Understanding Users' Perception of Speech Recognition Errors in Mobile Communication

1. Introduction

Over the last decade, mobile devices have evolved from business tools and luxury gadgets to an integral means of communication. Making mobile devices smaller and more portable brings convenience to access information and entertainment anytime, anywhere. When mobile devices become more compact and capable, the user interface based on small screen and keypad brings usability issues. The convenience of an ultra-compact cell phone is particularly offset by the difficulty of using the device to enter text and manipulate data. Finding an efficient way to enter text on a cell phone is one of the essential UI design challenges in mobile industry.

Technologies for speech recognition and synthesis are becoming increasingly sophisticated, which has the potential for making mobile device easier to use when entering text and controlling applications. The speech-based interface is especially important in hands-busy and/or eyes-busy situations. For example, via voice-based interactions, users can keep their hands on the steering wheel while making / receiving calls or text messaging.

However, the current speech recognition technology is inherently error prone (Bradford, 1995). The constrained system resources, such as limited memory and processing capabilities, bring challenges to the advancement of Automatic Speech Recognition (ASR) technology of the portable devices. Furthermore, mobile environments are more demanding than traditional computer-based environments. The higher background noise hinders the improvement of speech recognition accuracy (Alewine, Ruback and Deligne, 2004).

Despite the significant amount of research effort in the area of error correction (Suhm, Myers and Waibel, 1999 & 2001), it remains an unsuccessfully addressed usability problem in the design and implementation of speech user interfaces. Errors generated by speech recognition are usually easy to notice because they don't make sense. The manual correction, however, is difficult due to the constraints in text input capability, especially in mobile situations. Although re-speaking is a preferred repair strategy in human-human dialogue (Baber and Hone, 1993), it does not necessarily increase the probability that the same user's voice input can be correctly recognized by the same ASR engine under the same condition for the second time.

In this paper, we propose a potential solution to the problem discussed above. The proposed design is a cell phone based dictation that allows a user to speak in a free-form style. User's voice input will be converted into text messages by the embedded ASR engine. The recognized text information will be read out to the user via Text-To-Speech (TTS) before being sent. When the message is received, it will be visually displayed and also output via TTS. The design idea is based on the observation that an inaccurately recognized speech input often looks contextually incorrect, but it may phonetically make better sense. A user study was conducted to evaluate this design idea. The goal of this study is to explore users' understanding, perception, and acceptance of speech recognition errors in mobile communication. Findings from this study could help us better understand how to optimize between users' effort on error correction and the effectiveness of their daily communications via text messaging.

2. Literature Review

Mobile device manufacturers and carriers are striving to deliver functional components and systems with advanced features that are easy to access and use. As mobile devices grow smaller and as in-car computing platforms become more common, traditional interaction methods seem impractical and unsafe (Alewine, et al., 2004). 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 a speech-driven interface on mobile devices.

As a more interactive and less formal expression media, speech is preferred for handling the complex, equivocal and emotional aspects of collaborative tasks (Chalfonte, Fish and Kraut, 1991). Speech also offers a natural interface for accomplishing tasks such as dialing a contact number, searching and playing songs, composing messages, or accessing data on a mobile phone. However, the current ASR technology is not yet satisfactory. One major challenge is the constrained system resources available on portable devices, such as limited memory and processing power. ASR typically involves extensive computation. Mobile phones have modest computing resources and battery power, as compared with a desktop computer or server. Network-based speech recognition seems to be a solution, where the mobile device must connect to the server to use speech recognition. However, network-based solutions are not well suited for applications requiring data that reside on the mobile devices (Marcussen, 2003). Context-awareness has been considered as another solution to improve the accuracy of speech recognition 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 (Intille, 2004; Danninger, 2005; Mantoro, 2003; Desilets, 2006), which makes them less applicable for mobile devices. The mobile environment also challenges the utilization of ASR technology, given the higher background noise and user's cognitive load when interacting with the device under a mobile situation.

On the other hand, consumers highly demand for speech solutions on mobile phones. A recent survey (http://www.nuance.com/unlockthepower/) reports a clear interest from end-users for applications that simplify access to functionality on the phone and enhance productivity while increasing safety. Hands-free and eyes-free access to mobile phone features are highly desired for convenience.

