SPEECH ASSISTED MOBILE TEXT ENTRY

Sami Ronkainen *Jonna Häkkilä* *René Hexel*

Nokia Corporation
P.O.Box 300, SF-90401 Oulu, Finland
{sami.ronkainen,jonna.hakkila}@nokia.com

Griffith University
Nathan QLD 4111, Australia
r.hexel@griffith.edu.au

ABSTRACT

The ever-increasing number of mobile applications increases also the likelihood that these applications need to be used in contexts where usage of traditional visual-manual user interfaces is difficult. In this paper, a multimodal user interface utilising manual input, and both spoken and visual output for entering text is presented. Results from user tests show that in stationary usage context, the spoken output is not helpful, but disturbs text production. In a mobile context, novice users were not able to take full advantage of the spoken output. An expert user could type text rather quickly, even blindfolded, and the comments from novice users showed that the potential for assisting in difficult to use contexts was recognised in the concept.

1. INTRODUCTION

The usage of mobile phones has become an integrated part of every day life in our society. In addition to calling, communication with text messaging has gained enormous popularity, and for many users, this is a significant way of interaction [1]. As mobile devices, mobile phones are used in various different kinds of situations, where a user's attention may not be fully focussed on the device. Such multitasking situations may appear, for instance, when walking in a crowd and using the phone simultaneously. Mobile phone usage while driving a car has already caused active discussion and legal actions, as drivers concentrating on phone usage have created risks in traffic [2]. Even without going to such extremes, there are numerous examples, where mobile phone usage is inconvenient in multitasking situations and limits the efficiency of all the tasks that should be performed simultaneously.

With mobile handheld devices, text entry is challenging due to both varying usage situations and limited device size. Several studies have been made on mobile phone text entry with different keypad solutions (see e.g. [3]) or combining different touch based techniques, such as pen input and key strokes [4]. Text input with using a gestural interface has been investigated by Wigdor and Balakrishnan [5], where text is entered by tilting the device.

Our research combines different modalities, when auditory feedback is used together with a conventional mobile phone keyboard and display setup. Previous research on mobile handheld devices has considered the use of audio modality as an assisting technology for navigation. In [6], different non-speech audio cues are identified with distinct menu selections to support the navigation. In a simple form, auditory feedback is already used in mobile devices also, for example, for keypad tones. We employ additional auditory aid from key presses to assist the userin text entry.

In the following, we first describe the application design, then its evaluation through user tests, and finally we analyse and discuss the experiment results.

2. CONCEPT DESCRIPTION

Text entry in mobile phones is commonly arranged by utilising so-called multi-tap text entry on a standard telephone keypad [7], where each number key is associated with one or more symbols or letters. For instance, key '2' could be associated with letters 'a', 'b' and 'c' (see Figure 1). In order to create, e.g., the letter 'c', a user must press key '2' on the keypad three times. In order to enter successive letters from the same key – e.g. two letters 'c' in a row, or an 'a' followed by a 'b' – the keys need to have different input modes: one for selecting the desired one from the available letters associated with that key, another one for entering successive characters from the same key. In many implementations, the mode change is automatic. There is a time period, during which the successive key presses are interpreted such that the user wants to change the subsequent letter produced by that key, e.g. switching from 'b' to 'c'. After a time-out, when no user interaction has occurred during the time period, the next key press from the same key is interpreted as the user wanting to enter a new letter from the same key, e.g. a 'b' followed by a 'c'. This mode change is usually indicated with some visual cue – e.g. through the appearance of a visual cursor on Nokia's mobile phones.

Figure 1. *The Nokia 6210 mobile phone keypad used in the test.*

The described mode switching requires constant visual monitoring from the user. The task of text entry currently requires a lot of visual attention simply because there is only a visual presentation of the letters the user has typed.

However, the text entry task needs not be as restricted to the visual modality as it currently is. The keypad also provides, if designed to accommodate also the needs of visually impaired

users, tactile cues about the placement of different keys [9]. This combined with an appropriate auditory feedback could provide users with the means for entering text when the visual modality is not fully available, which often is the case in a mobile usage context.

Ronkainen and Marila have studied an auditory indication of the automatic mode switch [8]. In the concept presented in this paper, the idea was taken further by generating spoken feedback from every key press when entering text. For instance, pressing the '2' key three times in order to enter letter 'c', caused the system to speak out: "a, b, c". The occurrence of the time-out that changed the input mode (interpretation of successive same key presses) was indicated by an audible signal consisting of two 10 ms beeps of 659Hz, separated by a 45ms silence, in a similar manner as in the previous study [8].

