

Investigating the Use of Voice and Ink for Mobile Micronote Capture

Adrienne H. Andrew¹, Amy K. Karlson², and A.J. Bernheim Brush²

¹ Computer Science and Engineering, University of Washington, Seattle WA 98195 USA

² Microsoft Research, One Microsoft Way, Redmond WA 98052 USA
aha@cs.washington.edu, {karlson,ajbrush}@microsoft.com

Abstract. Despite the potential benefits of digital note taking tools, research has found that people continue to use paper for creating micronotes, informal personal notes such as reminders and to-dos. Design recommendations from formative studies suggest that “natural” input modalities such as voice and digital ink could help to overcome the drawbacks of text entry on phones and PDAs. We conducted an 18-person lab study to understand the perceived and actual trade-offs that these non-traditional input methods offer for micronote capture. We found that people preferred ink (8 participants) and voice (8 participants) input over keyboard (2 participants) input. Half our participants varied the input method they used in different environments, while the rest did not. However, paper remains popular and was preferred by 8 participants when given the option. The 9 participants whose ink and voice micronotes were transcribed with higher error rates had a noticeably different experience using voice including slower capture times, and higher mental and physical demand survey responses. The percentage of participants that preferred ink, voice, and keyboard was the same for both transcription quality groups.

Keywords: Mobile input, voice input, digital ink, micronotes, mobile note taking.

1 Introduction

Despite the large number of technologies that are now available to help us digitally capture and manage small pieces of information, such as to-do lists and phone numbers, people continue to use paper for these notes. Micronotes, a term Lin et al. [1], adopted to “cover the host of personal jottings to ourselves that we all make every day” often manifest themselves on the scraps of paper that fill our desks, purses and bags. Digital versions take the form of a small piece of information emailed to yourself, or typed into Notepad or digital post-its. As such, micronotes (both physical and digital) are frequently distinguished from formal notes or tasks by their inability to fit easily into traditional Personal Information Management (PIM) tools. The result is that micronotes are often lost and difficult to maintain over time.

Researchers have been captivated by the potential benefits that digital capture might offer micronote creators in terms of editing, searching, sharing, and archiving. Several research studies have explored the content and purpose of micronotes [2], the lifecycle of micronotes [1], the management of micronotes [3,4] and how people use

(or do not use) PDAs for capturing and retrieving micronotes [1,4,5,6]. Their results highlight that a key barrier to digitizing micronotes is getting them in digital form in the first place; that is, people find that digital devices, even pervasive ones such as mobile phones, generally do not meet their needs during micronote creation. Design recommendations from these studies suggest numerous ways in which future tools might lower the barrier to digital micronote capture. Of particular interest to us, several studies recommend that support for input methods such as digital ink and voice could be valuable in making micronote capture natural and fast, while automatic transcription of the resulting pen-strokes and audio could assist people in retrieving, organizing and searching their captured notes [1,3,5,6,7].

Inspired by these design recommendations and to complement previous research, which has primarily focused on understanding current micronote taking practices, we conducted a controlled laboratory study to lend formal insight into micronote capture. Specifically, we were interested in understanding what factors might influence people's choice of input modality for capturing a micronote on a mobile device that supported digital ink and voice input in addition to standard keyboard entry. Our study was designed to allow us to explore the effect of three factors on participant preference: transcription quality, time to enter a note, and the physical environment. Eighteen participants created micronotes on a phone using digital ink, voice, and a virtual soft keypad for lists of varying length in three different settings: a lab, café, and while walking. The ink and voice micronotes of half our participants were transcribed with error rates similar to current day technology, while the micronotes of the other half were transcribed with near perfect accuracy that might be possible in the future.

Our participants preferred voice (8 participants) and ink (8 participants) over keyboard input (2 participants). Surprisingly, although participants in different transcription conditions did have measurably different experiences, particularly for voice input, the distribution of participant preferences across input modalities was the same in both conditions (e.g. 4 from each condition preferred voice, 4 preferred ink and 1 preferred keyboard). Even after trying the "natural" input modalities of ink and voice, paper remained appealing to 8 participants who ranked it as their most preferred input when given the option. Participant preference did not appear to be related to how fast they were at capturing notes using a particular modality, while the environment did seem to affect which modality some participants selected (e.g., using voice while walking, but using ink or keyboard in the nosy café). However half of our participants did not change their input modality based on the environment suggesting some had a strong personal preference. Our results lend evidence that note taking systems should offer integrated experiences to allow users flexibility in choosing an appropriate input modality based on their preferences and situational needs.

