Modality Preferences in an Instrumented Environment

Rainer Wasinger  
DFKI GmbH  
Intelligent User Interfaces Department  
66123 Saarbrücken, Germany  
+49 (681) 302-3393  
rainer.wasinger@dfki.de

Antonio Krüger  
University of Münster  
Institute for Geoinformatics  
48149 Münster, Germany  
+49 (251) 83-33073  
antonio.krueger@uni-muenster.de

ABSTRACT
In this paper, we describe the results of a usability study on user preferences for multimodal interaction in an instrumented environment. The study was conducted in a public setting, and provides insight into modality preferences among users, and specific to men and women. The returned results are also contrasted to the results of a former study based on the same evaluation procedures but conducted under a laboratory setting.

Categories and Subject Descriptors
H.5.2 [User Interfaces]: Input devices and strategies, interaction styles, user-centered design. I.2.1 [Applications and Expert Systems]: Natural language interfaces.

General Terms
Design, Measurement, Human Factors.

Keywords
Multimodal Interaction, Mobile, Ubiquitous Computing.

1. INTRODUCTION
Work on multimodal interfaces is slowly moving from resource-rich desktop computers to that of mobile devices in the form of client-server configurations (e.g. Smartkom [4], QuickSet [1]), and standalone configurations (e.g. the Mobile ShopAssist [5]). Mobile, ubiquitous, and pervasive computing environments stand to benefit significantly from multimodal interfaces that cover a large proportion of a user's day-to-day interaction. These environments do however differ significantly from the traditional desktop computing scenarios. Users in these environments are commonly standing or moving (but not sitting at a desk), and have to deal with constantly changing environment characteristics and differing types of background noise. Devices for these environments also differ to their desktop counterparts, and are generally characterized by their restrictive limitations such as a small display, and minimal I/O processing and storage capability.

Past studies on multimodal interaction have focused on efficiency and recognition accuracy gains [3], but are often based on stationary settings and desktop computers, or Wizard of Oz mockups. One study dedicated to a mobile setting is that conducted by Kumar et al. [2], which analyzes the effects of exertion on users interacting multimodally in an outdoor setting.

Other lines of work have focused on creating guidelines for the advantages and disadvantages of modalities (e.g. Sun's Java Speech API guidelines, see http://java.sun.com/products/java-media/speech/). Although such reports outline factors that can affect modality use (e.g. background noise can have a negative effect on speech), they only target individual modalities in isolation, rather than the combination of multiple modalities.

This work focuses on user preferences for different multimodal input combinations. It is based on a shopping scenario and takes 23 different unimodal and multimodal combinations into account. These combinations include both those in which the underlying semantics of the user input may be non-overlapped or overlapped (e.g. where speech and handwriting input are used to provide duplicate information rather than complementary information). All combinations are derived from three elementary modalities: speech, handwriting, and gesture, whereby gesture may be further categorized as being either intra-gesture (e.g. pointing to a referent on the Pocket PC's display), or extra-gesture (e.g. picking an object up in the physical world). The underlying service used as the basis for this study is the Mobile ShopAssist [5], which allows a user to compare different products like digital cameras, and to query their attributes. This information is semantically interpreted and then represented as <FEATURE> and <OBJECT> elements in a modality free language. The main modality-free language template used for this study is <FEATURE><OBJECT>+, where “feature” refers to a camera’s attribute like “mega pixels” and “object” refers to an actual product like the “PowerShot S80”. As an example, the above attributes would result from the user speaking the utterance: “How many mega pixels does the PowerShot S80 have?”. The concepts derived from the given shopping scenario are also expected to apply to other day-to-day scenario contexts.

2. USABILITY STUDY
This study builds extensively upon a prior study [5] which looked at modality preferences in a “laboratory” setting. The two fundamental differences of the study described here are that it took place in a “real-world” setting, and that it encompassed twice the number of test subjects. These differences allow important conclusions to be formed on the effects that a laboratory and real-world setting have on multimodal interaction, and also permit conclusions to be drawn regarding the effects of gender on multimodal interaction in a real-world setting.

The study was conducted inside a retail shop called “Conrad Electronic”, which sells more than 50,000 technology and electronic products. Our sample of test persons consisted of 28

Copyright is held by the author/owner(s).

IUI’06, January 29–February 1, 2006, Sydney, Australia.
ACM 1-59593-287-9/06/0001.

1 Conrad Electronic, Saarbrücken, http://www.conrad.de
people (16 female and 12 male) ranging in age from 19 to 55 years (mean: 28.3 years). 17 subjects were classified as experienced computer users, but only 5 subjects had had previous experience with a Personal Digital Assistant (PDA). All subjects bar two were unfamiliar with the system. The testing was conducted over a two week period, during which time a total of 1489 interactions were recorded.