Considering the limitations of mobile speech recognition technology and the growing user needs for a speech-driven mobile interface, it becomes critical to make error correction easier on mobile devices. Researchers in speech communication have been actively exploring the error handling mechanisms in ASR systems. Their effort has focused on three major issues: error reduction or prevention; error detection; and error correction (Mankoff and Abowd, 1999).

Most techniques for error reduction and detection are architecture-oriented, which embed the error handling mechanism in components such as a recognizer engine, an acoustic model, or a dialog manager. For example, confidence measures are often obtained during the speech recognition processes of a dialog system to guide the behavior of the dialog manager. Torres, Hurtado, Garcia, Sanchis, and Segarra (2005) describe an approach that allows the system to ask the user for confirmation about the data that have low confidence values associated to them. Their evaluation of the proposed error handling through confidence measures used in a stochastic

dialog system indicates that this technique could reduce the system recognition errors. Prodanov and Drygajlo (2005) introduce a probabilistic model based architecture for error handling. They use a Bayesian network framework to interpret multimodal signals in the spoken dialog between a tour-guide robot and visitors in exhibition conditions. They indicate that correct understanding of a user's goal or intention is the key for a successful communication between the robot and human visitors. Similarly, McTear, O'Neill, Hanna, and Liu (2005) present an approach based on the theory of grounding. An architecture is proposed in their approach to incorporate generic confirmation strategies with domain specific heuristics to complete a transaction.

A large group of researchers have explored the error correction techniques from the users' perspective by evaluating the impact of different correction interfaces on users' perception and behavior. User-initiated error correction methods can be categorized into four types: (1) re-speaking the misrecognized word; (2) replacing the wrong word by typing; (3) choosing the correct word from a list of alternatives; and (4) using multi-modal interaction that supports various combinations of the above methods. In their study of error correction with a multimodal transaction system, Oviatt and VanGent (1996) have examined how users adapt and integrate input modes and lexical expressions when correcting recognition errors. The results indicate that speech is preferred over writing as input mode. Subjects initially try to correct the errors by re-speaking. If the correction by re-speaking fails, subjects then switch to the typing mode (Suhm et al., 2001). As a preferred repair strategy in human-human conversation (Brinton, Fujiki and Sonnenberg, 1988), re-speaking is believed to be the most intuitive correction method (Chen and Tremaine, 2006; Larson and Mowatt, 2003; Robbe, Carbonell and Valot, 1994). However, re-speaking does not increase the accuracy of the re-recognition. Some researchers (Ainsworth and Pratt, 1992; Murray, Frankish and Jones, 1993) suggest increasing the recognition accuracy of re-speaking by eliminating alternatives that are known to be incorrect. They introduce the correction method as "choosing from a list of alternative words". Sturm and Boves (2005) propose an error correction strategy by using a web-based form-filling interface. With a speech overlay that recognizes pen and speech input, the proposed multimodal 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 (Oviatt, 1999) also shows that redundant multimodal (speech and manual) input can increase interpretation accuracy on a map interaction task.

Although the multimodal error correction seems to be promising among other techniques, it is challenging to use it for error correction on mobile phones. The main reasons are: (1) cell phone interfaces make manual selection and typing difficult; and (2) users have limited attentional resources in some mobile contexts (such as driving) where speech interaction is mostly appreciated.

3. Proposed Design and Research Questions

As discussed above, text entry remains difficult on cell phones. Voice recognition provides a potential solution to this problem. However, ASR accuracy is not yet satisfactory. Meanwhile, error correction methods are less effective on mobile devices. Thus, mobile interface designers are challenged to provide an easy and intuitive tool to facilitate mobile communication such as text messaging. While previous research has been focusing on how to improve mobile usability with innovative technologies, few studies have attempted to solve the problem from the

perspective of users' cognition and perception. Little is known regarding whether a receiver can understand text messages that contain speech recognition errors. Will audible readout of the message improve users' comprehension? What kind of recognition errors are considered as critical by senders and/or receivers? How will speech recognition errors affect users' satisfaction and perceived effectiveness of mobile communication?

