Modality Preference – Learning from Users

Rainer Wasinger[1], Antonio Krüger[2]

[1] DFKI GmbH
66123 Saarbrücken Germany
rainer.wasinger@dfki.de
[2] University of Münster
48149 Münster, Germany
antonio.krueger@uni-muenster.de

Abstract. An important constituent for mobile and ubiquitous computing systems is the interface and the associated human-computer interactions. Mobile contexts have different design requirements to stationary desktop contexts. Although previous work has concentrated on stationary domains and unimodal systems, and more recently on multimodal systems, user evaluation for the use of different modality combinations is limited. In this paper, we outline the qualitative results from a recent usability study. These results form a general guideline in determining which base modalities and modality combinations to use when designing for mobile and ubiquitous environments.

1 Introduction

Stationary computing scenarios relating to document processing and web browsing are currently nourished by interactions based on keyboard and WIMP (Windows, Icon, Mouse and Pointer). Mobile scenarios in contrast quite often have restrictions placed on them that render such interaction as inadequate. A user may for example be standing or walking, and computing resources may be limited to the use of a mobile phone, Personal Digital Assistant (PDA), or the environment's public infrastructure.

Speech is one modality that has received a lot of attention in the recent past as processing power for mobile devices has increased. Similar to WIMP, speech input is however not always appropriate (e.g. in a noisy environment). Multimodal interaction is the current trend for modern mobile applications, and this is partly due to the flexibility and expressiveness [2] that such systems exhibit (e.g. Smartkom Mobile [3], QuickSet [1], and the Mobile ShopAssist [4]). User studies are however still limited, due to the large task of evaluating the very rich range of modality combinations that may be formed from even a limited number of base modalities such as speech (S), handwriting (H), and gesture (intra GI, and extra GE), as outlined in [4].

In this paper, we report on qualitative results that were obtained from two studies, the first conducted in a laboratory setting on 14 users, and the second conducted in a real-world electronics store of the Conrad[1] chain on 27 users. During the testing, our

[1] Conrad Electronic, Saarbrücken.

subjects were required to trial 23 different modality combinations, ranging from unimodal to overlaid and conflicting multimodal combinations. A unimodal modality is for example speech-only interaction, while an overlaid (and conflicting) modality combination is one in which (conflicting) input is provided through multiple modality channels such as speech and handwriting. The test scenario was that of buying a digital camera, and can be seen in Figure 1. User dialogues generally consisted of feature (e.g. price, mega pixels) and object (e.g. product name) information, for example: "What is the *price* of this camera <gesture=*EOS 10D*>".

Figure 1 – Usability study: Buying a digital camera.

2 Qualitative Usability Results

During the usability studies, our subjects provided insight into the advantages and disadvantages of the different modality combinations that were used while interacting with the digital camera objects. In this section, we provide discussion on the individual modalities. This is followed by a summary of criterion seen by our subjects to be relevant for multimodal interaction.

Speech: Some users found the camera names (i.e. the object referents) such as "PowerShot S1IS" and "FinePixA202" unintuitive to pronounce via speech. Other users expressed concern about their spoken dialects, despite the system correctly understanding them and despite being told to disregard system failures. Users mentioned that they would find the modality better if no buttons were required to start and stop the speech engine, which was observed to require training. A single press to start (and an automatic stop via silence detection) was also said to be better. Users stated that additional semantically similar dialogue structures for providing speech input would be an improvement, for example "what is the price of this?", and "what does this cost?", and both shorter and longer utterances such as "price?", and "what is the price of this camera?". One person in our study was left-handed and found the place-

ment of the start/stop button for speech – which was positioned on the left side of the PDA – to be cumbersome. Users did however also state that speech was an "excellent no-fuss modality combination". They said it was fast, comfortable, interactive, and one user made it very clear that she liked to talk.

Handwriting: Users commonly stated that it took too long to write and that the use of abbreviations for the feature and object names would be an improvement, e.g. "S70" instead of "PowerShot S70". When writing both feature and object information, the problems of handwriting were amplified, with users stating that the display space was too small. One user said that they would prefer not to have to write on top of the product images shown on the PDA's display. More than in the other modalities, users were extremely self-conscious of their handwriting and feared that their input would be falsely recognized by the system. For the number of objects present in our database (13), gesture was seen to be a better modality to use for object resolution. It was however noted that handwriting (as too intra-gesture) was unobtrusive, and thus perhaps better to use if privacy were required.