The visual presentation on the phone screen was untouched, i.e. it was the same as in an off-the-shelf phone.

The rationale behind the design was that in a mobile usage situation, users are often restricted in the way they can utilise their vision for using a mobile device – but they are not completely unable to look at the device. For instance, when walking, one usually needs to monitor the surrounding environment, but one also has the time to occasionally glance at the mobile device.

The advantage in the concept was presumed to be that it is possible to type when not looking at the screen, allowing the user to look around when, e.g., walking and typing simultaneously. User tests were arranged to evaluate the application. Adding auditory feedback to text input was assumed to enhance the possibility to focus on other tasks than writing. Another hypothesis was that the feedback would increase the efficiency in multitasking situations.

The concept is not without downsides, however. It is previously known from cognitive psychology that when creating text, the inner speech of a human being plays an important role. It is also known that the inner speech utilises the same part of working memory – the phonological loop – when speech is heard [10]. Therefore, it was presumed that hearing spoken letters continuously while typing might be disturbing. The extent of disturbance should appear from results of user testing.

3. USER TESTING

3.1. Participants

The concept was tested with 11 participants (Finnish university students from different fields, aged 20-29, 6 male, 5 female) who were all experienced in creating text messages on a mobile phone. Every user's native language was Finnish. The concept was new to all participants. For comparison, the concept was also tested on one expert user who had had a chance to practice and learn the concept thoroughly.

All subjects were owners of mobile phones and were active users of the text messaging application. Two subjects reported to write a text message almost every day, eight participants reported 1-5 and one over 10 text messages a day.

3.2. Test setup and measurements

The concept was implemented on a Nokia 6210 mobile phone connected to a PC with a serial cable. On the PC, a piece of software kept track of all key presses the participants made. The key presses were time labelled to an accuracy of 100ms. The PC software also created the auditory feedback, speaking out the letter related to each key press and playing the beeps, indicating the input mode change. The speech that was used was sampled from a live speaker. Finnish pronunciation of the letters was used. The feedback was played through a pair of headphones.

At first, all users were introduced to the test setup and the tested concept. Before the actual test, they tried out typing some text using the concept in order to familiarise themselves with it.

The actual test consisted of two phases. The first phase was a typing test in quiet laboratory conditions. In that phase, all participants typed the first strophe of the Finnish national anthem, which was assumed to be familiar to all participants. The length of the strophe was 139 characters (including spaces), but the actual number of entered characters varied slightly, depending on the error correction the participants made. 6 participants typed the text first with no auditory feedback, 5 started with typing with the auditory feedback turned on. Then the test was repeated, with the feedback conditions reversed. This was repeated twice. The participants were instructed to correct errors they noticed right away, but to not go back in the text to correct errors they had not noticed immediately. The typing speed (characters per second, CPS) and the percentage of corrected and remaining errors were measured.

The second phase was carried out so that the participants were typing the same text as in the first part, now carrying the laptop PC in a bag. While typing, the participants were also walking along a predefined route, which included going through a corridor followed by stairs leading down to a lower level, and the same route back. Before any typing task, the users walked the route without typing, and the time was measured. This was done in order to familiarise the participants with the route, and to get a nominal walking speed preferred by each user, to which we could compare the speed of walking while typing simultaneously. A similar metric (PPWS, Percentage Preferred Walking Speed) for evaluating mobile user interfaces has been utilized also previous research, e.g. by Pirhonen and Brewster on the design of a gestural interface for a portable music player [11]. After this, all participants performed the task of simultaneous typing and walking the route. Just as in the first part, half of the group started with the auditory feedback turned on, the other half with the feedback off. The conditions were then reversed and the task was repeated. This was done twice, so in total the participants performed the task four times, twice with each feedback condition.

The number of characters entered in each condition was whatever the participant had managed to type during walking the route. The number of characters produced was not the same for all participants, as it depended on the typing speed of each participant. These differences were compensated for in the calculation of the results, as is explained in chapter 3.3. Again, the typing speed and percentages of corrected and remaining errors were also measured. In addition, the time to walk the route was measured.