2 Related Work

Lin et al. [1] conducted one of the earliest studies investigating micronotes, identifying the lifecycle of a micronote. This includes trigger, record, transfer, maintain, refer, complete, discard, and archive. In our study we focus on the record stage and how people capture micronotes using mobile devices. Research by Lin and others (e.g., [1,3,5,8]) highlights the importance of being able to quickly access the capture device

(for example, taking a mobile phone out of a pocket), start taking the note, and finish making the note. Several studies (e.g., [3,4,5,6,8]) have also found that some note-taking applications require people to specify detailed information about the note after entering it (e.g., a reminder time, date or importance rating) which are potential barriers to the quick capture of a micronote.

Previous research (e.g., [1,3,5,6]) has suggested that people can more quickly create a micronote using either voice or writing with a pen, rather than, for example, multi-tap on a cell phone. However, voice and ink input both have drawbacks in terms of post-entry consumption and management. For example, voice notes are easy to enter, but time-consuming to listen to. Ink is natural to enter on a touchscreen device, but handwriting can look messy and tends to be large so fewer notes can fit on a screen. Research by [1] and [2] suggests both voice and ink notes can benefit from post-processing that transforms them into digital characters (transcription), which then allows them to be searched, quickly visually inspected, organized and incorporated into reminder systems. As Lin et al. wrote, "In summary, the optimal mobile micronote system combines the ubiquitous convenience of paper, the intuitive writing process of a digital pen, and the computational functionality of a PDA" [1].

Of course mobile micronote tools are being developed. Commercial offerings include tools such as Jott [9], which allows users to call a number and follow structured voice prompts to leave a voice message which is then transcribed and made available to the user, and Microsoft's OneNote Mobile that accepts voice and digital ink input, as well as research prototypes (e.g., [1]). Evernote [10], which supports mobile note-taking using voice, pictures, and text, is one of the most compelling mobile micronote applications, but it does not support ink input or transcribe voice notes. The widespread continuing use of bits of paper, post-it notes and other scraps for many micronotes highlights the challenges of developing a digital micronote application that is widely adopted [3,8]. Given that previous studies have identified quick entry and natural input modalities as important, in this study we sought to better understand in a systematic way people's preference for using ink and voice to capture micronotes.

3 Study

We designed and executed a controlled quantitative study to answer research questions regarding how transcription quality, capture time, and environment affect user choice and preference for using Ink and Voice for entering micronotes on a touchscreen-based mobile device. Our three specific research questions were:

Q1: How accurately can users build a mental model of what input modality is fastest for them? Does transcription quality affect this?

Q2: Does the environment influence the choice of input modality or do participants seem to have a stable personal preference across environments?

Q3: What input modalities do participants prefer? Does transcription quality or speed of capture seem to influence users' preference?

To explore these questions we recruited eighteen participants (9 men, 9 women; ages 17 to 56, mean = 39, median = 46) from the general population to take part in our study in July 2008. We specifically recruited participants who used paper to

capture micronotes at least once a week and owned a standard mobile phone (not a Smartphone or Pocket PC) so that they would not be biased by previous mobile micronote capture experiences. During the study, participants captured micronotes of three different lengths (short, medium, long) using Ink, Voice and a virtual Keyboard. The transcription quality of the Ink and Voice notes was a between-subjects condition. Participants in the CurrentDay (CD) condition received transcriptions with error rates similar to those produced by standard voice and ink recognizers [11,12], while those in the NearPerfect (NP) transcription condition received transcriptions that were close to perfect and represented possible future error rates. We split the men and women in the study across the two transcription conditions as equally as possible (CD: 5M, 4W; NP: 4M, 5W). We now describe VINO, a prototype mobile note taking application we built for our study and then our study procedure.

3.1 VINO Prototype

For the study we built the VINO (Voice and InkNOTes) note taking prototype. The note capture screen shown in Fig. 1a can be reached directly by using the hardware note-taking button or from an initial screen (not shown) listing all previously entered notes. From the capture screen the user can create a note in whatever modality she wishes—either by starting the recorder using the audio recording bar (top), writing in Ink (middle), or typing on the Keyboard (bottom). Once input has begun, the display swaps into either audio-capture mode, Ink-capture mode (Fig. 1b), or typing mode to maximize the space available for capturing the micronote, which is especially valuable during Ink input mode. After an Ink or Voice note has been captured, an editable transcription appears in a new tab (Fig. 1c), allowing the user to toggle between the original Ink or Voice note and its transcription. VINO does not currently support real-time transcription. All transcriptions were pre-computed for each user task because of the need to control for errors in our study, and we incorporated event logging capabilities and study control logic into the prototype.