Intra-gesture: Regarding the use of intra-gesture for "feature" resolution, users commonly stated that the system would be better if they could see all of the options at the one time. It was stated that the modality was very fast, but that one had to wait at times depending on whether the relevant keyword was currently visible. Some users stated that they would prefer the text's font to be bigger and that the text should not scroll, while others mentioned that it might be better if the scrolling text's speed could be changed, especially for large data sets. The general consensus was that having to wait for a feature to become visible was not good and that the modality would suffer if the user was under time pressure. Intra-gesture for the "objects" was seen by most users as an excellent modality, and even likened to a natural reflex.

Extra-gesture: Several users mentioned that they liked to handle products before purchasing them. It was suggested that just pointing at an object (or using the PDA as a pointing device) would improve the usability, especially if the hands were also required by other modalities such as handwriting. In our scenario, one hand was required to hold the PDA and to start and stop the speech recognizer, while the other hand was required to differing degrees by the modalities intra-gesture, extra-gesture, and handwriting. Although the weight of our empty camera boxes was insignificant, large or heavy objects would pose a limitation for extra-gesture interaction. Some users stated that extra-gestures would be even better if the scenario entailed only the user and the objects (i.e. no PDA), or a headset instead of the PDA to free up the user's hands. It was stated that the location of the objects on the shelf should be mapped to a similar order on the PDA's display to ease finding objects when navigating in a mixed-reality world. Product placement (e.g. high-up or low-down products) was also seen to affect the use of the modality, and although extra-gesture was seen to be interactive, it was also one of the slower modalities.

Overlapped modality combinations: Some users formed fixed ideas early on in the study that all of these modality combinations were terrible. Users stated that the overlapped modality combinations were too complicated, took a lot of understanding and coordination, were time-consuming, and that the system should understand them without needing redundant information. Despite most of these modalities receiving a poor rating, our subjects were generally aware that redundant information might

benefit the system. Subjects for example said that overlapped information might be good to ensure that the input was correct, and to aid in the recovery of errors.

During the studies, users identified several multimodal interaction aspects as being important. These included comfort, enjoyment, familiarity, speed, scale, accessibility, privacy, intuition, and the complexity of a modality combination. Speech for example was seen as being ***comfortable*** to use in comparison to handwriting. Extra-gesture was described as being ***enjoyable*** by many users in comparison to intra-gesture where users said that "clicking is boring". Users were however very ***familiar*** with the modality of intra-gesture which closely resembles mouse interaction. The ***speed*** of handwriting was seen to be slow when compared to speech and intra-gesture, and the use of handwriting when overlapped with speech input caused users to slow down and slur their spoken input. The (perceived) accuracy of handwriting was also low, despite the recognized results being quite good. Speech and handwriting were said to ***scale*** better than gesture for large databases. Speech and handwriting were also said to be better if an object were not ***accessible*** (e.g. behind glass or out of stock). The obtrusive modalities speech and extra-gesture were seen to disregard ***privacy***, while handwriting and intra-gesture were noted as being better suited to dealing with sensitive objects. Privacy was stated to be a greater concern for object information than feature information. Some multimodal combinations (e.g. H,S and GI,S) and most overlaid modality combinations were seen to be less ***intuitive*** than their non-overlapped and unimodal counterparts. Several modality combinations also incurred ***complexity*** costs arising through modality switching, which was particularly evident for combinations consisting of on- and off-device interactions such as H,GE.

3 Conclusions

This paper highlights the concerns that users had while interacting in a mobile and multimodal fashion within a shopping environment. We discuss the advantages and disadvantages of the individual modalities, and outline a set of criteria that our users felt were important, ranging from comfort to privacy. This criterion may be taken as a starting point for future empirical studies.

References

1. Cohen, P., Johnston, M., McGee, M., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J.: Quickset: Multimodal interaction for distributed applications. In Proc. of ACM International Multimedia Conference (1997), pp. 31-40.
2. Oviatt, S.L. and Cohen, P.R.: Multimodal Interfaces That Process What Comes Naturally. In Communications of the ACM, Vol. 43, No. 3 (2000), pp. 45-53.
3. Wahlster, W.: SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions. In 1st International Workshop on Man-Machine Symbiotic Systems (2002), pp. 213-225.
4. Wasinger, R., Krüger, A., Jacobs, O.: Integrating Intra and Extra Gestures into a Mobile and Multimodal Shopping Assistant. In Proc. of the 3rd International Conference on Pervasive Computing (2005), pp. 297-314.