Our experiment setup consisted of a single instrumented shelf placed in a central part of the store as shown in Figure 1. In comparison to the prior study, our subjects were subjected to a real-world environment, in which on average 13.8 people could be seen from the shelf’s location during any given session. Typical real-world disturbances included the shop’s loudspeaker system, goods trolleys, sales assistants, and customers.

Sessions took between 45 and 60 minutes to complete. At the beginning of each session, the subjects were informed about the system and the type of interactions that were possible. Notably, this consisted of FEATURE-OBJECT interactions based on the modalities speech, handwriting, and gesture. A total of 12 non-overlapped modality combinations and 11 overlapped modality combinations were tested (see Figure 2). The data set contained 10 different camera objects and 13 different attribute features.

At the start of a session, test subjects were given a Pocket PC and a headset through which they could speak and listen to the system. Subjects were then asked to commence interacting with the system. After each interaction, the subjects were asked to rate the modality combinations by answering the question “Would you use this modality combination?”. The rating scale used for this purpose (and iterated throughout the results section of this study) was: “0=prefer not, 1=maybe not, 2=maybe yes, and 3=prefer yes”. Upon completing the study, subjects were further asked to comment on the intuitiveness of each modality combination (yes intuitive, or no not intuitive), and their comfortability in using the base modalities (embarrassed, hesitant, or comfortable).

### 2.1 Results

This section summarizes the significant differences between modality combinations within this real-world study, between the real-world and the laboratory study conducted in [5], and between men and women. For convenience, the individual modality combinations referred to in the results are abbreviated as follows: speech (S), handwriting (H), intra-gesture (GI), and extra-gesture (GE). As an example, the interaction: `<FEATURE modality=speech><OBJECT modality=speech>` is analogous to the modality combination SS.

![Figure 1. Instrumented shelf, located at Conrad Electronic.](image)

![Figure 2. Preference comparisons between laboratory and real-world modality combinations (non-overlapped on the left, overlapped on the right).](image)

**2.1.1 Preferred Modality Combinations**

Results show that subjects in a real-world setting preferred non-overlapped combinations ($A_V=1.87$) over overlapped modality combinations ($A_V=0.43$), and a Mann-Whitney U test showed this to be significant in 23 out of 26 subjects: $U(12,11)<35, p<0.05$. The non-overlapped combinations were further grouped by the modality used for providing the feature attribute - speech, handwriting, and intra-gesture. In contrast to the laboratory study, in which speech was the preferred modality group ($S_AV=2.09$, $GI_AV=1.39$, $H_AV=1.25$), intra-gesture was the preferred modality group in this real-world study ($GI_AV=1.98$, $S_AV=1.83$, $H_AV=1.80$).

Ranked by preference, GIGI (2.65) was the most preferred out of all the 23 modality combinations in this study, followed by HGI (2.19) and SGI (2.15). The modality combinations SGE (2.00), SS (1.96), and SGI (2.15), which ranked best in the laboratory setting decreased in ranking in the real-world study to 5th, 6th, and 3rd respectively. This shows the preference that subjects had for the non-obtrusive modalities such as handwriting and intra-gesture, over speech and extra gesture. The benefit of allowing a subject to provide deictic input was also seen in that 6 out of the top 7 modality combinations use gesture to identify the object. Similar to the laboratory study, SGE (overlapped object) and SGI (overlapped object) were the most preferred overlapped modality combinations.

The main differences in modality preference between laboratory and real-world settings are shown in Table 2, where the 9 most differing modality combinations between studies are listed. The significance values (also shown in Figure 2) were calculated using the Mann-Whitney U test and are based on the combined set of 40 samples (14 from the laboratory, 26 from the real-world study).
Table 2. Modality combinations bearing significantly different preference ratings between usability studies

<table>
<thead>
<tr>
<th>Modality Combination</th>
<th>Lab Values</th>
<th>Real-world Values</th>
<th>Difference</th>
<th>Asymp. Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>2.57</td>
<td>1.96</td>
<td>-0.61</td>
<td>U(14, 26)=123.5, p=0.073</td>
</tr>
<tr>
<td>SH</td>
<td>0.64</td>
<td>1.19</td>
<td>+0.55</td>
<td>U(14, 26)=123.0, p=0.077</td>
</tr>
<tr>
<td>SGE</td>
<td>2.64</td>
<td>2.00</td>
<td>-0.64</td>
<td>U(14, 26)=101.0, p=0.013</td>
</tr>
<tr>
<td>HGI</td>
<td>1.36</td>
<td>2.19</td>
<td>+0.83</td>
<td>U(14, 26)=104.0, p=0.020</td>
</tr>
<tr>
<td>HGE</td>
<td>1.00</td>
<td>1.92</td>
<td>+0.92</td>
<td>U(14, 26)=92.0, p=0.008</td>
</tr>
<tr>
<td>GH</td>
<td>0.93</td>
<td>1.5</td>
<td>+0.57</td>
<td>U(14, 26)=121.5, p=0.068</td>
</tr>
<tr>
<td>GIGI</td>
<td>1.86</td>
<td>2.65</td>
<td>+0.79</td>
<td>U(14, 26)=99.5, p=0.006</td>
</tr>
<tr>
<td>SGI (OBJ Ov.)</td>
<td>1.43</td>
<td>0.73</td>
<td>-0.70</td>
<td>U(14, 26)=107.5, p=0.025</td>
</tr>
<tr>
<td>SGE (OBJ Ov.)</td>
<td>1.71</td>
<td>0.81</td>
<td>-0.90</td>
<td>U(14, 26)=83.5, p=0.003</td>
</tr>
</tbody>
</table>

