. Therefore the devices which advertise high productivity (e.g., RIM's BlackBerry series, Motorola's Q series and the Samsung's BlackJack series phones) tout the QWERTY keypads. But with the limited space available on the mobile devices, the keys remain small thus limiting the text entry speed. All the above issues indicate an inefficient human computer interaction. All these hard-to-use solutions contribute to a dismal user experience which inhibits adoption of technology.

Compared to the above typing speeds, the normal speech of humans which reaches 125-150 wpm [6] makes a compelling argument for high efficiency and utility if used for text entry. Speech based user interfaces can provide an interaction experience that is similar to human-to-human communication. The interaction is more natural and efficient for transfer of information between the human and the machine [10]. Also, manual text entry methods are not conducive to hands-free and eyes-free usage. They

demand high cognitive attention by keeping the user involved with the UI both manually as well as perceptually. Hence, the speech recognition systems have been proposed as a better alternative to existing text entry methods [24].

Rudnicky et al in their survey of speech technologies [19] have indicated a wide variety of applications for which ASR has been utilized. From voice-based personal computer control [13] to large vocabulary dictation systems on personal computers such as Dragon Naturally Speaking [4] and from automated call routing to customer service applications such as automated flight status inquiry, there are a number of implementations of speech recognition and speech synthesis technologies [19]. So far, the implementation of ASR in the mobile phone industry (Motorola, Nokia, Samsung phones etc…) has been limited so far to simple voice command features such as name dialing, number dialing and application shortcuts. This is mainly due to the low processing power and memory of mobile phones. With the advancement of technology, ASR is expected to perform better and support features such as text input via voice. Therefore, dictation using ASR is the next frontier for text entry in the mobile phones. Samsung A900 model phone has already brought to market a limited vocabulary dictation feature which enables word-by-word text input.

Despite the progress in ASR technologies and their applications, the complexity of mobile environments hinders the performance in terms of speech recognition accuracy and error handling. Also there are the traditional ASR system problems such as turn-taking (between human and machine) and errors that occur with the opening and closing of the speech recognition window. The challenge for dictation based text input is to capture large vocabulary spoken by diverse speakers in various acoustic conditions. Without an extensive vocabulary it is hard to reach satisfactory recognition accuracy. Since speech-based text input is error prone, three kinds of interactions are needed with users to complete tasks: dictating, navigating to errors and error correction. It has been reported that users spend only 25-33% of the time on dictation of a document and the rest is spent in error detection, navigation to error and error correction [9, 20]. Research by Munteanu et al, [15] shows that state-of-the-art ASR systems trained by an individual can achieve about 3% of Word Error Rate (WER). However, as acoustic conditions degrade, WER can increase to 40-45%. Therefore it is imperative that the dictation feature should also have a robust error correction method that is easy and efficient.

Prior research thus shows us the advantages and problems associated with ASR. It has been shown that users have different expectations and needs based on their previous interaction experience with ASR based systems. Novice users need good instructions to make a task easier, while the experts want efficiency in task completion [8]. Despite the abundant studies about the ASR technology itself, little has been known or explored about users' needs and expectations. For examples, for what applications will the ASR technology be useful? why? How tolerable are dictation errors? Will users tolerate errors at all? How will they prefer to correct recognition errors on mobile devices? This research is therefore motivated to explore and find users' needs, expectations, preferences, and potential concerns from different population segments.

3 Methodology

Four (4) focus groups were conducted in this study, followed by a data analysis workshop. The two independent variables in this study are: (1) users' experience with SMS; and (2) users' experience with ASR. Participants were screened carefully for a controlled balance, which resulted in the four group classifications as shown in Table 1.

Table 1. Focus group classifications

		Text Messaging (SMS) experience	
		Yes	**No**
Automatic Speech Recognition (ASR/VR) Experience	**Yes**	Focus Group 1	Focus Group 3
	No	Focus Group 4	Focus Group 4

A dictation based SMS application was prototyped and used to gather data (see Figure 1). This prototype allows the user to dictate a text message. It simulates users' interaction experience with dictation on a cell phone. During the study, participants were shown demonstration of two tasks and later were asked to discuss the ideas with respect to their own needs and lifestyles. The experiment moderator followed a pre-defined script to ensure that the discussion was carried out consistent questions across the four groups. Each group discussion lasted for about two (2) hours. All data was video taped for later review.

Fig. 1. Screenshot of the prototype

Participants. A total of twenty (20) participants were recruited for this study, five (5) in each focus group. All participants were randomly selected by a marketing research company based on pre-defined criteria. Among the nineteen (19) people who attended the focus groups, there were 11 males and 8 females with a mean age of 35.1 (SD = 10.1). They were all cell phone users and were homogeneous regarding their experience with SMS and ASR within each group. Their cell phone usage varied from minimal use (calling and phonebook only) to extensive use (calling, phonebook, instant messaging, emails, SMS, voice dialing, etc.). Their experience levels were collected via a set of questions during the recruitment screening. Remuneration was paid to each participant for their time.