Researchers in cognitive psychology have reported that phonological activation provides an early source of constraints in visual identification of printed words (Tan and Perfetti, 1999; Ziegler and Jacobs, 1995). Phonological Hypothesis assumes that phonological information is a constituent of visual word identification and is essential in understanding the word. This hypothesis has been supported by a variety of empirical studies. Van Orden's studies (1987 & 1994) have indicated that participants detected homophone imposters by verifying target foil spellings against their knowledge of the correct spellings of category exemplars. Swinney (1979) confirms that prior semantic context facilitates participants' comprehension of aurally presented sentences that contain lexical ambiguities. Luo, Johnson and Gallo (1998) find phonological recoding occurs automatically and mediates lexical access in visual word recognition and reading. Lukatela and Turvey's experiments (1994) suggest that lexical access is initially phonological. They also find that phonologically activated representations are eventually suppressed when the input orthography does not match the addressed spelling of the words. Research in phonological ambiguity suggests that printed homographs provide confirmation for fast phonetic recoding in reading (Frost and Kampf, 1993; Intille et al., 2004). Other studies (Frost, 1995; Lesch and Pollatsek, 1993 & 1998) 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 of word identity. In other words, a short message may be understood by the receiver if it makes sense phonologically, despite spelling mistakes.

Besides the research findings in cognitive psychology, it is also observed that an inaccurately recognized voice input often looks incorrect, but it may make better sense phonetically (Lieberman, Faaborg, Daher and Espinosa, 2005). Some examples are listed as follows:
[1.Wrong] "Please *stand* the driving directions to my *self all*."
[1.Correct] "Please *send* the driving directions to my *cell phone*."
[2.Wrong] "The baseball game was cancelled due to the *under stone*."
[2.Correct] "The baseball game was cancelled due to the *thunder storm*."
The errors do not prevent readers 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 (Bell, 1991).

In this study, we propose a mobile application design of dictation to improve the efficiency of text messaging. Dictation is an application that recognizes a user's free-formed speech input and converts the information into text. In the proposed design, a sender uses ASR to dictate a message on his/her cell phone. While the text message is recognized and displayed on the cell phone screen, it will also be read out to the sender via TTS. The sender can send this message if it *sounds* close enough to the original sentence, otherwise the sender can re-dictate the entire message or make corrections manually. When the text message is received, it will be displayed on the cell phone screen and readout via TTS as well.

A user study was carried out to validate this design idea. The main research question of this study was whether users would understand and accept text messages that contain recognition errors. We were also interested in what kind of recognition errors would be considered as severe,

and how the VR errors would impact the satisfaction and effectiveness of mobile communication. The main research question is broken down into the following measurable research questions:
(1) *Will the audio presentation help receivers understand the misrecognized text messages?*
We believed that it would be easier for the receivers to identify the recognition errors if an audio readout was available. We predicted that receivers would understand the message despite the recognition errors, as long as the message was phonetically close to the original text.
(2) *Will different types of errors affect users' acceptance of the dictated text messages?*
Different types of errors may play an important role in users' acceptance. We believed users might show higher acceptance for errors if their misunderstanding of the message would not cause severe consequences.
(3) *Will senders and receivers indicate different overall satisfaction of using dictation for text messaging?*
We believed that senders might have higher satisfaction because voice dictation made it easier to enter text messages on cell phones, whereas the receivers might have lower satisfaction if the recognition error hindered their understanding.

4. Experiment

A computer-based prototype was developed to simulate users' interaction experience with the proposed design on a mobile platform. As shown in Figure 1a and 1b, a Motorola ROKR E2 simulator was used to display and read out the misrecognized messages to a sender in the experiment. A Motorola RAZR V3 simulator was used to present the received messages to a receiver.

Figure 1a. Interface for senders with a misrecognized text message.

Figure 1b. Interface for receivers with a mis-recognized message.

4.1 Participants

Seventeen users (9 females) were recruited externally to participate in this one-hour study, ranging in age from 20 to 60 years. The participants were all native English speakers with no sight or hearing disabilities. All participants currently own a cell phone and have used text messaging before. Their experience with text messaging varied from novice, intermediate, to veteran (mean 6.5 messages/day). Nine participants considered themselves novice users of voice recognition technology, and the rest claimed intermediate VR experience with voice-dialing on cell phones, computer applications, or navigation systems in the car. They were paid for their individual participation.

