An Empirical Study on Users' Acceptance of Speech Recognition Errors in Text-Messaging

Shuang Xu, Santosh Basapur, Mark Ahlenius, and Deborah Matteo

Human Interaction Research, Motorola Labs, Schaumburg, IL 60196, USA
{shuangxu,sbasapur,mark.ahlenius,deborah.matteo}@motorola.com

Abstract. Although speech recognition technology and voice synthesis systems have become readily available, recognition accuracy remain a serious problem in the design and implementation of voice-based user interfaces. Error correction becomes particularly difficult on mobile devices due to the limited system resources and constrained input methods. This research is aimed to investigate users' acceptance of speech recognition errors in mobile text messaging. Our results show that even though the audio presentation of the text messages does help users understand the speech recognition errors, users indicate low satisfaction when sending or receiving text messages with errors. Specifically, senders show significantly lower acceptance than the receivers due to the concerns of follow-up clarifications and the reflection of the sender's personality. We also find that different types of recognition errors greatly affect users' overall acceptance of the received message.

1 Introduction

Driven by the increasing user needs for staying connected, fueled by new technologies, decreased retail price, and broadband wireless networks, the mobile device market is experiencing an exponential growth. Making mobile devices smaller and more portable brings convenience to access information and entertainment away from the office or home. Today's mobile devices are combining the capabilities of cell phones, text messaging, Internet browsing, information downloading, media playing, digital cameras, and much more. When mobile devices become more compact and capable, the user interface based on small screen and keypad can cause problems. The convenience of an ultra-compact cell phone is particularly offset by the difficulty of using the device to enter text and manipulate data. According to the figures announced by the Mobile Data Association [1], the monthly text messaging in UK broke through the 4 billion barrier for the first time during December 2006. Finding an efficient way to enter text on cell phones is one of the critical usability challenges in mobile industry. Many compelling text input techniques have been previously proposed to address the challenge in mobile interaction design [21, 22, and 41]. However, with the inherent hardware constraints of the cell phone interface, these techniques cannot significantly increase the input speed, reduce cognitive workload, or support hands-free and eyes-free interaction. As speech recognition technology and voice synthesis systems becoming readily available, Voice User Interfaces (VUI) seem to be an inviting solution, but not without problems. Speech recognition

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 232–242, 2007.

accuracy remains a serious issue due to the limited memory and processing capabilities available on cell phones, as well as the background noise in typical mobile contexts [3,7]. Furthermore, to correct recognition errors is particularly hard because: (1) the cell phone interfaces make manual selection and typing difficult [32]; (2) users have limited attentional resources in mobile contexts where speech interaction is mostly appreciated [33]; and (3) with the same user and noisy environment, re-speaking does not necessarily increase the recognition accuracy for the second time [15].

In contrast to the significant amount of research effort in the area of voice recognition, less is known about users' acceptance or reaction to the voice recognition errors. An inaccurately recognized speech input often looks contextually ridiculous, but it may phonetically make better sense. For examples, "The baseball game is canceled due to the *under stone* (thunderstorm)", or "Please send the driving directions to *myself all* (my cell phone)." This study investigates users' perception and acceptance of speech recognition errors in the text messages sent or received on cell phones. We aim to examine: (1) which presentation mode (visual, auditory, or visual and auditory) helps the receiver better understand the text messages that have speech recognition errors; (2) whether different types of errors (e.g., misrecognized names, locations, or requested actions) affect users' acceptance; and (3) what are the potential concerns users may have while sending or receiving text messages that contain recognition errors. The understanding of users' acceptance of recognition errors could potentially help us improve their mobile experience by optimizing between users' effort on error correction and the efficiency of their daily communications.

2 Related Work

The following sections explore the previous research with a focus on the three domains: (1) the inherent difficulties in text input on mobile devices and proposed solutions; (2) the current status and problems with speech recognition technology; and (3) the review of error correction techniques available for mobile devices.