Task Scenarios. Two tasks were demonstrated to the users using the low-fidelity prototype of a Motorola RAZR phone as described in Table 2. The first scenario was sending of a text message to John Doe and no ASR errors were introduced in this scenario. The second scenario was again sending of a text message to John Doe but now two pre-defined ASR errors occurred back to back in this scenario. Both demos were canned interaction experiences with the moderator performing the tasks in front of the participant group.

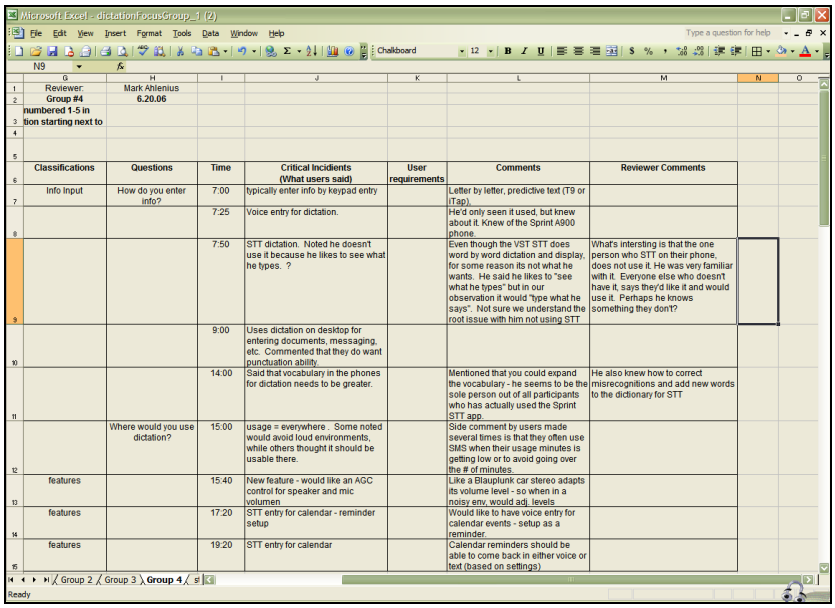
Table 2. Use Cases used in focus groups

Use Case 1: Dictation feature with no errors 1) Moderator presses a voice command key 2) Phone speaks: "Say a command" 3) Moderator speaks: "Send a message to John Doe." 4) Phone speaks: "Say the message" 5) Moderator speaks: "I will be there in 15 minutes." 6) Phone displays texts and speaks: "I will be there in fifteen minutes."
Use Case 2: Dictation feature with errors 1 - 4) Same as the use case 1 5) Moderator speaks: "Did he sign the agreement?" 6) Phone displays texts and speaks: "Disney signed agreements" 7) Moderator presses 'back soft key' 8) Phone speaks: "Repeat the message" 9) Moderator speaks: "Did he sign the agreement?" 10) Phone displays texts and speaks: "Did she sign the agreement"

Participants were encouraged to discuss their initial reactions to the demonstrated task experiences. They were also asked to discuss the usefulness of mobile dictation in relation to their lifestyles. The second demonstration was split into two stages to gather participants' reaction to ASR errors. First an error was shown in the dictated text and then was followed-up by an error correction attempt. Users' reactions to errors were discussed along with different error correction options.

4 Data Analysis

The video-taped sessions were reviewed and analyzed by each of the four researchers individually. Each member identified critical incidents, transcribed the relevant verbal data into text, and summarized it using a template as shown in Figure 2. Upon the completion of the individual review, an affinity diagram was completed by the four researchers to cluster the qualitative findings. Labels were assigned to each cluster that represented a common perspective, followed by development of summary statements for each cluster.