4.2 Tasks

Task-based interviews were conducted in this study. Each participant was asked to act as a message sender in one task section, and a message receiver in the other task section. Five pre-selected but randomized text messages were given to the sender to dictate on the prototype. The "recognized" text message was shown on the screen with an automatic voice readout via TTS. Participants were not aware that all recognition errors were predefined, their reactions and comments regarding the task experience were explored via a set of follow-up interview questions. As a message receiver, the participant reviewed fifteen individually received messages on the prototype, with predefined recognition errors. Among these messages, five were presented as audio readout only; five were presented in text only; the rest were presented simultaneously in text and audio modes. Similarly, receivers' understanding of the misrecognized text messages and their task experience were captured by the follow-up interview questions. The sequence of the two task sections, the order of the messages given in each section, and the order of the three presentation modes were randomized with a controlled balance among the participants.

4.3 Quantitative Measures

For senders, we examined how different types of recognition errors affect their acceptance. The five types of errors predefined in this study included errors in (1) location; (2) requested action; (3) event or occasion; (4) requested information; and (5) person's name. Senders' error acceptance was measured by their answers to the question "*Will you send this message without correction?*" in the interview.

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. 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 was defined as the percentage of successfully corrected errors out of the total predefined errors in each received message.

Overall satisfaction of participants' task experience was measured for both senders and receivers, separately. A System Usability Score (SUS) questionnaire was given after each task section to collect participants' overall satisfaction of their task experience.

4.4 Procedures

After a participant signed a consent form and filled out a background questionnaire, she was given instructions regarding the concept of dictation and how to interact with the prototype. As a sender, a participant was told to press the VR key on the cell phone interface, then read the given text message loud and clear. In order to objectively capture participants' reaction, they were led to believe that their speech input was recognized by the prototype. As a receiver, a participant was told that all received messages were entered by a sender via voice dictation. These messages may or may not have recognition errors. 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 video recorded.

5. Results and Discussion

5.1 Quantitative Findings

As previously mentioned, we examined the following variables in this study: senders' error acceptance; receivers' understanding and error acceptance, and participants' overall satisfaction. Each measure was analyzed using a single-factor ANOVA. The following sections discuss the quantitative findings from this experiment.

As shown in Figure 2a, a significant difference ($F_{1,66}$=30.01, p<.001) between senders' and receivers' overall acceptance was reported in this study. This finding was confirmed by users' comments during the interview. When sending text messages containing recognition errors, most users were concerned about their self-reflection, and 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.