2.1 Text Input on Mobile Device

As mobile phones become an indispensable part of our daily life, text input is frequently used to enter notes, contacts, text messages, and other information. Although the computing and imaging capabilities of cell phones have significantly increased, the dominant input interface is still limited to a 12-button keypad and a discrete four-direction joystick. This compact form provides users the portability, but also greatly constrains the efficiency of information entering. On many mobile devices, there has been a need for simple, easy, and intuitive text entry methods. This need becomes particularly urgent due to the increasing usage of text messaging and other integrated functions now available on cell phones. Several compelling interaction techniques have been proposed to address this challenge in mobile interface design. Stylus-based handwriting recognition techniques are widely adopted by mobile devices that support touch screens. For example, Graffiti on Palm requires users to learn and memorize the predefined letter strokes. Motorola's WisdomPen [24] further supports natural handwriting recognition of Chinese and Japanese

characters. EdgeWrite [39, 41] proposes a uni-stroke alphabet that enables users to write by moving the stylus along the physical edges and into the corners of a square. EdgeWrite's stroke recognition by detecting the order of the corner-hits can be adopted by other interfaces such as keypad [40]. However, adopting EdgeWrite on cell phones means up to 3 or 4 button clicks for each letter, which makes it slower and less intuitive than the traditional keypad text entry. Thumbwheel provides another solution for text entry on mobile devices with a navigation wheel and a select key [21]. The wheel 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. As a text entry method designed for cell phones, Thumbwheel is easy to learn but slow to use, depending on the device used, the text entry rate varies between 3 to 5 words per minute (wpm) [36]. Other solutions have been proposed to reduce the amount of scrolling [5, 22]. But these methods require more attention from the user on the letter selection, therefore do not improve the text entry speed. Prediction algorithms are used on many mobile devices to improve the efficiency of text entry. An effective prediction program can help the user complete the spelling of a word after the first few letters are manually entered. It can also provide candidates for the next word to complete a phrase. An intelligent prediction algorithm is usually based on a language model, statistical correlations among words, context-awareness, and the user's previous text input patterns [10, 11, 14, and 25]. Similar to a successful speech recognition engine, a successful prediction algorithm may require higher computing capability and more memory capacity, which can be costly for portable devices such as cell phones.

The above discussion indicates that many researchers are exploring techniques from different aspects to improve the efficiency of text entry on mobile devices. With the inherent constraints of the cell phone interface, however, it remains challenging to increase the text input speed and reduce the user's cognitive workload. Furthermore, none of the discussed text entry techniques can be useful in a hands-busy or eyes-busy scenario. With the recent improvement of speech recognition technology, voice-based interaction becomes an inviting solution to this challenge, but not without problems.

2.2 Speech Recognition Technology

As mobile devices grow smaller and as in-car computing platforms become more common, traditional interaction methods seem impractical and unsafe in a mobile environment such as driving [3]. Many device makers are turning to solutions that overcome the 12-button keypad constraints. The advancement of speech technology has the potential to unlock the power of the next generation of mobile devices. A large body of research has focused on how to deliver a new level of convenience and accessibility with speech-drive interface on mobile device. Streeter [30] concludes that universality and mobile accessibility are the major advantages of speech-based interfaces. Speech offers a natural interface for tasks such as dialing a number, searching and playing songs, or composing messages. However, the current automatic speech recognition (ASR) technology is not yet satisfactory. One challenge is the limited memory and processing power available on portable devices. ASR typically involves extensive computation. Mobile phones have only modest computing resources and battery power compared with a desktop computer. Network-based speech recognition could be a solution, where the mobile device must connect to the server to use speech recognition. Unfortunately, speech signals transferred over a

wireless network tend to be noisy with occasional interruptions. Additionally, network-based solutions are not well-suited for applications requiring manipulation of data that reside on the mobile device itself [23]. Context-awareness has been considered as another solution to improve the speech recognition accuracy based on the knowledge of a user's everyday activities. Most of the flexible and robust systems use probabilistic detection algorithms that require extensive libraries of training data with labeled examples [14]. This requirement makes context-awareness less applicable for mobile devices. The mobile environment also brings difficulties to the utilization of ASR technology, given the higher background noise and user's cognitive load when interacting with the device under a mobile situation.

