

Evaluation of Five Interruption Methods for Speech Interaction in Wearable Computing Dual-Task Environments

Hendrik Witt
TZI, Wearable Computing Lab.
University of Bremen
hwitt@tzi.de

Abstract

Because users of wearable computers are often involved in dual-task situations with real-world tasks of a critical nature, interruption handling is a crucial design issue of wearable user interfaces. We present a study evaluating the impact of five interruption methods to determine how interruptions should be handled when speech interaction is available to control an application. The study uses the HotWire apparatus to simulate a primary manual task in a laboratory experiment. Results indicate that speech interaction can compensate for many typical challenges users have to deal with in a dual-task situation and can reduce the impact of an interrupting wearable computer.

1. Introduction

Unlike stationary computers where attention is focused on a single application, wearable computers expect users to accomplish two different tasks. A primary task involves real world physical actions while the wearable computer and the interaction with it are only secondary. Thus, wearable computers call for different user interfaces that regard their unique characteristic. To minimize cognitive load and to allow users to divide their attention between both tasks as easily and efficient as possible, the handling of interruptions is a crucial design issue of wearable user interfaces.

McFarlane [3] presented the first empirical study to coordinate interruptions in human-computer interaction (HCI) with multiple tasks. His method was based on a simple computer game that required constant user attention while users were randomly interrupted by a matching task. As a continuation of his work for wearable computing, in [1] a head-mounted display (HMD) was used to display matching tasks. It was found that a scheduled approach gave best performance, while notifications came second. User preferences on interruption were studied in [4]. Audio notifications appeared to give slightly better performance although

users considered them more stressful compared to more distracting visual signals. Drugge et al. [2] replaced the game task by a more realistic physical task represented by the HotWire [7] apparatus to study interruptions. They confirmed and complemented previous results for handling interruptions with gesture interaction.

Unlike novel gesture interaction, speech is probably the most natural way to interact with a computer. Nowadays, available speech recognition software has overcome many technical problems and can be used for simpler interaction in applications. A major advantage of speech interaction is its “hands-free” nature, i.e. users do not need their hands during interaction. Therefore it is a promising interaction technique for wearable computers [6]. However, little work on wearable audio interfaces and interruptions has been carried out so far (see e.g. [5]).

This paper is a continuation of a series of research [1, 2, 4] we have conducted studying interruptions for wearable computing. We will focus on examining different interruption methods while users perform a physical primary task and handle interruptions using speech input.

2. Experiment

The experiment addresses how different methods of interrupting the user of a wearable computer affect that person’s performance. The scenario involves the user performing a primary manual task in the real world while interruptions originate from the wearable computer and are handled with speaker-independent speech interaction.

2.1. Primary Task

In order to more easily relate the results of this study to previous work, we decided to rebuild the HotWire primary task setup successfully used by Drugge et al. [2].

Our apparatus is shown in Figure 1. It consists of a metallic wire that was bent in the same shape and mounted in the same way to a base plate as done in [2]. In contrast to the original ring tool of our previous work which had a too small ring diameter, our tool has a slightly bigger diameter

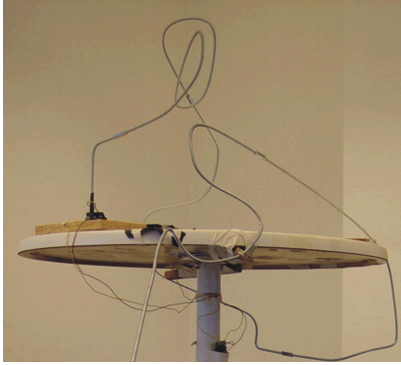


Figure 1. The HotWire apparatus used.

of 2.6 cm (an increase of 0.4 cm). Track length and mounting height are identical with [2].