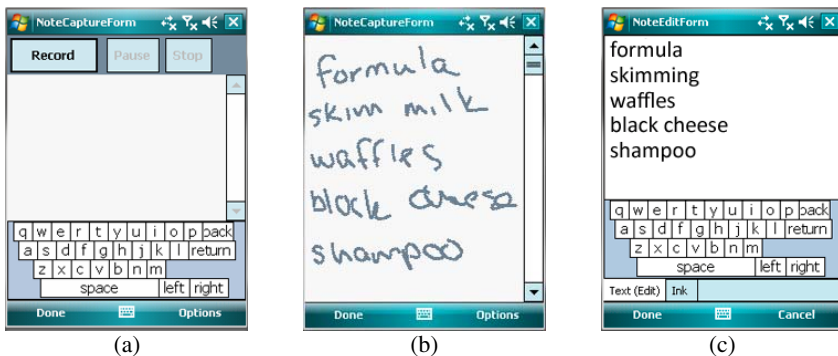


Fig. 1. The VINO note capture software. (a) The note capture screen supports voice (top), ink (middle) or keyboard entry (bottom); (b) ink entry mode; (c) the transcription screen for Voice or Ink notes.

To develop transcriptions for the CurrentDay and NearPerfect transcription conditions, six co-workers (3M, 3F) were recruited to speak and write (with digital Ink) all of the micronotes included in the study tasks. We then used standard Voice and Ink recognizers [11,12] to transcribe this data, and calculated error rates using a version of the Levenshtein distance function [13] modified to take into account the fact that the VINO text editor only allows insertions and deletions, not overwrites. This process helped us determine our error rates for the CurrentDay condition as 0.35 errors per character for Voice and 0.17 for Ink. For the NearPerfect condition, we decided upon error rates for Voice of 0.11 and 0.04 for Ink.

Using these rates, we introduced errors to the transcriptions we gave the participants using misrecognitions from the software-generated transcriptions whenever possible. To control for total number of errors across tasks in each condition, we occasionally created a transcription with the appropriate number of errors that was similar to but not exactly the same as any error generated by the transcription software. In a further attempt at realism, Voice transcriptions were always displayed as a single line, while Ink transcriptions showed the lists as 1, 3, or 5 lines, depending on the task length. Overall, this process allowed us to generate “fake” transcriptions that contained realistic errors, but which also had a controlled error rate. One aspect of errors we did not control for were “meaning-changing” errors. For example, with the micronote “bake cookies,” a transcription of “bak cookie” is likely to be understood by the creator, while “bark copies” might not be. We calculated that 40% of the 168 transcriptions we used had potentially meaning-changing errors. The participants did not communicate any concern or disbelief about the transcriptions.

We built VINO using C# (.NET Compact Framework) and it runs on Windows Mobile PocketPCs—mobile phones with touchscreens that function similar to Personal Digital Assistants (PDAs). We conducted the study using an HTC Touch Cruise, which has a large screen with a flush bezel that is easy to write on, a hardware button that could be used to launch VINO, and a voice recorder. However, because the Touch Cruise lacks a physical keyboard, we built a custom soft keyboard for inclusion in VINO. We implemented our own keyboard to control when and how it was displayed to users and avoid the predictive word choices that the built-in keyboard automatically presents to users as we were concerned it might confuse users and potentially confound our results.

3.2 Procedure

The study was conducted in our lab and took roughly 1.5 hours. The study consisted of five phases:

Current Behavior: We interviewed the participant about her use of micronotes, and discussed the personal micronotes we asked her to bring to the lab.

Training (3 micronotes): We introduced the participant to the VINO software and the protocol that would be used for performing the study tasks by having her enter micronotes using all three input modalities (Ink, Voice, Keyboard) and lengths. We specified the text of each micronote for all trials, which were either Short (S), a 1-item 2-word list, Medium (M), a 3-item 4-word list or Long (L), a 5-item 7-word list. We ensured that all micronotes in each length category had the same number of

characters. Research by Ludford et al. [14] showed that users in their study on location-based reminding created lists of things to do or get most of the time. Therefore, the micronotes we created for the study were lists of things to do (e.g., “wash windows”) or get (e.g., “cloth, bucket, rubber gloves”).

At the start of each task the participant was given a printed card showing the target micronote text, which was also displayed on the screen of the phone. The participant pressed a button on the screen to start the capture process, entered the content into VINO using the designated input method, and then pressed another button to indicate they were finished. The participant was then presented with a transcription of the note (Fig. 1c), which she corrected using the soft keyboard before submitting the final note. The participants were not aware that the transcriptions were pre-computed.

Participants were instructed to correct transcriptions to whatever degree “felt comfortable” in order to make the final transcription capture the essence of the list presented on the printed card. For consistency, we had considered asking participants to correct the transcription to exactly match the list given in the task. However, our two pilot participants reported that it felt arbitrary and contrived. Thus we decided to loosen the correction constraint, as this is closer to real world use where one might ignore a mistranslation that is clearly an error, but is still recognizable.

