User Expectations and User Experience with Different Modalities in a Mobile Phone Controlled Home Entertainment System

Markku Turunen[1], Aleksi Melto[1], Juho Hella[1], Tomi Heimonen[1], Jaakko Hakulinen[1], Erno Mäkinen[1], Tuuli Laivo[1], and Hannu Soronen[2]

[1]TAUCHI, University of Tampere
Kanslerinrinne 1, FI-33014 University of Tampere
+358 3 3551 8559
firstname.lastname@cs.uta.fi

[2]IHTE, Tampere University of Technology
Korkeakoulunkatu 10, FI-33720 Tampere
+358 3 3115 3096
firstname.lastname@tut.fi

ABSTRACT

Home environment is an exciting application domain for multimodal mobile interfaces. Instead of multiple remote controls, personal mobile devices could be used to operate home entertainment systems. This paper reports a subjective evaluation of multimodal inputs and outputs for controlling a home media center using a mobile phone. A within-subject evaluation with 26 participants revealed significant differences on user expectations on and experiences with different modalities. Speech input was received extremely well, even surpassing expectations in some cases, while gestures and haptic feedback were almost failing to meet the lowest expectations. The results can be applied for designing similar multimodal applications in home environments.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces And Presentation**]: User Interfaces – *Input devices and strategies, Interaction styles, Haptic I/O, Voice I/O.*

General Terms

Measurement, Performance, Experimentation, Human Factors.

Keywords

Gestures, haptic feedback, speech recognition, SUXES.

1. INTRODUCTION

Current technology has enabled the development of rich mobile applications with modalities such as speech input, gestures, and haptic feedback working together with standard graphical and touch interfaces. At the same time, the ubiquity of mobile phones has brought along new potential application domains, such as using mobile phones as personal control devices for home entertainment, our focus in this paper. These new application areas provide numerous challenges, not only for technology, but also for usability and user interface research. Since the new applications bring new kinds of contexts and styles of use, the attitudes and expectations people have towards these systems are not well known. Working with home environments is particularly challenging, since strong value systems are connected with homes. For example, while some people are keen to try out any new technology to improve their home, others are reluctant to adopt technical solutions for everyday tasks. Instead, they appreciate media silence at home. There is an urgent need to know more about user expectations and experiences related to novel modalities in these scenarios.

MobileHCI'09, September 15 - 18, 2009, Bonn, Germany.
ACM 978-1-60558-281-8.

Previously, similar issues have been studied especially in the interactive television setting. Ibrahim and Johansson [4] proposed a novel TV program guide system that provides multimodal dialogue by combining speech interaction and direct manipulation with a remote control. Their results indicate that users prefer the multimodal approach to pure spoken input or pure direct manipulation, as the modalities are better suited for different operations and hence support each other. However, when speech input is used, its complexity might become a problem. Wittenburg and others [10] studied unrestricted speech input for TV content search and found retrieval performance to be critical to user experience, indicating that unrestricted speech input is only viable if the user expectations can be met. Therefore, it might be advisable to use a domain specific grammar and specific vocabulary rather than allow free form speech, in order to avoid negative experiences due to recognition problems. With restricted speech, however, the amount of Out-Of-Vocabulary (OOV) sentences might become a problem. Here, we show how grammar-based speech interface can be used to control digital television.

In addition to speech, gesture-based interaction can be very efficient in a media center interface. For example, gestures can be performed efficiently in low-light conditions, *e.g.*, when watching movies. Ferscha and others [1] investigated how gestures could be used to express the most frequently used remote control commands. They highlight the importance of simplicity, affordance, and focused functionality. This suggests that when adding gesture interaction to a multimodal home entertainment system, we should be cautious not to overload the gesture channel with too many commands. Furthermore, care should also be taken to map the gestures intuitively to the remote control commands they are replacing to create effective mappings. In our approach, we use a simple set of personalized gestures.