Analysing gender differences in modality preference, the results show that women most preferred the combinations GIGI, HGE, and HGI, while men most preferred the combinations GIGI, SGI and GIGE. Preference ratings provided by women were lower for all but 3 modality combinations and provided an average difference per modality of -0.30 preference points. The largest differences between men and women were seen in the modality combinations SGE (Obj. Overlapped, -0.71), SS (-0.70), and SH (-0.61), perhaps attributable to the obtrusive nature of speech. From these three values, only SGE (Obj. Ov.) was however significant with respect to gender: U(15,11)=37.5, p=0.018.

2.1.2 Modality Intuition

The results showed that 6 out of the 12 non-overlapped modality combinations (SS, GIGI, SGI, HH, SGE, and HGI) were rated significantly differently by our subjects: \(\chi^2(1,N=25)>6.760, p<0.009\). An additional 5 combinations (HS, GIS, GH, SH, and GIGE) were rated intuitive by more than half of the subjects. In comparison, 9 from 11 overlapped modality combinations were rated significantly non-intuitive: \(\chi^2(1,N=25)>9, p<0.003\) (i.e. all except for SGI (Obj. Ov.) and SGE (Obj. Ov.)).

2.1.3 Public and Private Settings and the use of (Non) Obtrusive Modality Combinations

Chi-square tests show that our subjects felt comfortable using intra-gesture, handwriting, and extra-gesture (but not speech) within a public environment: \(\chi^2(2,N=27)>12.667, p<0.002\), and all base modalities within a private environment.

Entirely obtrusive (or observable) and entirely non-obtrusive (or non-observable) modality combinations can be taken as an index for measuring concerns that a user might have when interacting in public or private settings. An entirely obtrusive modality combination is comprised of only modalities from the set \{speech, extra-gesture\} while an entirely non-obtrusive modality combination is comprised of only modalities from the set \{handwriting, intra-gesture\}. Within a real-world setting, subjects marginally preferred the entirely non-obtrusive modalities \(\text{A}_{Ov.}=2.00\) over the entirely obtrusive ones \(\text{A}_{Ob.}=1.98\). This contrasts to the results from the laboratory settings in which a greater difference between the entirely obtrusive modalities \(\text{A}_{Obv.}=2.61\) and the entirely non-obtrusive ones \(\text{A}_{NonOb.}=1.43\) exists. Comparing the results from both studies shows that the entirely obtrusive modalities were significantly preferred in the laboratory setting: U(14,26)=96, p=0.011 (2-tailed), while the entirely non-obtrusive modalities were significantly preferred in the real-world setting: U(14,26)=90.5, p=0.009 (2-tailed). These results imply a preference shift towards non-observable modalities when in a public environment, and a preference shift towards observable modalities (which subjects stated were the simpler and more efficient modalities) when in a private environment.

While both men and women rated the non-obtrusive modalities similarly, women were less keen to use the obtrusive modalities. Within a public setting, only 33% of women stated that they would be comfortable using speech (in comparison to 67% of men), while 40% said they would feel hesitant (in comparison to 25% of men) and 27% embarrassed (in comparison to 8% of men). Similarly, only 40% of women said they would be comfortable using extra-gesture (in comparison to 83% of men), while 53% would be hesitant (in comparison to 17% of men) and 7% embarrassed (in comparison to 0% of men).

3. Conclusions

This study has highlighted several important facts about modality combination preferences that will impact the design of future mobile and multimodal interfaces. Most importantly, from the 23 tested modality combinations, intra-gesture and intra-gesture (GIGI), handwriting and intra-gesture (HGI), and speech and intra-gesture (SGI) were the most preferred combinations in a real-world or public setting, while speech and extra-gesture (SGE), speech and extra-gesture (GIGE), and speech and intra-gesture (SGI) were the most preferred combinations in a laboratory or private setting. The results on modality intuition are expected to be useful for designers of interfaces that need to be learnt quickly by users (e.g. exhibitions at a museum), and the differences between obtrusive and non-obtrusive combinations highlight that interfaces may need to cater for users differently when situated in public and private environments.

4. Acknowledgments

This work was partially funded by the German Federal Ministry for Education and Research (BMBF) under the contract no. 01 IN C02, as part of COLLADE II.

5. References