The phone was equipped with a miniature camera focussed on the user's face (see Figure 2). From the video, the number of times the participants glanced away from the phone screen was

counted. The rationale was that the more often the user can look away from the display while typing, the more benefit the concept yields.

After user testing, subjects filled out a written questionnaire where they commented on the concept.

The expert user performed tasks only in the lab – i.e. no walking while typing task was performed. Additionally, the user typed the text completely blindfolded, relying only on the auditory feedback and feeling the keys with his fingers.

Figure 2. *The test equipment: mobile phone with a miniature video camera facing the user, laptop computer in bag, headphones.*

3.3. Analysis

In the first phase, when typing text in a stationary laboratory context, the typing speed, and amount of corrected and remaining errors in both conditions (with and without auditory feedback) were compared. These results were then used as the nominal performance, to which we compared the typing performance in the walking condition.

In the second phase, the differences in walking times between the nominal condition, and each typing-while-walking condition was calculated for each user. Also, the difference in typing speed and accuracy compared to each user's performance in the first test phase was calculated.

Naturally, the participants could spend more effort either on the walking task (walking faster, producing less text) or on the typing task (walking slower, producing more text). To compensate for these differences, the typing speed in the walking tasks was calculated as the number of characters produced, divided by the time spent walking. This, combined with the fact that the participants were instructed that both tasks were equally important, was trusted to produce comparable results.

In the second phase, the number of times each user glanced away from the phone display was compared between feedback conditions.

4. RESULTS

4.1. Measured data

In the laboratory conditions, there was a clear difference in typing speed between the silent and auditory feedback conditions. Typing speed with auditory feedback was, on average, 0.93 characters per second, which is 88% of the typing speed without auditory feedback (avg. 1.05 CPS). The difference is statistically significant (two-tailed Student's t-test, $p<0.05$). This confirms the presumption that the spoken feedback from key presses disturbs the task of language creation. In error percentages there were no clear differences (6.8% in the quiet condition vs. 6.7% in the auditory feedback condition)

In the walking task, the typing speed difference vanished. In fact, the average typing speed while walking with the auditory feedback condition (avg. 0.93 CPS) was slightly higher than with the silent condition (avg. 0.91 CPS) but this difference is not statistically significant. An interesting finding, however, is that when walking and typing with no auditory feedback, the typing speed dropped to 87% compared to the laboratory condition. This result was statistically significant ($p<0.01$). But with the auditory feedback turned on, the typing speed was almost exactly the same as in the lab (99% CPS compared to lab). This implies that the disturbances caused by the auditory feedback and the multitasking situation did not cumulate.

The differences between auditory and silent feedback conditions, when considering the percentage of total typing errors (corrected and remaining), was not near statistical significance in any of the tested conditions. However, the number of remaining errors in the walking conditions was interesting. The average percentage of remaining errors was 0.99% in the walking without auditory feedback condition. In the walking with auditory feedback situation, the average percentatge of remaining errors was 0.55%. This result, while not being statistically significant (two-tailed Student's t-test, $p=0.16$), still hints that auditory feedback might have allowed the users to better notice the errors they made in typing while walking.

The expert user reached a typing speed of 2.3 CPS in a laboratory setting, both with and without auditory feedback. In other words, the disturbance resulting from the spoken feedback can be overcome, but not without increased cognitive effort, as reported by the user. When completely blindfolded, the expert user reached a typing speed of 1.6 CPS, which is 70% of the nominal performance. The main problem reported by that user was being able to simultaneously keep track of what has been typed so far, which is the next word to be typed, and which letter to type next. It must also be kept in mind that this result was gained in a laboratory setting. With external disturbance resulting e.g. from walking in traffic, it is unlikely that the concept would work satisfactorily completely without display – which on the other hand was not an objective for the concept in the first place.

The feedback type did not affect the walking speed. The average walking time spent in typing while walking was 150% of the nominal time (walking without typing) in both cases with and without the auditory feedback. This was probably due to the fact that during the test, the participants did not learn to take

advantage of the provided feedback, but spent almost all of their time staring at the display. This may also partially be due to the fact that none of the test participants used exactly the same phone model in their everyday lives as the one tested, making it more difficult to locate the keys by touch alone.