2.3 Error Correction Methods

Considering the limitations of mobile speech recognition technology and the growing user demands for a speech-driven mobile interface, it becomes a paramount need to make the error correction easier for mobile devices. A large group of researchers have explored the error correction techniques by evaluating the impact of different correction interfaces on users' perception and behavior.

User-initiated error correction methods vary across system platforms but can generally be categorized into four types: (1) re-speaking the misrecognized word or sentence; (2) replacing the wrong word by typing; (3) choosing the correct word from a list of alternatives; and (4) using multi-modal interaction that may support various combinations of the above methods. In their study of error correction with a multi-modal transaction system, Oviatt and VanGent [27] have examined how users adapt and integrate input modes and lexical expressions when correcting recognition errors. Their results indicate that speech is preferred over writing as input method. Users initially try to correct the errors by re-speaking. If the correction by re-speaking fails, they switch to the typing mode [33]. As a preferred repair strategy in human-human conversation [8], re-speaking is believed to be the most intuitive correction method [9,15, and 29]. However, re-speaking does not increase the accuracy of the re-recognition. Some researchers [2,26] suggest increasing the recognition accuracy of re-speaking by eliminating alternatives that are known to be incorrect. They further introduce the correction method as "choosing from a list of alternative words". Sturm and Boves [31] introduce a multi-modal interface used as a web-based form-filling error correction strategy. With a speech overlay that recognizes pen and speech input, the proposed interface allows the user to select the first letter of the target word from a soft-keyboard, after which the utterance is recognized again with a limited language model and lexicon. Their evaluation indicates that this method is perceived to be more effective and less frustrating as the participants feel more in control. Other research [28] also shows that redundant multimodal (speech and manual) input can increase interpretation accuracy on a map interaction task.

Regardless of the significant amount of effort that has been spent on the exploration of error correction techniques, it is often hard to compare these techniques objectively. The performance of correction method is closely related to its implementation, and evaluation criteria often change to suit different applications and domains [4, 20]. Although the multimodal error correction seems to be promising among other techniques, it is more challenging to use it for error correction of speech input on mobile phones. The main reasons are: (1) the constrained cell phone

interface makes manual selection and typing more difficult; and (2) users have limited attentional resources in some mobile contexts (such as driving) where speech interaction is mostly appreciated.

3 Proposed Hypotheses

As discussed in the previous sections, text input remains difficult on cell phones. Speech-To-Text, or dictation, provides a potential solution to this problem. However, automatic speech recognition accuracy is not yet satisfactory. Meanwhile, error correction methods are less effective on mobile devices as compared to desktop or laptop computers. While current research has mainly focused on how to improve usability on mobile interfaces with innovative technologies, very few studies have attempted to solve the problem from users' cognition perspective. For example, it is not known whether the misrecognized text message will be sent because it sounds right. Will audible play back improve receivers' comprehension of the text message? We are also interested in what kind of recognition errors are considered as critical by the senders and receivers, and whether using voice recognition in mobile text messaging will affect the satisfaction and perceived effectiveness of users' everyday communication. Our hypotheses are:

Understanding: H1. The audio presentation will improve receivers' understanding of the mis-recognized text message.

We predict that it will be easier for the receivers to identify the recognition errors if the text messages are presented in the auditory mode. A misrecognized voice input often looks strange, but it may make sense phonetically [18]. Some examples are:

[1.Wrong]"How *meant it ticks* do you want me to buy for the white sox game next week?"
[1.Correct] "How *many tickets* do you want me to buy for the white sox game next week?"
[2.Wrong] "We are on our way, will be at *look what the* airport around noon."
[2.Correct] "We are on our way, will be at *LaGuardia* airport around noon"

The errors do not prevent the receivers from understanding the meaning delivered in the messages. Gestalt Imagery theory explains the above observation as the result of human's ability to create an imaged whole during language comprehension [6]. Research in cognitive psychology has reported that phonological activation provides an early source of constraints in visual identification of printed words [35, 42]. It has also been confirmed that semantic context facilitates users' comprehension of aurally presented sentences with lexical ambiguities. [12, 13, 34, 37, and 38].