	G	H	I	J	K	L	M	N	O
1	Reviewer:	Mark Ahlenius							
2	**Group #4**	**6.20.06**							
3	**numbered 1-5 in**								
4	**tion starting next to**								
5									
6	**Classifications**	**Questions**	**Time**	**Critical Incidients (What users said)**	**User requirements**	**Comments**	**Reviewer Comments**		
7	Info Input	How do you enter info?	7:00	typically enter info by keypad entry		Letter by letter, predictive text (T9 or iTap).			
8			7:25	Voice entry for dictation.		He'd only seen it used, but knew about it. Knew of the Sprint A900 phone.			
9			7:50	STT dictation. Noted he doesn't use it because he likes to see what he types. ?		Even though the VST STT does word by word dictation and display, for some reason its not what he wants. He said he likes to "see what he types" but in our observation it would "type what he says". Not sure we understand the root issue with him not using STT	What's intersting is that the one person who STT on their phone, does not use it. He was very familiar with it. Everyone else who doesn't have it, says they'd like it and would use it. Perhaps he knows something they don't?		
10			9:00	Uses dictation on desktop for entering documents, messaging, etc. Commented that they do want punctuation ability.					
11			14:00	Said that vocabulary in the phones for dictation needs to be greater.		Mentioned that you could expand the vocabulary - he seems to be the sole person out of all participants who has actually used the Sprint STT app.	He also knew how to correct misrecognitions and add new words to the dictionary for STT		
12		Where would you use dictation?	15:00	usage = everywhere . Some noted would avoid loud environments, while others thought it should be usable there.		Side comment by users made several times is that they often use SMS when their usage minutes is getting low or to avoid going over the # of minutes.			
13	features		15:40	New feature - would like an AGC control for speaker and mic volumen		Like a Blauplunk car stereo adapts its volume level - so when in a noisy env, would adj. levels			
14	features		17:20	STT entry for calendar - reminder setup		Would like to have voice entry for calendar events - setup as a reminder.			
15	features		19:20	STT entry for calendar		Calendar reminders should be able to come back in either voice or text (based on settings)			

Fig. 2. Screenshot of the template used to identify critical incident clusters

5 Results and Discussion

Some of the interesting findings are reported in this section. Regardless of users' experience level with SMS or ASR, the use of key pad for data entry is reported to be time-consuming. Data entry on mobile devices seems to be more difficult for the participants in the first group, who had no ASR and SMS experience, as compared to the other three groups. Participants expected the dictation feature be very accessible and intuitive to use. Some requested that the dictation feature be accessible via a single click. Although the dictation feature was desired for the convenience and safety (e.g., driving), most participants were concerned about the current performance of the speech recognition. Based on their subjective responses in the discussion, a minimum accuracy of 90% was expected by most of the participants to adopt ASR as a text entry method. Users with previous experience (desktop based dictation and/or phone based voice dialing) were more skeptical about the recognition accuracy but were willing to tolerate the errors if the error correction was easy and efficient.

On being asked to name some applications where dictation will be useful, participants listed a variety of daily tasks which were mundane but difficult to accomplish on cell phones due to the slow text input mechanisms. Analysis of data for potential applications found two distinctive trends namely ***communications*** use and ***personal*** use as shown in Table 3. Errors were deemed to matter most when communicating with other people. However, participants indicated higher acceptance, if errors occurred in information that was meant for personal use.

Table 3. User needs for dictation on mobile phones

Communication use	Personal use
• SMS/ Text Messaging • Email • Instant Message • Timed outgoing message • Fax an order • Shared calendar event	• Reminders/ Alarms • To-do list/ Grocery List/ Shopping lists • Calendar • Data entry to phonebook • Recording personal journal • Games • Internet • Quick note • Tracking travel expenses

When potential dictation errors were demonstrated, participants mentioned that they would try to correct the errors by using voice commands. This finding was a bit surprising as the researchers had expected the keypad to be the first resort for error correction. Only a few participants said that they would try and correct errors manually, which again reflects a higher performance expectation.

Other results included the perceived high usefulness of the message read back function using a text-to-speech (TTS) engine after the user is done dictating the text. The mechanical nature of the voice of the TTS engine was not favored by some participants. Lack of punctuation and inflexion in the feedback voice was pointed out as being annoying. The ASR was expected to correctly recognize different accents of non-native English speakers and was also expected to perform in noisy conditions such as on public transportation and in personal automobiles.

6 Design Recommendations and Conclusion

Based on the findings from this study, we recommend the following guidelines to improve the design of applications using dictation on mobile devices:

(1) users are more tolerant to speech recognition errors if the dictated information is for personal use rather than for communication use;
(2) high recognition accuracy is required for dictation feature to be adopted as a text input medium for communication use;
(3) to facilitate the adoption of dictation feature, introduce dictation on mobile phones for personal use prior to communications use; and
(4) participants liked the audible feedback (Text-To-Speech) of recognized text since it enables use of mobile phone in hands-free and eyes-free situations (e.g., driving).

Additionally, the following guidelines are recommended to improve the ease of use and efficiency of error correction methods on mobile devices:

(a) provide "re-speak the entire phrase' technique, as it was the initial reaction for error correction from all users;
(b) provide multi-modal interaction where users are allowed to select a mistaken word manually and speak a target word to replace the error; and
(c) allow users to type individual letters to correct errors when methods (a) and (b) do not work.

The error correction related findings are supported by literatures in this area, which further indicates that speech as an input medium is most effective when combined with other modalities [17, 20]. This was also confirmed by participants' comments on error correction techniques in this study.