The number of glances away from the display during one walking task varied from zero (even though most of the route took place on staircases!) to 18. No consistent difference between the number of glances with different feedback types was observed, with the exception of two users, who also reported after the test that they had consciously tried if they could walk and type without looking at the display. In those cases the difference was clear: (maximum differences: 11 glances vs. 3 glances with user #2, 12 glances vs. 1 glance with user #11).

The findings (average numbers of typing speed in percentage of the preferred walking speed, percentages of remaining and total number of typing errors, walking time in percentage of the time spent in the nominal case, and the number of glances away from the display) are presented in Table 1.

Data / FB type	Typing speed (%)	Rem. Errors (%)	Tot. Errors (%)	Walk time (%)	Glances away from display
Silent	87.0	0.99	7.8	149.5	5
Speech	88.2	0.55	7.1	150.2	6.7

Table 1: Findings (% of nominal performance, or error % in case of typing errors) in different feedback (FB) types.

4.2. User feedback

The first feedback coming forward concerned the disturbing or irritating effect of the audio feedback. Altogether seven subjects (#1 #3 #4 #5 #8 #9 #10) commented that they felt audio intrusive, especially in the beginning. Despite of this impression, seven participants (#1 #2 #3 #6 #7 #10 #11) commented that they believed the usability of the application would improve and benefits rise if they would get more familiar with the application. *"Auditory feedback felt bad when I didn't walk. To my surprise, it didn't disturb the writing so badly when I was walking. However, I felt it (generally) easier to write without audio feedback. I believe this however was caused mainly because I wasn't used to it."* (participant #1).

All participants found general positive sides or potential usage situations where they felt the feature might be useful. The usefulness of the application considered more flexible eye contact and focus of attention. A comment from participant #7 represents a typical answer: *"The phone could be used more easily in situations, where one cannot look at it, but usually I don't have a need for that."*

Participant #11 commented that she would actively study to use the application if it was implemented in her own mobile phone, as she saw potential advantages in it.

Generally, all participants found that the application would benefit users in situations, where eye contact or focus of the attention could not fully be on the phone. In addition, two participants (#4 #7) suggested the application could help visually impaired users. Although the participants saw that the feature had advantages, only five (#2 #3 #8 #10 #11) reported situations where they might use it. *"[I would use the feature] in situations where my sight should concentrate to something else, but first I should get familiar with the application"* (#10) and *"for example in city or in traffic, where one has to monitor also other things than the phone. In all situations, where eyes should be 'free'."* (#3).

Despite of the low feedback on this, eight participants reported that they would welcome such a feature on their phone if they could turn if on/off, even when two (#3 #6) added that they were not sure if they would ever use it and two (#5 #7) commented that they would not pay any extra for it.

Six participants suggested a limited or modified implementation of the auditory feedback, where the feature would be applied only to some specific function. Participants #2 and #11 wished that audio feedback could be combined with predictive text input (T9). Similar feedback was received from users #4 and #8, who commented that not all characters needed audio feedback and that repeating a syllable or word would be better. Participant #9 wanted to use just an audio signal for indicating the end of the time-out in key mode change, i.e., when writing two letters in succession from the same key. This is consistent with the user comments collected in a previous study [8]. Participant #1 suggested that audio output would be useful when browsing the names in the device phonebook.

Five participants commented on the need of earphones (#2 #3 #7 #8 #11), which might limit the use of the application.

5. DISCUSSION

The tests show that the concept where spoken auditory feedback is added to mobile phone text input does not show any immediate enhancement for novice users, neither in text input efficiency nor in sharing the focus of attention in multitasking situations. The number of remaining, uncorrected errors in the text seemed to be smaller, though, with auditory feedback, but the difference did not reach a level of clear statistical significance. This may be due to the small number of participants. A larger number of subjects would need to be tested to clarify this.

Although the statistics don't show any clear advantage, qualitative user feedback is quite encouraging and suggests that there is definite potential to enhance the usability with this kind of application. Auditory feedback requires learning, though, and presumes that the user is accustomed with the feature.

This conclusion can be drawn from the general comments, as all participants commented that the feature could be useful even when they did not identify a specific situation where it would benefit them. The findings suggest that the application should be tested for a longer time period and with a larger audience to find out true use cases and potential advantages in real life use situations. A longer test period would also offer a change for the user to get over the difficulties that arise from the unfamiliarity with the feature, which was reported to be disturbing in testing situations.