In multimodal mobile interfaces, and there are many situations where haptic feedback is more feasible than auditory or graphic feedback. For example, haptic feedback can provide silent, non-

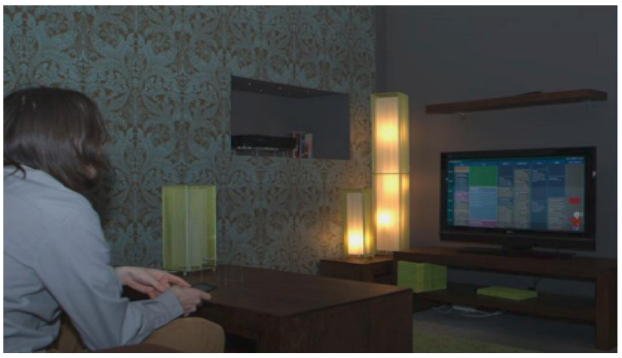

Figure 1. The system in use in a living room entertainment.

visible feedback about the speech recognition process, without disturbing the users or filling up the display with graphical icons, saving space for the actual content. We have found that even simple haptic feedback can be very useful in multimodal mobile user interfaces [5]. On the other hand, research has shown that learning more complex haptic patterns requires training [3]. Here, we use five different haptic patterns to provide feedback.

In this paper, we investigate user expectations and user perceptions of speech input, gesture input, and haptic feedback in a mobile phone controlled multimodal home entertainment system. The paper first introduces the home entertainment system and the multimodal interaction. Next, we describe a user evaluation that was carried out to study user expectations and experiences of the three modalities. After results, we conclude by discussing the implications for the use of multimodal interaction in home entertainment systems.

2. MULTIMODAL MEDIA CENTER

In TÄPLÄ (Ambient intelligence based on sound, speech and multisensor interaction) project we are developing methods for ubiquitous multimodal applications based on sound, speech, gestures, machine vision, and their rich multimodal use. In the beginning of the project, a large consumer survey was conducted to guide the design [7]. In the second phase, a prototype system was implemented and piloted in a public place. The implemented home media center includes a high-definition digital television program guide with full control over its content. The system utilizes a Nokia N95 mobile phone as a remote control. The pilot prototype demonstrates how speech, gestures, and haptic feedback can be used together. An example usage situation is shown in Figure 1. Users are able to control the media center by speaking, e.g., saying commands such as *"show me all children programs tomorrow"*, performing gestures by moving the mobile phone, and using mobile phone keys. In addition, haptic icons are used to provide five types of tactile feedback to the users. The system is described in more detail in [9].

A living laboratory containing the pilot application was constructed in a local media museum for the first public study, started in June 2008. It was used to gather an extensive amount of feedback from real users, with issues ranging from acceptability of the different input modalities to desirable features in such a system. Museum visitors could test the system and provide feedback using a web questionnaire and some participated in user studies. Additionally, user experiments with the system were carried out in controlled laboratory conditions.

2.1 Interface Modalities

For speech input, a 'push-to-talk' approach is used, *i.e.*, a user presses the left selection button of the mobile phone and utters a command. The recorded audio is sent to a server-based speech recognizer. The speech recognition grammar (a context-free grammar with about 900 words) includes navigation commands for the graphical interface (*e.g.*, *"go to program guide"*), navigation inside the electronic program guide (*"channel four"*) and also more complex commands. For example, there are spoken commands to record multiple episodes of a television series (*"record all Tom the Tractor shows this week"*) or highlight programs based on their genre (*"show me all the children programs"*). These commands were found highly appropriate among the test users during the pilot studies.

In addition to speech, commands can be issued with the mobile phone keypad (*e.g.*, arrow keys) or with gestures, by moving the mobile phone in specific patterns. The accelerometer-based gesture recognizer supports seven gestures: (i-ii) tilting of the phone up and down for moving the selection up and down on the screen, (iii-iv), swinging left and right for left and right movements, (v) swinging forward for selection, (vi) upwards for cancel, and (vii) shake to summon a pop-up with instructions. A combination of rule-based methods and Hidden Markov Model (HMM) based statistical model is used for gesture recognition, similar to the approach presented in [6]. The swing and shake gestures are recognized using HMMs, while the tilting gestures are recognized with rule-based methods.