Input Trials (18 micronotes): The participant captured 2 micronotes for each note length (S, M, L) using each of the three input modalities (Ink, Voice, Keyboard). All 6 tasks for a particular input modality were performed together, but the presentation order of the input modalities was counterbalanced across participants. Tasks were randomized within each input modality. After each input modality, the participant filled out a NASA TLX-based survey about her experience.

Timed Competition (6 micronotes): The participant captured 6 micronotes (2 for each of the 3 note lengths) using whatever input modality she felt would allow her to capture a note fastest. We required that the final note match the task text exactly, and specified that the time to enter the note included the edit time. Participants could choose a different input modality for each of the 6 tasks. To motivate participants to choose the input modality they perceived to be fastest we told them that an additional gratuity would be awarded to the participant with the fastest average capture time.

Environment Phase (12 micronotes): We took the participant to a café located in our building to enter 6 micronotes (2 of each length) using her preferred input method, and then had her walk around the building while entering another 6 micronotes (2 of each length), again using whatever method she wished.

Wrap-up Phase: At the end of the study participants filled out a survey about which input methods they thought were fastest and which they preferred.

4 Results

Our 18 participants (M:9, F:9) ranged in age from 17 to 56 (mean = 39, med. = 46) and had wide variety of occupations including student, artist, fire-fighter, healthcare and technology related (e.g. programmer, IT specialist). Despite this diversity, capturing micronotes was a common occurrence for our participants and twelve told us they

deliberately carried note-taking mechanisms with them. We asked participants how frequently they captured micronotes for a variety of methods with the options of never, once ever, monthly, weekly, daily, and several times per day. For our participants, paper was the most commonly used method for micronotes (med. = 'several times per day'). This was followed by the use of personal computers (med. = 'weekly'), but phone, PDA, and voice recorder all had a median response of 'never.' We were not surprised that our participants did not use their phones for micronote capture, given that we had explicitly recruited participants with standard mobile phones so they would not be biased for or against capturing micronote notes on their mobile phone. However, 17 of the 18 participants used their mobile phones daily or multiple times per day, so we know they are carrying mobile devices with them. We now describe results of our Input Trials, followed by the Timed Competition, capture by participants in different Environments, and their Preferences.

4.1 Input Trials

The Input Trials gave participants experience with all three input modalities.

4.1.1 Total Capture Time

To understand how the Ink, Voice and Keyboard input methods compared to each other, we first considered the total time it took participants to capture micronotes using each input modality. Fig. 2 shows the average total capture time for the three input modalities and two transcription conditions broken down by input and edit time.

We first compared capture times within transcription conditions by conducting a 3 (InputMode: Ink, Voice, Keyboard) x 3 (Length) RM-ANOVA for each transcription condition. However, other than the expected main effects of Length on capture time for both groups, no significant effects of InputMode were present for either NearPerfect or CurrentDay participants. So overall, capture times for participants in each condition were not significantly different across the three input modalities, which we found somewhat surprising.

Next, we compared how transcription quality affected total capture times, by comparing capture times between participants in the two transcription conditions for each input method. Looking at the right hand side of Fig. 2, we can see that participants performed comparably in the Keyboard condition across the two transcription conditions. Given that the Keyboard condition lacked a transcription/edit phase, this performance similarity is what we would have expected and gives us confidence that our random assignment of participants into the two transcription conditions avoided any unintentional speed bias.

For Ink and Voice, where we would have expected transcription quality to make a difference, we compared the capture times between the two transcription conditions. We conducted a one-way RM-ANOVA with a within-subjects factor of Length and a between-subjects factor of transcription quality (TxQuality) on the mean total micronote capture time for Ink and Voice. Both tests yielded the expected significant main effects for Length, but only Voice tasks showed significant effects of TxQuality ($F(1,16)=8.6$, $p=.01$), with NP participants completing Voice tasks faster than CD participants (29.9s. v. 43.2s).

4.1.2 Input and Edit Times

We also independently analyzed the time spent to input and edit the micronote. To explore input time, a 3 (InputMode) x 3 (Length) RM-ANOVA with a between-subjects factor of TxQuality was performed on mean input time. Significant main effects were found for InputMode ($F(2,15)=53.89$, $p<.001$) and again for Length ($F(2,15)=62.71$, $p<.001$). However, as per our design, TxQuality did not affect input times, which again gives us confidence that participants' input speeds were balanced across the transcription conditions. Post hoc tests on InputMode using Bonferroni correction revealed significant differences between input speeds of all modalities; Voice supported faster micronote input than Ink (11.15s v. 26.69s, $p<.001$), which in turn was faster than Keyboard (34.70s, $p=.002$).