As the spoken feedback was found to be disruptive for the typing task, it should be considered if the feedback from key presses could be modified, e.g., into consisting of both or either of speech and non-speech audio. For instance, every key press could produce just a beep (possibly a different beep for each key

press or character). Alternatively, full words could be read out. It should also be studied whether synthetic speech, or speech processed e.g. by removing all intonation and possibly making it more machine-like would be as intrusive as speech sampled from a live speaker.

It must also be noticed that the typing task requires the mental creation of language. The disturbing effect of speech output most likely results from that. For other tasks – e.g. navigation – speech most probably will not be as disruptive.

6. CONCLUSIONS

In this paper we described a concept where text entry on a mobile phone is assisted with spoken feedback from key presses, and a non-speech auditory signal of key input mode changes. The concept was tested with 11 participants novice to the concept, and one participant with a lot of practice with it. The tests with novice users were carried out both in laboratory and mobile usage contexts.

In laboratory conditions the concept of providing spoken feedback from key presses while typing test proved to be disruptive, as was expected.

Disruptions to typing speed resulting on one hand from the spoken feedback and on the other hand from a mobile usage context did not cumulate. Typing speed without auditory feedback was significantly lower in the mobile context than in laboratory conditions. With auditory feedback the typing speed stayed virtually the same in both conditions.

An expert user reached the same typing speed in the laboratory with the auditory feedback as without it. However, the disturbance caused by the spoken feedback did result in an increased cognitive effort required. An expert user also managed to type with the concept with no visual feedback at all, but at a reduced typing speed. The biggest potential for the concept therefore lies in a situation where the usage of eyes is restricted but not completely impossible.

In a mobile context while typing and walking simultaneously, novice users did not learn to take advantage of the concept. However, most of the participants recognised the potential of the concept for such situations.

As proven by the expert user, the concept can be used also without looking at the display, but how much cognitive potential it leaves for e.g. monitoring the surroundings remains to be verified.

The large number of suggestions from the test participants indicates that the concept should be developed further to chart the most potential form for the auditory feedback. The availability of the feature would enlarge the number of possible usage situations. For instance cases such as writing a text message when the phone is in a pocket and typing is done completely out of sight are currently not possible.

7. REFERENCES

[1] R. E. Grinter and M. Eldridge, "Wan2tlk?: Everyday Text Messaging," *CHI letters,* vol. 5, iss. 1, pp. 441-448, Apr 2003.

[2] D. A. Redelmeier and R. J. Tibshirani, "Association Between Cellular-Telephone Calls and Motor Vehicle Collisions," *The New England Journal of Medicine,* vol. 336, no. 7, pp. 453-458, Feb 1997.

[3] M. Silfverberg, I. S. MacKenzie and P. Korhonen, "Predicting Text Entry Speed on Mobile Phones", *CHI Letters*, vol. 2, iss. 1, pp. 9-16, 2000.

[4] L. K. Seng, "Hybrid Stroke/Vowel Input System for Mobile Devices", *in Proceedings of OZCHI 2003.*

[5] D. Wigdor and R. Balakrishnan, "TiltText: Using Tilt for Text Input to Mobile Phones", in *Proceedings of UIST 2003.*

[6] G. Leplatre and S. A. Brewster, "Designing Non-Speech Sounds to Support Navigation in Mobile Phone Menus" in *Proceedings of ICAD 2000,* pp. 190-199.

[7] -, "Human Factors (HF); Assignment of alphabetic letters to digits on standard telephone keypad arrays", ETS 300640, European Telecommunications Standards Institute, August 1996

[8] S.Ronkainen and J.Marila, "Effects Of Auditory Feedback On Multitap Text Input Using Standard Telephone Keypad", *Proceedings of 8th International Conference on Auditory Display (ICAD)*, July 2-5 2002, pp. 125-129

[9] - " Human Factors (HF); Telecommunications keypads and keyboards; Tactile identifiers", ES 201 381 V1.1.1, 1998

[10] M. W. Eysenck, M. T. Keane, *Cognitive Psychology - A Student's Handbook.* Psychology Press, United Kingdom, 2000, p.157, p. 381.

[11] A. Pirhonen, S.A. Brewster, C. Holguin, "Gestural and Audio Metaphors as a Means of Control for Mobile Devices". In *Proceedings of ACM CHI2002,* Minneapolis, MN, April 20-25 2002, ACM Press Addison-Wesley, pp 291-298.