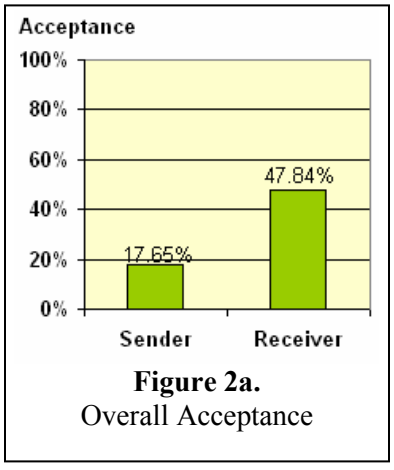

Figure 2a.
Overall Acceptance

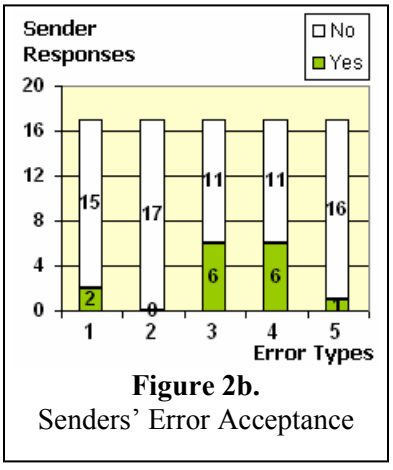

Figure 2b.
Senders' Error Acceptance

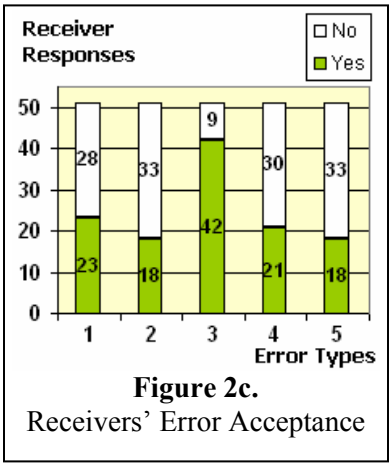

Figure 2c.
Receivers' Error Acceptance

A closer examination of participants' error acceptance revealed a significant impact of different errors types on senders' acceptance ($F_{4,80}$=3.60, p=.010) and receivers' acceptance ($F_{4,250}$=8.92, p<.001), as indicated in Figure 2b and 2c. Among these seeded errors (see definitions of error types in "Quantitative Measures"), senders showed very low tolerance for errors in the requested actions and person's names. Receivers indicated a different pattern in acceptance, where they

showed much higher tolerance for general informative messages regarding upcoming events and occasions.

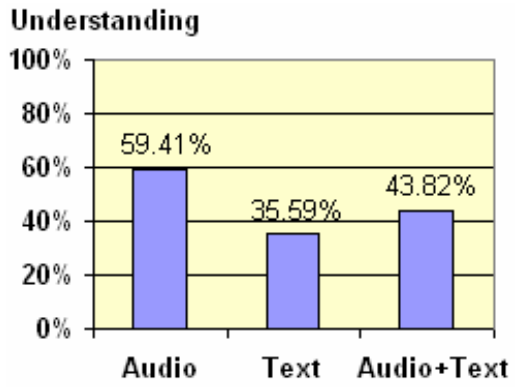

Figure 3.
Understanding vs. Presentation Modes

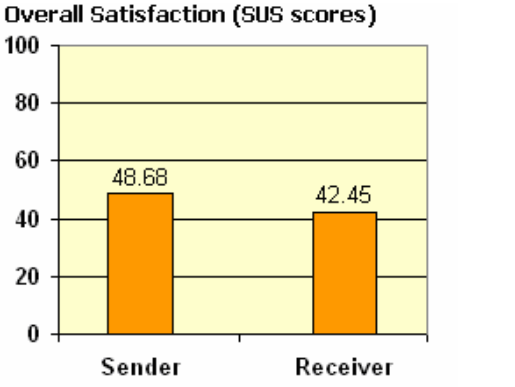

Figure 4.
Overall Satisfaction

Receivers' understanding of the misrecognized text messages was measured by the number of corrected errors divided by the number of total errors contained in each message. The results of ANOVA indicates the audio presentation did significantly improve receivers' understanding of the misrecognized text messages ($F_{2,48}$=10.33, p<.001), as shown in Figure. 3. The findings are consistent with many research studies in cognitive psychology, as discussed previously. Some interesting findings about how receivers managed to understand text messages with recognition errors were revealed during the debrief, which are discussed in details in the next section.

System Usability Scores (SUS) were used in this study to examine users' overall satisfaction with their task experience. As we expected, senders reported slightly higher satisfaction than receivers. We believe 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 statistically significant ($F_{1,66}$=1.06, p=.308). Overall, participants were not satisfied ($mean_{sender}$=48.68 and $mean_{receiveer}$=42.45 on the scale of 1~100) about the voice recognition accuracy demonstrated in this study. They were particularly concerned about the potential cost associated with the miscommunication caused by speech recognition errors.

5.2 Qualitative Findings

To better understand the reasons behind participants' performance and satisfaction, we explored their perception in detail during the one-on-one interview. An affinity diagram was used in the data analysis of user's comments. Over 1200 incidents were collected and clustered. We summarize the qualitative findings in the following four categories: (1) Why are users frustrated? (2) How do users decipher the errors? (3) When are errors more acceptable or unacceptable? And (4) What would users do to use voice recognition more effectively?

Why are users frustrated?

In general, participants did not like the fact that they had to guess or decipher a received message. Some said, "Even if I did understand the message correctly, I don't like spending time to listen to it again and again. I don't like guessing, I would rather the message to be accurate and clear."