Given that input times varied significantly by InputMode, but total capture times did not, we conclude that edit times were inversely proportional to input times for these three modalities. This interpretation is also supported by Fig. 2 which shows Voice edits generally took longer than Ink edits, which took longer than Keyboard edits. This is perhaps not surprising given our design incorporated more transcription errors in Voice tasks than in Ink tasks, as well as more errors in CurrentDay tasks than in NearPerfect tasks. However, given that we asked users to correct tasks to something they were "comfortable" with, we worried that participants might have chosen not to edit the transcriptions. To check this, we ran a 2 (InputMode: Ink, Voice) x 3 (Length) RM-ANOVA with a between-subject factor TxQuality on mean number of corrections. Keyboard was excluded from the analysis because it did not have an edit phase. Significant main effects for InputMode ($F(1,16)=140.8$, $p<.001$), TxQuality ($F(1,16)=85.2$, $p<.001$) and Length ($F(2,15)=91.3$, $p<.001$) were present.

Post hoc analyses found that users made significantly fewer corrections during Ink tasks than Voice tasks (3.6 v. 11.0, $p<.001$) and corrected fewer errors in the NearPerfect condition than the CurrentDay condition (3.5 v 11.1). Thus, our analysis shows that users indeed varied the amount of correction applied to tasks according to the number of errors present in the transcription, preserving the relative edit effort we had intentionally designed into the study. Equivalent analyses on edit times (vs. number of corrections) yielded the same findings.

4.1.3 Survey Responses

We surveyed participants after they had used each input modality using a NASA TLX-based survey. Participants answered questions on a 7-point Likert scale, with 1='Very Low' and 7='Very High' about mental demand (Voice: med.=2 Ink: med.=2, Keyboard: med.=3), physical demand (V:2, I:3, K:2.5), whether they became discouraged during the tasks (V:2, I:2, K:2), or had to work hard (V:3, I:2.5, K:2). We also asked how easy the input was to use¹ (V:2, I:3, K:3.5), to learn (V:2, I:2, K:1), and lastly how quickly they felt they could do the tasks (V:3, I:3, K:3). We conducted Friedman tests to compare the responses across the three input modalities and saw a statistically significant difference only for physical demand ($\chi^2(2, N=18) = 11.36$,

¹ Questions about ease of use and learning used the scale of 1='Not Easy' to 7='Very Easy' while the 'quickly' question had the scale 1='Very slowly' to 7='Very Quickly.' All three questions have been reverse coded for analysis to be consistent with other questions where lower scores indicate less effort and higher scores indicate more effort.

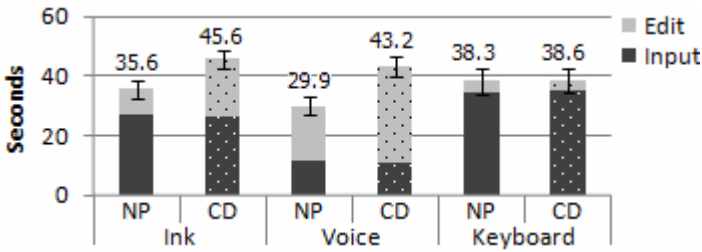


Fig. 2. Total capture time (input+edit) averaged across micronote lengths by input mode and transcription condition (NP=NearPerfect, CD=CurrentDay)

$p=.003$). Follow-up pairwise comparisons using a Wilcoxon test showed participants felt Voice required significantly less physical exertion than Ink ($z=-2.97$, $p=0.003$).

To compare responses of participants in the CurrentDay (CD) condition to those in the NearPerfect (NP) condition we conducted Mann-Whitney U tests for each of the survey questions. We saw statistically significant differences only for the Voice input modality, specifically for the questions on mental demand (NP:1, CD:3, $z=-2.72$, $p=.007$), physical demand (NP:1, CD:3, $z=-3.00$, $p=.003$), how discouraged participants were (NP:1, CD:4, $z=-2.70$, $p=.007$) and how hard they felt they had to work (NP:2, CD:5, $z=-2.89$, $p=.004$). Taken together these ratings suggest that differences in transcription quality for the Voice input did make a noticeable difference in the perceptual impact on the participants.

4.2 Timed Competition

The Timed Competition, which participants completed right after the Input Trials, was designed to help us answer our first research question and determine whether participants could build a correct mental model of how long it took them to capture a micronote. For each of the 108 tasks in the Timed Competition (6 tasks for 18 participants), we calculated the input modality choice that would have been optimal for the participant to use based on the input modality that was fastest for them for tasks of the same length in the Input Trials.