In conclusion, the use of key pad for data entry is still time-consuming to most mobile phone users despite technologies like iTAP or T9 [16]. A strong need for dictation feature was identified in this study. All participants perceived the dictation feature to be very useful, but it has to be easily accessible. Potential applications of the dictation feature were found in two areas: communication use and personal use. Despite the potential usefulness, participants were concerned about the speech recognition performance and expected a high accuracy for them to adopt this technology.

References

1. Card, S.K., Moran, T.P., Newell, A.: The Psychology of Human-Computer Interaction, pp. 259–297. Lawrence Erlbaum Associates, New Jersey (1983)
2. CTIA Website: http://ctia.org/research_statistics/statistics/index.cfm/AID/10202
3. Cox, A., Walton, A.: Evaluating the viability of speech recognition for mobile text entry. In: Proceedings of HCI 2004: Design for Life. pp. 25–28 (2004)
4. Dragon Naturally Speaking Software website: http://www.nuance.com/naturallyspeaking/
5. Dunlop, M.D., Crossan, A.: Predictive text entry methods for mobile phones. Personal Technologies London: Springer Verlag vol. 4, pp.134–143 (2000)
6. Feng, J., Karat, C.-M., Sears, A.: How productivity improves in hands-free continuous dictation tasks: lessons learned from a longitudinal study. Interacting with Computers 17, 265–289 (2005)
7. Grinter, R., Eldridge, M.: y do tngrs luv 2 txt msg. In: Prinz, W., et al.: (eds.)Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work ECSCW '01, Dordecht, Netherlands: Kluwer, pp. 219–238 (2001)
8. Kamm, C.: User interfaces for voice applications. Paper presented in Colloquim: Human-Machine Communication by Voice at National Academy of Sciences at the Arnold and Mabel Beckman Center, Irvine, CA, February 8-9 (1993)
9. Karat, C.-M., Halverson, C., Karat, J., Horn, D.: Patterns of entry and correction in large vocabulary continuous speech recognition systems. In: Proceedings of CHI'99, pp. 568–575 (1999)
10. Leiser, R.G.: Improving natural language and speech interfaces by the use of metalinguistic phenomena. Applied Ergonomics 20, 168–173 (1989)

11. MacKenzie, I.S., Soukoreff, R.W.: Text entry for mobile computing: Models and methods, theory and practice. Human-Computer Interaction 17, 147–198 (2002)
12. Marturano, L., Wheatley, D.: User centered research and design at Motorola. In: Proceedings of CHI'2000, pp. 221–222 (2000)
13. Microsoft Vista: http://www.microsoft.com/enable/products/windowsvista/
14. Mobile Data Association Website: http://www.mda-mobiledata.org
15. Munteanu, C., Baecker, R., Penn, G., Toms, E., James, D.: The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives. In: Proceedings of Computer Human Interaction Conference, Montreal, Canada, pp. 493–502. ACM Press, New York (2006)
16. Oniszczak, A., MacKenzie, S.I.: A Comparison of Two Input Methods for Keypads on Mobile Devices. In: Proceedings of the third Nordic conference on Human-computer interaction, Tampere Finland, pp. 101–104. ACM Press, New York (2004)
17. Oviatt, S.L., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winogard, T., Landay, J., Larson, J., Ferro, D.: Designing the user interface for multimodal speech and gesture applications: state-of-the-art systems and research directions. Human-Computer Interaction 15(4), 263–322 (2000)
18. Palm Handheld Products website: http://www.palm.com/us/products/input/
19. Rudnicky, A.I., Lee, K-F., Hauptmann, A.G.: Survey of current speech technology. Communications of the ACM 37(3), 52–57 (1994)
20. Sears, A., Karat, C.-M., Oseitutu, K., Karimullah, A., Feng, J.: Productivity, satisfaction, and interaction strategies of individual with spinal cord injuries and traditional users interacting with speech recognition software. Universal Access in the information Society, vol. 1, pp. 4–15 (2001)
21. Silfverberg, M., MacKenzie, I.S., Korhonen, P.: Perdicting text entry speed on mobile phones. In: Proceedings of CHI 2000. Amsterdam: ACM Press, pp. 9–16 (2000)
22. Suhm, B., Myers, B., Waibel, A.: Multimodal error correction for speech user interfaces. ACM Transactions on Computer-Human Interaction 8(1), 60–98 (2001)
23. Tarasewich, P.: Evaluation of thumbwheel text entry methods. Extended Abstracts of the CHI 2003 Conference, pp. 756–757 (2003)
24. Waibel, A., Suhm, B., Vo, M.T., Yang, J.: Multimodal Interfaces For Multimedia Information Agents. In: International Conference on Acoustics, Speech, and Signal Processing 1997, ICASSP 1997, Munich, Germany, 04 (1997)