Many participants indicated that there could be consequences of misunderstanding the received messages. For examples, "If the sender was expecting me to do something, but I didn't understand the message and did the wrong thing or did nothing at all, that will be a big problem." "Key information (location, time, date, subject, etc.) must be crystal clear, especially if the message is a direction (regarding where to go, what to do, etc.) or a question for the receiver." "Short messages require higher accuracy because concise information is hard to decode." If the receiver is confused by the errors in the message, it may trigger a series of back and forth messages or follow-up phone calls for clarification, which defeats the purpose of communicating via quick and short text messages. As summarized by several participants, "If the recognition accuracy is this low, I would either call, leave a voice message, or manually text the message (depending on how I can get the receiver sooner) to make sure the information is delivered accurately in the first shot." Things that contribute to users' frustration include:

- Being confused by the wrong name
- Cost of not taking the right action as requested
- Cost of getting to the wrong place at the wrong time, or taking too much stress to get there
- Fear of miscommunication
  - Sender's fear of the message being misunderstood
  - Receiver's fear of not being able to clarify the message with the sender
- Effort needed for re-communication
- Effort needed for decoding the errors
- Input frustration (entry and correction)

How do users decipher the errors?

Despite the frustration discussed above, receivers' understanding performance (see Figure 3) indicates that about half of the seeded errors were correctly identified in this study. How did they do it? Various decoding techniques were revealed from participants' comments, as summarized below:

(a) Using common sense
   - Certain associated concepts, or context, may help receivers decipher (e.g., "hospital"-"accident"; "cancel"-"appointment"; "buy"-"tickets"; "lunch together"-"restaurant name"; "baseball game cancelled"-"thunderstorm"; etc.)
   - Basic knowledge of language / grammars also helps receivers identify errors in the message (e.g., "she's seen town" → "she's in town"; "a bit of picture" → "a beautiful picture"; etc.)

(b) Background knowledge of the message

In order to objectively measure participants' under-standing of the misrecognized messages, we did not provide any information about the context of the message in this study. However, participants mentioned that certain before-hand knowledge of the context of the received message was available in real life, which could have made the deciphering easier. Typically, the sender and receiver would have mutual awareness of things such as birthday party, wedding ceremony, where to have lunch together, which train station or airport to pick up a person, etc.

(c) The familiarity between the sender and receiver

The understanding of certain habits or language patterns between the sender and the receiver plays a role in deciphering as well. For example, a young lady mentioned that if she got a message from her mom asking "are you good?", she knew the question was actually "are you

home?" (Note: "home" and "good" have the same key entry as "4663" on keypad. The difference between textual errors and speech recognition errors is discussed below.)

(d) Textual errors vs. speech recognition errors

Typos, misspelled or abbreviated words are easier to recognize than the voice recognition errors. Many people are experienced in decoding the typos in emails or text messages, but very few will try to decipher a word according to how it sounds like. (Examples that can be decoded with previous text messaging skills include "come from"→"confirm" and "ticks"→"tickets". However, similar skills did not work for ASR errors like "a huge end"→"agenda" and "that free meal"→"fat free milk".) Therefore, if the recipient knew the text message had been composed with voice dictation, she might be able to decipher the message more effectively.

When are errors acceptable or unacceptable?

We reported that different error types had an impact on users' acceptance of recognition errors in messaging. An in-depth discussion during the interview revealed the following factors that may also influence users' error acceptance:

(a) The key information in a message

- Errors in information such as name, address, time, and action would change the meaning of a message, and were considered as serious.
- Errors were not acceptable if they allowed ambiguous interpretation of the message.
- Descriptive words or emphasis in the message were less important than the key information. Errors in these words might affect receivers' perception, but wouldn't mislead them into wrong actions.

(b) Situation dependent

- There were split opinions regarding errors in urgent messages. Some participants would send the message as soon as possible to get the words out. Others would avoid any potential confusion if the message was critical.
- Senders indicated higher acceptance if she was in a rush or an eyes-busy hands-busy situation such as driving or multitasking.
- Errors are more acceptable if the receiver and the sender can clarify it afterwards.
- Senders may send the message if they believe the receiver can probably understand it. A participant said that if he was in a hurry, he would send the misrecognized message to his teenager son, but not to his wife. Because his son has developed superior decoding skills over the years of text messaging experience.

(c) Message types

For an informative message, participants showed higher acceptance if they believed the receiver would get the gist. However, very low acceptance was given to messages that are business related, messages that request actions or solicit information from the receiver. Participants would rarely send business messages with errors due to their concerns of reputation and professionism.