Based on these calculations, 8 of the 18 participants, all from the NearPerfect condition, were expected to use Voice for all task lengths. Seven of the participants should have switched between 2 different modalities, and 3 were expected to switch between all 3 modalities. While 3 participants (NP:3, CD:0) used the optimal input modality for all tasks, our data suggests that most participants did not necessarily have a clear understanding of what input modality would be fastest for them for different task lengths. The remaining 15 participants (NP:6, CD:9) chose the non-optimal input modality in 51 tasks (47%). We were somewhat surprised that the incorrect choices were split relatively evenly between task lengths (Short:25%, Medium:37%, Long:37%) as we expected participants might have more trouble determining the optimal input mode for shorter tasks where capture times using different modalities might be more similar. However, we did find that participants in the CurrentDay condition appeared to have more trouble. Of the 51 incorrect choices, 63% of them were made by participants in the CurrentDay condition while only 37% were

made by participants in the NearPerfect condition. Finally, when participants made a non-optimal choice they did not appear to favor any one particular input modality, using Voice 37%, Ink 37% and Keyboard 26% of the time. People in the CurrentDay condition who chose non-optimally chose Voice 41% (13), Ink 38% (12) and Keyboard 22% (7) of the time. NearPerfect participants used Voice 32% (6), Ink 37% (7) and Keyboard 32% (6) of the time.

Survey responses also support the notion that participants had trouble determining what method was fastest for them. On the final survey we asked participants to rank input modalities from fastest to slowest for capturing short lists and long lists. We compared participants' reported fastest methods to their actual performance in the timed trials and again found they correctly identified their fastest method correctly only 53% of the time.

4.3 Different Environments

Our second research question asked about the affect of environment on the participants' choice of input modality. While understanding realistic usage across different environments requires a field study, we felt it was valuable to take our lab study participants into two additional environments (café and walking) to explore in a structured way the impact of environment on their choice of input modality.

Of the 108 micronotes captured in the café (in both conditions), participants used Voice for 38% of the micronotes, Ink for 39%, and Keyboard for 23%. When walking, participants used Voice for 72%, Ink for 17% and Keyboard for 11%. Fig. 3 shows participants' input choices by transcription condition. Participants in the NearPerfect condition had a mostly even split between Ink and Voice inputs in the café, while CurrentDay participants chose more evenly among the three inputs. A few participants expressed concerns about Voice in public venues. Comments included "It would seem weird to command your phone while you are in a café or public place" (P9), and "you don't always want everybody to hear what you are saying" (P7).

The worse transcriptions that the CD participants received may have negatively affected the experience with Voice and Ink enough to encourage the use of Keyboard input. However, while walking, the majority of participants in both transcription conditions chose Voice input, so some users in both conditions switched from another input to Voice, presumably prioritizing Voice's low visual and physical demands over

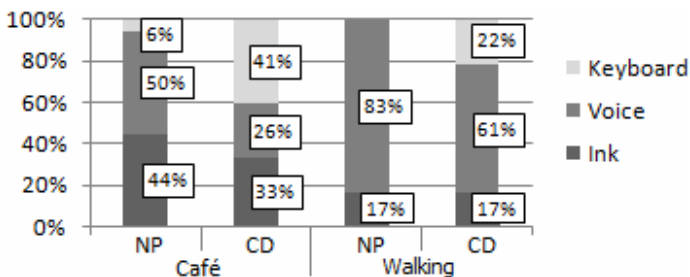


Fig. 3. Percentage of environment tasks captured with the three input modalities, by environment (Café v. Walking) and transcription quality group

other considerations while mobile. Despite the high visual and physical attention required for Keyboard input, Keyboard was not entirely abandoned by the CD participants during walking as it was by the NP participants. This again demonstrates the seemingly negative influence that the worse transcription quality had on CD participants' likelihood of choosing one of the more "natural" input methods.

We were also interested in how consistently participants used the same input modality in each environment. In this analysis we considered users to have a consistent input for an environment if they used a particular input method for at least 4 of the 6 tasks performed. Nine participants (NP:6, CD:3) did not change their preferred input method based on the environment; in 7 of these cases, the method chosen was the one they reported as their overall preferred capture method among the three studied, and in two cases the method chosen was the one they considered fastest, based on their Timed Competition data. Of the other 9, 4 (NP:1, CD:3) did not exhibit a consistent input method in one or both environments. The 5 remaining participants switched from another input method to Voice for walking.