Acceptance: H2. Different types of errors will affect users' acceptance of sending and receiving text messages that are misrecognized.

Different types of error may play an important role that affects users' acceptance of the text messages containing speech recognition errors. For example, if the sender is requesting particular information or actions from the receiver via a text message, errors in key information can cause confusion and will likely be unacceptable. On the other hand, users may show higher acceptance for errors in general messages where there is no potential cost associated with the misunderstanding of the messages.

Satisfaction: H3. Users' overall satisfaction of sending and receiving voice dictated text messages will be different.

We believe that senders may have higher satisfaction because the voice dictation makes it easier to enter text messages on cell phones. On the other hand, the receivers may have lower satisfaction if the recognition errors hinder their understanding.

4 Methodology

To test our hypotheses, we proposed an application design of dictation. Dictation is a cell phone based application that recognizes a user's speech input and converts the information into text. In this application, a sender uses ASR to dictate a text message on the cell phone. While the message is recognized and displayed on the screen, it will also be read back to the sender via Text-To-Speech (TTS). The sender can send this message if it *sounds* close enough to the original sentence, or correct the errors before sending. When the text message is received, it will be visually displayed and read back to the receiver via TTS as well. A prototype was developed to simulate users' interaction experience with the dictation on a mobile platform.

Participants. A total of eighteen (18) people were recruited to participate in this experiment. They ranged in age from 18 to 59 years. All were fluent speakers of English and reported no visual or auditory disabilities. All participants currently owned a cell phone and have used text messaging before. Participants' experience with mobile text messaging varied from novice to expert, while their experience with voice recognition varied from novice to moderately experienced. Other background information was also collected to ensure a controlled balance in demographic characteristics. All were paid for their participation in this one-hour study.

Experiment Design and Task. The experiment was a within-subject task-based one-on-one interview. There were two sections in the interview. Each participant was told to play the role of a message sender in one section, and the role of a message receiver in the other section. As a sender, the participant was given five predefined and randomized text messages to dictate using the prototype. The "recognized" text message was displayed on the screen with an automatic voice playback via TTS. Participants' reaction to the predefined errors in the message was explored by a set of interview questions. As a receiver, the participant reviewed fifteen individual text messages on the prototype, with predefined recognition errors. Among these messages, five were presented as audio playbacks only; five were presented in text only; the other five were presented simultaneously in text and audio modes. The task sequence was randomized as shown in Table 1:

Table 1. Participants Assignment and Task Sections

Participants Assignment			**Task Section 1**	**Task Section 2**
18~29 yrs	30~39 yrs	40~59 yrs		
S#2(M)	S#8(F)	S#3(M)	Sender	Receiver-Audio, Text, A+T
S#1(F)	S#4(M)	S#14(F)	Sender	Receiver-Text, A+T, Audio
S#7(M)	S#16(F)	S#9(M)	Sender	Receiver-A+T, Audio, Text
S#5(F)	S#6(M)	S#15(F)	Receiver-Audio, Text, A+T	Sender
S#10(M)*	S#17(F)	S#13(M)	Receiver-Text, A+T, Audio	Sender
S#12(F)	S#11(M)	S#18(F)	Receiver-A+T ,Audio, Text	Sender

*S#10 did not show up in the study.