(d) Personal traits

We also found that participants' error acceptance was highly related to individual's personal traits. Some interesting findings include:

- Participants who were experienced in text messaging indicated higher error acceptance, although their previous deciphering skills did not help them better identify the speech recognition errors in our study.

- Some participants were confident about their own decoding skills as a receiver, but showed lower error acceptance as a sender. They did not believe other people would understand the misrecognized messages.
- Self-claimed professionals or perfectionists reported lowest error tolerance.

What would users do to use VR more effectively?

Some interesting behaviors were observed during participants' interaction with the dictation prototype in this experiment. Although a high accuracy (95% or above) was anticipated by most participants for them to adopt VR as a primary text entry method, they were willing to take additional effort in order to use dictation more effectively. Some observations and comments include:

- If the dictation did not recognize a particular word correctly, users would try to avoid using (by either substituting or removing) this word in the future.
- If the recognition was not accurate for the first time, users would speak much slowly for the second time, assuming that would help the system understand it more easily.
- For certain names that the system always failed to recognize, users would like to manually save these frequently used names in the system or train the VR system in advance.
- Keep messages short to avoid misrecognition.
- To correct errors efficiently, users would manually select and type over the errors if there were a few errors, otherwise, they would like to re-dictate the entire message.

6. Conclusions

By evaluating our proposed idea of a dictation design of text messaging, we investigated users' perception and acceptance of speech recognition errors in mobile communication. Our findings indicated that an audio readout significantly improved users' understanding, and the "audio + text" presentation was preferred by most of the users. Users showed overall low acceptance for errors in text messaging. Their main concern was the cost of misunderstanding: a confusing message may trigger a series of follow-up phone calls, which defeats the purpose of quick communication via text messaging. This explains why users showed much lower tolerance for errors in messages that request 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 VR to dictate text messages for its convenience and safety concerns, they preferred to correct errors before sending the messages to ensure clear and efficient communications.

The potential benefits of dictating text messages on mobile devices include (1) helping users stay connected via mobile communications in hands-busy and eyes-busy situations; and (2) reducing the transmission cost by sending texts instead of voice attachments. Based on the understanding of users' acceptance and reaction to recognition errors in 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. Findings of this experiment could be applied in the multimodal interface design to facilitate information input on mobile devices. However, this study is only the 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

Ainsworth, W. A. and Pratt, S. R. (1992). Feedback strategies for error correction in speech recognition systems. *International Journal of Man-Machine Studies*, 36(6), pp.833-842.

Alewine, N., Ruback, H., and Deligne, S. (2004). Pervasive Speech Recognition. *IEEE Pervasive Computing*, 3(4), pp.78-81.

Baber, C. and Hone, K. S. (1993). Modeling error recovery and repair in automatic speech recognition. *International Journal of Man-Machine Study*, 39(3), pp.495–515.

Bell, N. (1991). Gestalt imagery: A critical factor in language comprehension. *Annals of Dyslexia*, 41, pp.246-260.

Bradford, J. H. (1995). The human factors of speech-based interfaces: a research agenda. *ACM SIGCHI Bulletin*, 27(2), pp.61-67.

Brinton, B., Fujiki, M., and Sonnenberg, E. A. (1988). Responses to requests for clarification in linguistically normal and language-impaired children in conversation. *J. Sp. Hrg. Dis*. 53, pp.383–391.

Chalfonte, B., Fish, R., and Kraut, R. (1991). Expressive richness: a comparison of speech and text as media for revision. In *Proceedings of CHI'91*, pp.21-26.

Chen, X. and Tremaine, M. (2006). Patterns of multimodal input usage on non-visual information navigation. In *Proceedings of HICSS'06*, Track 6.

Consumer Demand for Speech Solutions on Mobile Phones, Available at http://www.nuance.com/unlockthepower/.

Danninger, M., Flaherty, G., Bernardin, K., Ekenel, H., Khler, T., Malkin, R., Stiefelhagen, R., and Waibel, A. (2005). The connector - facilitating context-aware communication. In *Proceedings of the International Conference on Mumtimodal Interfaces*, Trento, Italy.

Frost, R. (1995). Phonological computation and missing vowels: Mapping lexical involvement in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, pp.398-408.

Frost, R. and Kampf, M. (1993). Phonetic recoding of phonological ambiguous printed words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, pp.23-33.