To summarize, participants in the NP condition were more likely to choose Voice or Ink over Keyboard for micronote capture and tended to be more stable in their choices. In contrast, CD participants were more divided among three input methods in the café, had more participants who made inconsistent task-to-task choices, and had more participants switch input choice between environments. These fluctuations observed in the CD participants suggest that the worse transcription made it more difficult for them to determine or decide on a preferred input method. For both conditions, however, we saw that environment could influence choice, with users favoring the lower demands of Voice for capturing micronotes while walking.

4.4 Participant Preference

Our final research questions asked what input modalities participants prefer and whether it appeared that either transcription quality or speed of capture had an effect on participants' preferences. On the final survey we asked participants to rank the input modalities from most to least favorite. Responses strongly demonstrate participants' preference for Ink and Voice input over Keyboard. As Fig. 4a shows, 8 participants preferred Voice, 8 participants preferred Ink and 2 preferred Keyboard for micronotes when paper was not option. When asked what they liked about their favorite method, common responses for both Ink and Voice included ease of use (Voice:6 participants, Ink:2), accuracy of transcription (Voice:5, Ink:4), and speed (Voice:3, Ink: 2). Very surprisingly, the number of participants who preferred Voice, Ink and Keyboard between the NearPerfect and CurrentDay conditions was exactly the same when paper was not an option.

However, Fig. 4b highlights that Paper was still a very popular choice for micronotes. Given the option, 8 participants (4 from each condition), ranked Paper as their most preferred input method. Participants who originally favored Ink seemed most likely to prefer Paper (5 moved from Ink to Paper). Again transcription quality did not seem to greatly affect preferences as the same number of participants from both transcription conditions preferred paper.

We also examined whether participants' preferred input modality might correlate with the modality they felt they were fastest using. Given that half (8) the participants

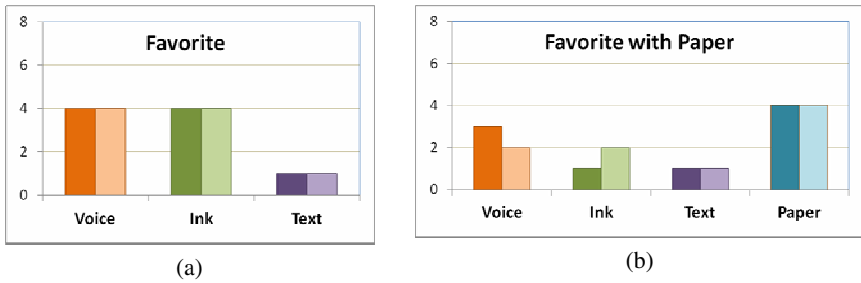


Fig. 4. Participants' favorite input modality (a) without and (b) with paper. Darker bars are users in the NearPerfect condition, lighter bars are CurrentDay.

felt that two different input methods were fastest for short vs. long tasks, it is impossible to draw conclusions about whether those participants' single preferred capture method was related to their perception of the fastest input method. But out of the 8 participants who felt that a single capture method was fastest for both long and short tasks, 7 chose that same method as their preferred capture method overall, providing some evidence that for those individuals, perceived capture speed influenced their stated modality preference.

We found it interesting that Ink was preferred as often as it was (8 participants) considering it was fastest on average for only one participant. Some of this discrepancy can be explained by the fact that 6 of those participants (incorrectly) *thought* Ink was fastest for at least some tasks, while 2 participants knew Ink was slower than others, but still preferred it overall. Even so, the relative popularity of Ink in the face of modest performance suggests that users appreciate a broader range of capture qualities beyond speed, such as similarity to paper-based note taking, and discreet capture. Comments on the final survey about Ink included “I like writing in my own handwriting” (P11), “seeing written words helps jog ideas and gives time to think” (P13), and “I always write/draw, so it's very familiar and amusing” (P16).

Finally, while not the focus of our study, participants provided several important pieces of feedback about the usability of the VINO prototype. First, while VINO required participants to input and then edit, some participants wanted to have the ink or voice transcription available immediately so they could see and correct any problems as they were inputting the micronote. Additionally, a few participants were very adamant about wanting to take notes in multiple different colors and easily changing colors mid-note. Lastly, a few participants felt the custom-keyboard was too small for older eyes and would have liked the text to be larger.

5 Discussion

We now discuss more broadly what our findings suggest for future mobile micronote capture technologies. First, participants' preferences, and their use of Ink and Voice input when given the option, strongly support the value of providing users with these “natural” input modalities for digitally capturing micronotes. More specifically, given that half our participants switched which input modality they used in different

environments, as well as the almost equal split in preferences for Ink or Voice, we believe that our study highlights the importance of multi-modal micronote capture applications that allow users to select whatever capture modality is appropriate for their current context and needs. This is in contrast to separate end-to-end applications for different input modalities that seem to be the focus of current development (e.g., Jott supports only Voice input). Supporting Ink, Voice and Keyboard input equally, and allowing participants to switch between them as we prototyped with VINO, would better support the capture behaviors we observed where some participants switched depending on the environment. Even as predictive text entry improves and becomes widely available, we anticipate different input modalities will remain valuable due to the need for low attention interfaces in certain environments and participants' varied preferences.