The haptic feedback is given using the vibration component of the mobile phone. We defined a markup language to make different kinds of haptic patterns. We mapped the recognition results, user interface actions, and the state of the application into different patterns. For example, the phone vibrates when it recognizes a gesture, when recording of speech ends and when it receives a response from the speech recognition server. There are five types of haptic patterns, most types containing a pattern pair. In overall, there are nine different hapticons.

3. EVALUATION

In order to study the expectations and user experience of the system, and its different input and output modalities, we arranged a controlled user experiment in our usability laboratory.

3.1 Participants

We recruited 26 students (10 male, 16 female) from the local university to participate in the evaluation. They were aged 19 to 33 (mean = 22.6 years, *SD* = 3.0). The participants received extra credit towards the completion of an undergraduate course as compensation.

3.2 Method

We used a subjective user experience method SUXES [8] for the evaluation. SUXES is based on the modified SERVQUAL method [2]. It produces a subjective measure of the gap between the pre-test *expectations* and the post-test *perceptions* (experiences). The questionnaire contained various statements about the quality of the application and each of the modalities. For example, one statement was *"Speech input is quick to use"*. Before the use, participants mark two values between 1 and 7, an acceptable level and a desired level of quality for each statement. As its name implies, the acceptable level means the lowest

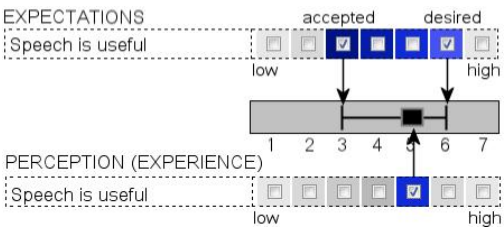

Figure 2. Interpreting expectations and perceptions.

acceptable quality level, while the desired level is the uppermost level, *i.e.*, there is no point to go beyond it. After the use, the participants mark the perceived level. Figure 2 illustrates the expectations, perceptions and the gap between them. In this example the user has marked 4 as the accepted level for speech input, 7 as the desired level, and 2 as the perceived.

The gap can be expressed using two disconfirmation measures, the *Measure of Service Superiority* (MSS) and the *Measure of Service Adequacy* (MSA). MSS measures the difference between the perceived level and the desired level, and MSA the difference between the perceived level and the accepted level. If experiences are in the range of expectations, MSS values are negative and MSA values are positive. The range of the accepted and the desired level is called the *Zone of Tolerance*. For the example in Figure 2, The Zone of Tolerance is <4, 7>, MSS is -5, and MSA is -2, meaning the perceived user experience is not within the Zone of Tolerance, but rather below the accepted level. The SUXES method is particularly suitable for iterative development, since it has indicates what the strong features of the application are, and where further development efforts are needed.

3.3 Procedure

Before the actual test, the participants were introduced to the home entertainment system application and its input and output modalities with a web-based wizard. The main features of the application were presented, but the actual usage instructions were not revealed at that point. Also, the participants were not informed that the test was related to input and output methods, but instead that it was a regular usability test to discover problems in the software. After the introduction, user expectations were gathered with the web-based SUXES questionnaire.

Each participant was given three exercise tasks and 11 evaluation tasks with the home entertainment system prototype. The order of task presentation was the same for each participant. The tasks reflect typical usage scenarios, *e.g.*, selecting a stored program, setting up recordings, and switching channels in the electronic program guide. Participants were free to use any of the input modalities to complete the task. After completing the tasks, they filled in a questionnaire consisting of the same statements they were asked in the pre-test questionnaire. This time the participants gave only one value to indicate their perceived experience.