Intille, S. S., Bao, L., Tapia, E. M., and Rondoni, J. (2004). Acquiring in situ training data for context-aware ubiquitous computing applications. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp.1-8.

Larson, K. and Mowatt, D. (2003). Speech error correction: the story of the alternates list. International *Journal of Speech Technology*, 6(2), pp.183-194.

Lesch, M. F. and Pollatsek, A. (1993). Automatic access of semantic information by phonological codes in visual words recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, pp.285-294.

Lesch, M. F. and Pollatsek, A. (1998). Evidence for the use of assembly phonology in accessing the meaning of printed words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, pp.573-592.

Lieberman, H., Faaborg, A., Daher, W., and Espinosa, J. (2005). How to wreck a nice beach you sing calm incense. In *Proceedings of Intelligent User Interface'05*, pp.278-280.

Lukatela, G. & Turvey, M. T. (1994). Visual lexical access is initially phonological: 1. Evidence form associative priming by words, homophones, and pseudohomophones. *Journal of Experimental Psychology: General*, 123, pp.107-128.

Luo, C. R., Johnson, R. A., and Gallo, D. A. (1998). Automatic activation of phonological information in reading: Evidence from the semantic relatedness decision task. *Memory & Cognition*, 26, pp.833-843.

Mankoff, J. and Abowd, G. D. (1999). Error correction techniques for handwriting, speech and other ambiguous or error prone systems. *GVU Technical Report Number: GIT-GVU-99-18*.

McTear, A., O'Neill, I., Hanna, P., and Liu, X. (2005). Handling errors and determining confirmation strategies — An object-based. *Speech Communication*, 45(3), pp.249-269.

Marcussen, C. H. (2003). Mobile Phones, WAP and the Internet. - The European Market and Usage. Available at http://www.rcb.dk/uk/staff/chm/wap.htm.

Mantoro, T. and Johnson, C. (2003). Location history in low-cost context awareness environment. In: Johnson, C., Montague, P., and Steketee, C., eds. *Proceedings of the Australasian Information Security Workshop Conference on ACSW Frontiers*, 21, pp. 153-158.

Murray, A. C., Frankish, C. R., and Jones, D. M. (1993). Data-entry by voice: Facilitating correction of misrecognitions. In *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers*, C. Baber and J. M. Noyes, Eds. Taylor and Francis, Inc., Bristol, PA, pp.137–144.

Oviatt, S. and VanGent, R. (1996). Error resolution during multimodal human-computer interaction. In *Proceedings of International Conference on Spoken Language Process (ICSLP'96)*, pp.204-207.

Oviatt, S. (1999). Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the International Conference on Computer-Human Interaction*, pp.576–583.

Prodanov, P. and Drygajlo, A. (2005). Bayesian networks based multi-modality fusion for error handling in human–robot dialogues under noisy conditions. *Speech Communication*, 45(3), pp.231-248.

Robbe, S., Carbonell, N., and Valot, C. (1994). Towards usable multimodal command languages: Definition and ergonomic assessment of constraints on users' spontaneous speech and gestures. In *Proceedings of the International Conference on Spoken Language Processing*, pp.1655–1658.

Sturm, J. and Boves, L. (2005). Effective error recovery strategies for multimodal form-filling applications. *Speech Communication*, 45(3), pp.289-303.

Suhm, B., Myers, B., and Waibel, A. (1999). Model-based and empirical evaluation of multimodal interactive error correction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp.589-591.

Suhm, B., Myers, B., and Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction*, 8(1), pp.60-98.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18, pp.645-660.

Tan, L. H. and Perfetti, C. A. (1999). Phonological activation in visual identification of Chinse two-character words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(2), pp.382-393.

Torres, F., Hurtado,L. F., García, F., Sanchis, E., and Segarra, E. (2005). Error handling in a stochastic dialog system through confidence measures. *Speech Communication*, 45(3), pp.211-229.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling , sound, and reading. *Memory & Cognition*, 15, 181-198.

Van Orden, G. C. and Goldinger, S. D. (1994). Interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1269-1291.

Ziegler, J., and Jacobs, A. (1995). Phonological information provides early sources of constraint in the processing of letter strings. *Journal of Memory and Language*, 34, pp.567-593.