Independent Variable and Dependent Variables. For senders, we examined how different types of recognition errors affect their acceptance. For receivers, we examined (1) how presentation modes affect their understanding of the misrecognized messages; and (2) whether error types affect their acceptance of the received messages. Overall satisfaction of participants' task experience was measured for both senders and receivers, separately. The independent and dependent variables in this study are listed in Table 2:

Table 2. Independent and Dependent Variables

	Independent Variables	**Dependent Variables**
Senders	1. Error Types: Location, Requested Action, Event/occasion, Requested information, Names.	1. Users' Acceptance 2. Users' Satisfaction
Receivers	1. Presentation Modes: Audio, Text, Audio + Text 2. Error Types: Location, Requested Action, Event/occasion, Requested information, Names.	1. Users' Understanding 2. Users' Acceptance 3. Users' Satisfaction

Senders' error acceptance was measured by their answers to the question "*Will you send this message without correction?*" in the interview. After all errors in each message were exposed by the experimenter, receivers' error acceptance was measured by the question "*Are you OK with receiving this message?*" Receivers' understanding performance was defined as the percentage of successfully corrected errors out of the total predefined errors in the received message. A System Usability Score (SUS) questionnaire was given after each task section to collect participants' overall satisfaction of their task experience.

Procedures. Each subject was asked to sign a consent form before participation. Upon their completion of a background questionnaire, the experimenter explained the concept of dictation and how participants were expected to interact with the prototype. In the Sender task section, the participant was told to read out the given text message loud and clear. Although the recognition errors were predefined in the program, we allowed participants to believe that their speech input was recognized by the prototype. Therefore, senders' reaction to the errors was collected objectively after each message. In the Receiver task section, participants were told that all the received messages were entered by the sender via voice dictation. These messages may or may not have recognition errors. Three sets of messages, five in each, were used for the three presentation modes, respectively. Participants' understanding of the received messages was examined before the experimenter identified the errors, followed by a discussion of their perception and acceptance of the errors. Participants were asked to fill out a satisfaction questionnaire at the end of each task section. All interview sections were recorded by a video camera.

5 Results and Discussion

As previously discussed, the dependent variables in this experiment are: Senders' Error Acceptance and Satisfaction; and Receivers' Understanding, Error Acceptance, and Satisfaction. Each of our result measures was analyzed using a single-factor

ANOVA. F and P values are reported for each result to indicate its statistical significance. The following sections discuss the results for each of the dependent variables as they relate to our hypotheses.

Understanding. Receivers' understanding of the misrecognized messages was measured by the number of corrected errors divided by the number of total errors contained in each message. Hypothesis H1 was supported by the results of ANOVA, which indicates the audio presentation did significantly improve users' understanding of the received text messages ($F_{2,48}$=10.33, p<.001), as shown in Fig. 1a. Age was not one of the independent variables in this study, but was examined for precaution. Fig. 1b shows that users' age has an impact on their understanding of the errors. Compared to the younger groups who also had more experience with mobile text messaging, the older generation had a lower understanding performance ($F_{2,48}$=5.24, p=.009).

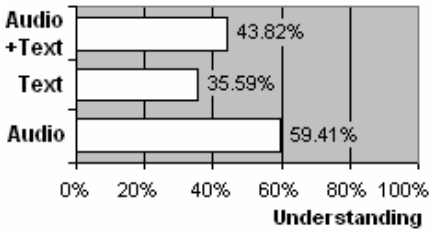


Fig. 1a. Presentation Modes vs. Understanding

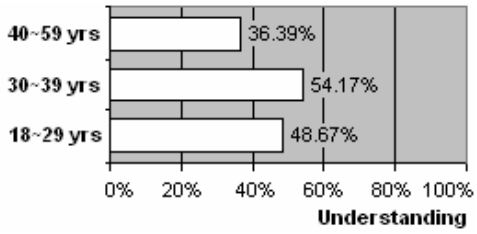
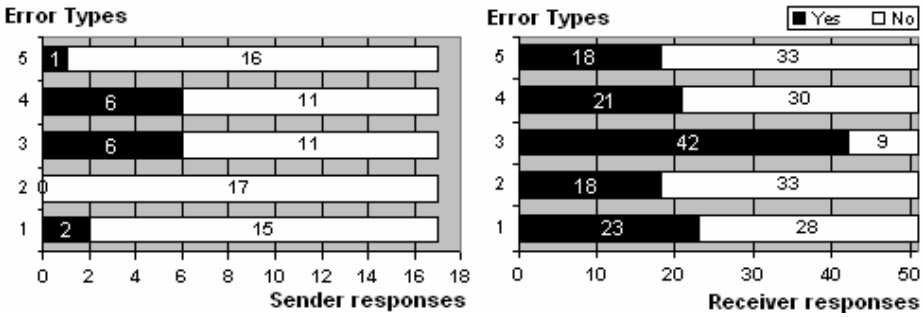