2.2. Secondary Task

The secondary task consists of different matching tasks presented in the user’s HMD. Two examples of these are shown in Figure 2. Figure 2(a) depicts the kind of matching tasks already used by Drugge et al. [2]. To increase cognitive workload on more levels than just shape and color matching, a second kind of matching task was added in form of an arithmetic exercise (see Figure 2(b)). The arithmetic task presents a mathematical expression of type

$$X < operator > Y, < operator > := + | - | * | /$$

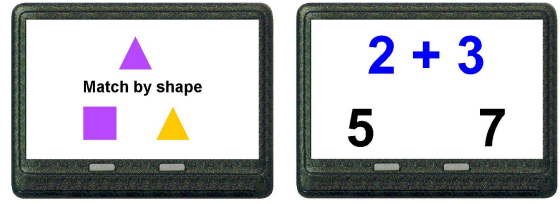
Expressions and answers are limited to integers. X and Y range from 1 to 9 for the sake of simplicity while still requiring enough mental effort. Matching tasks are created at random. If a visible matching task cannot be handled within 5 sec., it disappears and is counted as a wrong answer as suggested in [2]. If tasks are very frequently created and cannot be answered in time they are queued up.

2.3. Methods for Handling Interruptions

The methods used for managing the interruptions in our experiment are basically the same as those used in [2]:

- **Immediate:** Matching tasks are created at random and presented for the user in the instant they are created.
- **Negotiated:** On randomly created tasks, the user is notified by either a visual or audible signal and can then decide when to handle them.
- **Scheduled:** Matching tasks are created at random but presented only at specific time intervals of 25 seconds.
- **Mediated:** Presentation of matching tasks is withheld during difficult sections of the HotWire. The algorithm is very simple; it counts contacts occurred in a time window of 3 sec. as a heuristic for task difficulty.

Additionally, there are also the two base cases (HotWire only and Match only) included that serve as baselines.



(a) Figure matching

(b) Arithmetic matching

Figure 2. Different types of matching tasks.

3. User Study

A total of 21 subjects were selected for participation. These were taken from students and staff at the local university — 13 males and 8 females aged between 21–55 years (mean 29.05). All subjects were screened for color blindness. The study uses a within subject design with the interruption method as the single independent variable. To avoid bias and learning effects, subjects were divided into counterbalanced groups. A single test session consisted of one practice round to practice both tasks followed by one experimental round.

The technical setup consisted of the HotWire, a Microoptical SV-6 HMD, and an audio headset. HMD and headset were connected to a stationary computer running the experiment. For a realistic situation users had to wear a special textile vest designed to unobtrusively carry a wearable computer and all cables. Furthermore, an OQO computer in the vest simulated the weight of a wearable computer.

The headset served as the interface to control matching tasks through speech commands. All tasks could be answered in the same way. By saying “left” or “right”, the left or right answer of a matching task was selected. To provide feedback, an audio signal, deemed not to interfere with the negotiated audio method, was used. For negotiated methods, the third command needed was “show me”.

4. Results

For analysis the following metrics were considered

- **Time:** The time required for the subject to complete the HotWire track from start to end.
- **Contacts:** The total number of contacts the subject made between the ring and the wire.
- **Error rate:** The percentage of matching tasks the subject answered wrong or that timed out.
- **Average delay:** The average time between a matching task creation and the subject giving an answer.

The four graphs in figure 3 depict the overall user performance for each method by showing the achieved average of the metrics together with one standard error. A repeated measure ANOVA was used to see whether there exist any significant differences between the methods tested. Table 1

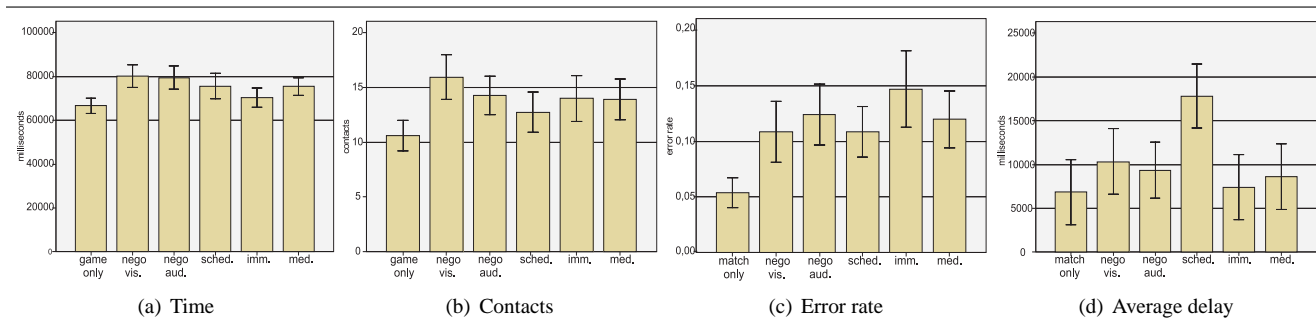


Figure 3. Averages of user performance.

shows the results. For all metrics except error rate, significance was found; average delay showed even strong significance ($p < .001$), indicating that differences do exist.