While clearly some research on improving transcription quality would be beneficial, having micronotes with near perfect transcription did not appear to make participants in the NearPerfect condition less resistant to the allure of paper (4 of 9 preferred paper when given the option). One consideration is that in our study participants were forced to correct the transcription after entering the complete note. Given that some participants wanted to have the Voice and Ink transcription available immediately so they could see and correct problems, we believe two approaches to transcription would be interesting for further study. First, providing immediate transcription even with a likely reduction in accuracy, and second emphasizing 'just-in-time' transcription where users would only edit or view transcription if needed (e.g., often the unrecognized Ink note might be fine) and ideally at a place where it might be easier to correct such as on a desktop or laptop computer.

6 Concluding Remarks

Our study participants preferred natural input modalities of Voice and Ink to a virtual Keyboard for making micronotes on a mobile device; however paper remains appealing to many. While participants in our CurrentDay transcription condition did find the tasks more challenging in some respects, particularly for Voice input, having worse transcription did not change the distribution of participants' preferences across the input modalities. Nor did more participants in the NearPerfect condition favor digital micronote capture over paper compared to those in the CurrentDay condition, suggesting dramatically better transcription quality on its own will not cause people to adopt digital micronote capture technologies.

However, given that our participants are already carrying mobile phones that will only grow more powerful in the years to come, we do believe there is an opportunity for developing mobile micronote capture technology that meets people's needs. We feel strongly these applications must support multiple input methods including Voice and Ink so that users can switch between different modalities as they desire. While we focused on input methods for capture in our study, it is critical to remember the entire lifecycle that Lin et al. identified, as emphasizing the additional benefits that digital captured notes might have (e.g., easy to share, search) will no doubt be important in developing a digital micronote capture system that people find useful.

Going forward we believe there are two valuable research directions that would build on our findings. First, conducting a similar study with users of smartphones to better understand their past experience with existing technology and compare their responses to VINO with the results from this study. Second, building a more robust version of VINO, perhaps with offline transcription, that could be deployed for short field studies in order to further explore the appeal of a multi-modal capture application in the wild.

References

1. Lin, M., Lutters, W.G., Kim, T.S.: Understanding the micronote lifecycle: improving mobile support for informal note taking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 687–694. ACM, New York (2004)
2. Bernstein, M., Van Kleek, M., Karger, D.R., Schraefel, M.C.: Information Scraps: How and Why Information Eludes our Personal Information Management Tools. *ACM Transactions on Information Systems* 26(4), 1–46 (2008)
3. Bernstein, M.S., Van Kleek, M., Schraefel, M.C., Karger, D.R.: Management of Personal Information Scraps. In: CHI 2007 Extended Abstracts on Human Factors in Computing Systems, pp. 2285–2290. ACM, New York (2007)
4. Dai, L., Lutters, W.G., Bower, C.: Why use memo for all?: restructuring mobile applications to support informal note taking. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1320–1323. ACM, New York (2005)
5. Campbell, C., Maglio, P.: Supporting notable information in office work. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems, pp. 902–903. ACM, New York (2003)
6. Hayes, G.R., Pierce, J.S., Abowd, G.D.: Practices for capturing short important thoughts. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems, pp. 904–905. ACM, New York (2003)
7. Zhou, L.: Natural Language Interface for Information Management on Mobile Device. *Behavior & Information Technology* 26(3), 197–207 (2007)
8. Falke, E.: The associative pda 2.0. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, pp. 3807–3812. ACM, New York (2008)
9. Jott, <http://www.jott.com>
10. Evernote, <http://www.evernote.com>
11. Microsoft Ink API, <http://msdn.microsoft.com/en-us/library/microsoft.ink.aspx>
12. Microsoft Speech API, <http://msdn.microsoft.com/en-us/library/system.speech.recognition.aspx>
13. Soukoreff, R.W., MacKenzie, I.S.: Measuring errors in text entry tasks: an application of the Levenshtein string distance statistic. In: CHI 2001 Extended Abstracts on Human Factors in Computing Systems, pp. 319–320. ACM, New York (2001)
14. Ludford, P.J., Frankowski, D., Reily, K., Wilms, K., Terveen, L.: Because I carry my cell phone anyway: functional location-based reminder applications. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 889–898. ACM, New York (2006)