Fig. 1b. Age Groups vs. Understanding

The findings are consistent with many research studies in cognitive psychology. Swinney confirms that prior semantic context facilitates participants' comprehension of aurally presented sentences that contain lexical ambiguities [34]. Luo, Johnson, and Gallo [19] find phonological recoding occurs automatically and mediates lexical access in visual word recognition and reading. Other studies [16, 17] reveal that semantic processing was evident before the acoustic signal was sufficient to identify the words uniquely. These findings indicate that semantic integration can begin to operate with incomplete or inaccurate information.

Acceptance. In our study, senders' error acceptance was measured by their answers to the interview question "Will you send this message without correction?" Receivers' error acceptance was measured by the question "Are you OK with receiving such a message?" after the errors in each message were identified by the experimenter. Hypotheses H2 concerned about the relation between different error types and users' acceptance. As shown in Fig. 2a and 2b, the occurrence of different kinds of errors (see Table 2) had a significant impact on senders' ($F_{4,80}$=3.60, p=.010). and receivers' ($F_{4,250}$=8.92, p<.001) acceptance. Senders showed much lower tolerance for errors in the requested actions and person's names, among the five controlled error types. Receivers indicated a different pattern in acceptance, where they showed significantly higher acceptance for general informative messages regarding upcoming events and occasions. H2 was thus supported by the findings.



1.Location; **2.**Requested Action; **3.**Event/Occasion; **4.**Requested information; **5.**Person's Name

Fig. 2a. Senders' Error Acceptance **Fig. 2b.** Receivers' Error Acceptance

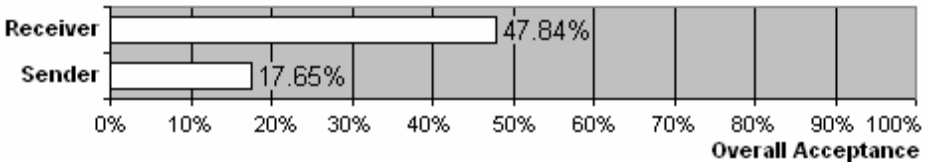


Fig. 2c. Users' Overall Error Acceptance (Receivers vs. Senders)

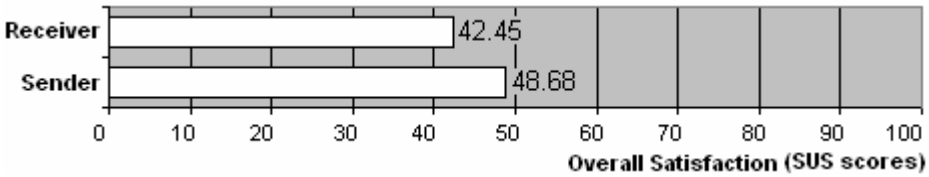


Fig. 3. Users' Overall Satisfaction (Receivers vs. Senders)

A significant difference ($F_{1,66}$=30.01, p<.001) between senders' and receivers' overall acceptance was reported in this study, as shown in Fig.2c. This finding was confirmed by users' debriefing comments during the interview. When sending text messages containing recognition errors, most users were concerned about their self-reflection, as well as the communications needed afterwards to clarify the confusion. As the receiver of misrecognized messages, users gradually developed deciphering skills based on phonetic similarity, common sense, and context of the messages.

Satisfaction. System Usability Scores (SUS) were used in this study to examine users' overall satisfaction with their task experience. As we predicted, the senders reported slightly higher satisfaction than the receivers. A plausible explanation is that senders' dissatisfaction with the inaccurate voice recognition was partially countered by the perceived convenience of entering text with speech input. However, the

difference was not significant enough ($F_{1,66}$=1.06, p=.308) to reject the null hypothesis, therefore H3 was not supported by our findings (see results in Fig.3).