Metric	df	F	P-value*
Time	125	5.840	0.001
Contacts	125	4.330	0.003
Error rate	125	2.171	0.083
Average delay	125	147.827	<0.001

* with Greenhouse-Geisser df adjustment applied.

Table 1. Repeated measures ANOVA.

To explore differences paired samples t-tests ($\alpha = .05$) were performed to compare the two base cases with each of the interruption methods. To accommodate for multiple comparisons a Bonferroni corrected α -value was used. Table 2 shows the significant results found.

Although we intuitively assumed that task complexity increases once being involved in a dual task situation, which likely causes errors and completion time to increase, our data does not generally support this. Neither the *scheduled* nor *immediate* treatment showed significant differences in completion time; all others did as expected though. This indicates that either our data could not uncover an existence or that some methods are more appropriate than others when using speech interaction.

Metric	Nego-Vis.	Nego-Aud.	Sch.	Imm.	Med.
Time	0.025	0.011	0.539	1.000	0.004
Contacts	0.023	0.835	0.472	0.997	1.000
Average delay	<0.001	0.002	<0.001	<0.001	0.001

Table 2. Base case comparison t-tests.

Regarding contacts, surprisingly we only found a significant difference for *negotiated visual*. One reason could be that the small increase in the ring diameter already resulted in a significant decrease of the HotWire’s difficulty. The hands free nature of speech interaction, which allowed subjects to hold the ring tool more accurately, is another possible explanation. In fact, this was what our post-hoc video analysis of user behaviors uncovered. Subjects used their free hand to either stabilize their bodies or the ring tool dependent on physical or visual attention demands of the primary task.

Although error rate results are similar to [2] (all being non significant), we expected differences for this study. Since there was a time out in matching tasks, this “forced” subjects to make more mistakes due to time pressure. This was at least what pilot studies indicated. In line with our expectations, there were the strongly significant ($p < .001$) differences in average delay. Since the user is involved in the HotWire task when matching tasks appear, and both the *scheduled* and *mediated* methods will by definition cause matching tasks to queue up, this causes age to increase.

In summary, adding interruptions to our dual task scenario with a manual primary task will make it more difficult to successfully carry out the tasks even with speech interaction. Some interruption methods do better than others though. Next, we examine the five interruption methods in more detail using Bonferroni corrected paired samples t-tests.

4.1. Time

Only *negotiated audio* vs. the *immediate* method exhibited a significant difference ($p = .002$). However, as *negotiated visual* was on average even 1 sec. slower than *negotiated audio* but with a smaller standard error, the data is trending to indicate that both negotiated methods differ from the immediate method. In general, the reason for the higher completion time of the negotiated methods is because of the extra effort required to present matching tasks with a second command. This result was therefore expected. Importantly, overhead (6.6 seconds higher, an increase of 6%) was much lower than expected. Although the recognizer needed 700ms to recognize a command, the overhead is still clearly smaller than found for gesture interaction [2]. With gesture there was an overhead of 24.8 sec. increased completion time about 26% for a similar experiment. This data indicates that with speech input the extra effort in the negotiated audio methods is almost negligible.

With respect to our previous work [2, 1], speech interaction is likely preferred in wearable computing where short completion time is important and where users are involved in complex primary tasks with substantial visual attention demands. In our dual-task scenarios, users performed faster

with speech than with gestures.

4.2. Contacts and Error Rate

Unlike time, the number of contacts subjects made on the HotWire did not allow to derive a clear grouping of methods. Collected data could not rival significant differences between methods. *Negotiated visual* gave on average the highest number of contacts (14.95) while *scheduled* gave the lowest number (12.76). The result of 74% less contacts on average compared to [2] were unexpected.