An affinity diagram was used in the data analysis of user's comments. Over 1000 incidents were collected and categorized. Some preliminary findings are:

- Users are concerned about the hidden cost associated with miscommunication (e.g., taking the wrong actions, additional time and effort spent on error-decoding, not being able to clarify, etc.)
- Some users showed higher error acceptance for urgent messages so to get the info. out as soon as possible, others showed lower acceptance because they believed words must be accurate in an important message.
- Critical information in a message must be error-free (e.g., when, where, what, who.)
- Users are willing to adapt to the voice recognition system for effective use (e.g., keep messages short and concise, avoid using words that often cause problems, train the voice recognition system to accurately pick up frequently used names, etc.)
- Personality traits and the familiarity between the sender and receiver also affect users' error acceptance in mobile messaging.
- Transfer of training does not work between identification of typing errors and identification of speech recognition errors.

6 Conclusions

This study investigated users' acceptance of speech recognition errors in text messaging. We hypothesized that audio presentation of the misrecognized message would improve receivers' understanding because of the phonetic similarity. We also predicted that different types of errors could affect users' acceptance of the message. Our findings revealed that the audio + text presentation was preferred by most of the users, and the audio playback significantly improved users' understanding. Users indicated overall low acceptance for errors in text messaging. The major concern was the consequence of misunderstanding: a confusing message may trigger a series of follow-up phone calls, which defeats the purpose of quick communication via SMS. This also explains why users showed much lower tolerance for errors in messages that elicit actions or information. In this within-subject study, interestingly, participants showed significantly lower acceptance as a message sender than as a receiver. Although the senders would like to use voice recognition to dictate text messages for convenience and safety concerns, they preferred to correct errors before sending the messages to ensure a clear and efficient communication. In conclusion, our hypotheses were supported by the results from this study. Based on the understanding of users' acceptance and reaction to recognition errors in mobile text messaging, we expect to develop guidelines for the interaction design of dictation to improve its effectiveness as a text input method on mobile devices. However, this study is only a first step towards this direction. Future work should further explore how to control error occurrence in critical information, and how to make error correction easier via a multi-modal interface.

References

Selected References (full references are available upon request)

3. Alewine, N., Ruback, H., Deligne, S.: Pervasive Speech Recognition. IEEE Pervasive Computing 3(4), 78–81 (2004)
6. Bell, N.: Gestalt imagery: A critical factor in language comprehension. Annals of Dyslexia 41, 246–260 (1991)
13. Frost, R., Kampf, M.: Phonetic recoding of phonological ambiguous printed words. Journal of Experimental Psychology: Learning, Memory, and Cognition 19, 23–33 (1993)
15. Larson, K., Mowatt, D.: Speech error correction: the story of the alternates list. International Journal of Speech Technology, 6(2), 183–194. Memory, and Cognition, 24, 573–592 (2003)
18. Lieberman, H., Faaborg, A., Daher, W., Espinosa, J.: How to wreck a nice beach you sing calm incense. In: Proceedings of Intelligent User Interface'05, pp. 278–280 (2005)
22. MacKenzie, I.S., Soukoreff, R.W.: Text entry for mobile computing: Models and methods, theory and practice. Human-Computer Interaction 17, 147–198 (2002)
27. Oviatt, S., VanGent, R.: Error resolution during multimodal human-computer interaction. In: Proceedings of International Conference on Spoken Language Process (ICSLP'96), pp. 204–207 (1996)
33. Suhm, B., Myers, B., Waibel, A.: Multimodal error correction for speech user interfaces. ACM Transactions on Computer-Human Interaction 8(1), 60–98 (2001)
37. Van Orden, G.C.: A ROWS is a ROSE: Spelling, sound, and reading. Memory and Cognition 15, 181–198 (1987)