Similar to our results in [2], the error rate exhibited no significant differences between methods. This was unexpected because our current changes in the matching task (adding arithmetic tasks and time limits) were deemed to make matching more challenging, i.e. causing more incorrect answers.

All in all, a clear ranking of methods based on our data on contact or error rate was not possible. Our observations uncovered an interesting aspect though. Almost all of the 8 female subjects very frequently confused left and right selections, i.e. females often stated the opposite command of what they apparently intended to state. As a consequence females often restated a second (now corrected) command which sometimes increased matching errors. Only one male subject showed similar problems. An in-depth analysis of this observation will be subject of future work.

4.3. Average Delay

Naturally, the average delay is expected to be the highest for the *scheduled* method, since the matching tasks are by definition queued and presented only in fixed time intervals. A difference was found with strong significance ($p < .001$) for all methods. The results also showed significant differences between the methods at the same level except between the two negotiated methods and *mediated*.

The *immediate* method provided on average quickest response (7.44 sec.), followed by the *mediated* (8.62 sec.), *negotiated audio* (9.36 sec.), and *negotiated visual* method (10.35 sec.). With an average delay of 17.8 seconds the *scheduled* method exhibited the longest average delay, but with an expected age of 12.5 seconds on average due to the queuing, this means the user only spends on average 5.2 seconds to respond to queued matching tasks, which is then the fastest response of all methods. We think this is because the need to mentally switch between primary and secondary task is reduced because of the clustering.

In line with our study in [1], the current data confirms the advantage of audio notifications for *negotiated methods* compared to visual ones. This indicates that with speech interaction users may compensate for typical challenges in dual-task environments with a manual primary task, like we tested, more easily than with gesture interaction. We think this is likely because hearing and speaking use the same stimulus.

5. Conclusion

The present study showed that the interaction device used has an impact on the selection of the best interruption method in wearable computing. Methods that were found superior for gesture interaction in dual-task situations [2] are different to those for speech interaction. Speech interaction is preferred for our complex primary tasks tested, particularly when task performance is concerned. Maintaining visual focus on a primary task is easy with speech interaction because it does not interfere much with manual activities.

Although each method has strengths and weaknesses, overall the *negotiated audio* method was most positively affected by speech. The extra effort needed to bring up matching tasks was very small in both negotiated methods and increased completion time only about 6% compared to the HotWire-only task. The number of contacts made, compared to HotWire only, did not increase significantly for any method except *negotiated visual*. *Negotiated audio* exhibited the lowest impact on the primary task likely because hearing and speaking both use the same stimulus. Subjects very easily kept focused with *negotiated audio*. With the *immediate* method subjects were fastest. Completion time did not increase significantly compared to HotWire only. *Scheduled* was found to give the lowest response time. If many tasks queue up *scheduled* allows fastest response to each task even with slow speech recognizers.

Acknowledgment This work has been funded by the European Commission through IST Project wearIT@work (No. IP 004216-2004).

References

- [1] M. Drugge, M. Nilsson, U. Liljedahl, K. Synnes, and P. Parnes. Methods for Interrupting a Wearable Computer User. In *ISWC'04*, November 2004.
- [2] M. Drugge, H. Witt, P. Parnes, and K. Synnes. Using the HotWire to study interruptions in wearable computing primary tasks. In *IEEE ISWC'06*, Montreux, Switzerland, October 11–14 2006.
- [3] D. C. McFarlane. Coordinating the interruption of people in human-computer interaction. In *Human-Computer Interaction - INTERACT'99*, pages 295–303. IOS Press, Inc., 1999.
- [4] M. Nilsson, M. Drugge, U. Liljedahl, K. Synnes, and P. Parnes. A Study on Users' Preference on Interruption When Using Wearable Computers and Head Mounted Displays. In *Communications IEEE PerCom'05*, March 2005.
- [5] N. N. Sawhney and C. Schmandt. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput.-Hum. Interact.*, 7(3):353–383, 2000.
- [6] T. E. Starner. The role of speech input in wearable computing. *IEEE Pervasive Computing*, 1(3):89–93, 2002.
- [7] H. Witt and M. Drugge. HotWire: An apparatus for simulating primary tasks in wearable computing. In *ACM CHI '06: Extended Abstracts*.