4. RESULTS

As presented in Section 3.2, the SUXES method makes it possible to estimate what is the current state of the application based on expectations and experiences. We calculated the expected values and the perceived value for different questions (speed, pleasantness, clearness, error free use, error free function, learning curve, naturalness, usefulness, and future use) corresponding to each multimodal *input/output method* (speech input, gestures, and haptic feedback). Figure 3 shows the Zones of Tolerance across the dimensions using the median values for the acceptable level (lower bound) and desired level (upper bound), and perceived level (black circles). Area to the left from the perceived level represents MSA and area to the right MSS for each modality. Application of Friedman's test on perception values shows that there are significant effects of the input/output method across all dimensions. Pair-wise comparisons between modalities were carried out by using Wilcoxon signed-rank test with Bonferonni corrected levels of observed significance.

In terms of perceived levels, speech was rated significantly higher than gestures in all dimensions ($p < 0.01$). It was also rated as the most pleasant, clearest, easiest to learn, most useful and likeliest modality to be used in the future. No significant difference was found between speech and haptic feedback in speed, error-free use, error-free function or naturalness. Similarly, haptic feedback was considered faster, more pleasant, less error-prone both in use and in function than gestures. The rest of pair-wise comparisons between haptic feedback and gestures were not significant.

Here, we focus on the two dimensions that we consider the most critical for adoption – usefulness and subjective future use.

Speech was considered by far the most useful modality in comparison to gestures ($z = -3.8$, $p < 0.01$) and haptic feedback ($z = -3.7$, $p < 0.01$), while the difference between gestures and haptic feedback is not significant. Unsurprisingly, this effect is mirrored in the MSA and MSS values. It is interesting that haptic feedback and gestures rank very low and barely meet the lowest acceptable level, while speech exceeds expectations. In terms of perceived future use potential, speech also dominates gestures and haptic feedback ($z = -4.0$, $p < 0.01$ and $z = -3.4$, $p < 0.05$ respectively). The difference between gestures and haptic feedback is not significant. As with usefulness, also the MSA and MSS values show a similar effect for speech. It is noteworthy that again speech exceeds the desired level, while gestures fall below the acceptable level. Informal feedback from the participants supports these findings.

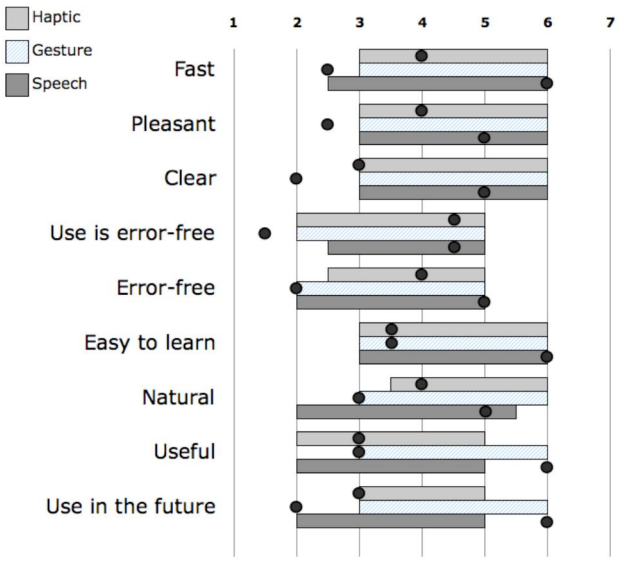

Figure 3. Ranges show Zones of Tolerance and circles the median perceived level for each modality across the dimensions.

5. CONCLUSIONS AND DISCUSSION

The results show that speech input was received very well as an input modality. Commands such as *"record all episodes of Friends"* were considered useful and the performance of speech recognition was perceived better than the upper limit of expectations, indicating that the participants seem to have been positively surprised. Speech recognition was, in fact, quite robust, with overall accuracy of 93%, and 97% when OOV sentences are removed. However, as previous studies in similar multimodal conditions have shown, there is no clear correlation between the speech recognition accuracy and user experience [5], which is the case here as well. More importantly, our results show that grammar-based recognition, *i.e.*, so-called restricted speech, can be used efficiently in this domain without user training. Otherwise, there would have been a significant amount of recognition errors due to OOV words as the utilized recognizer was not robust in rejecting OOV utterances.

A comparison to results from a previous study with the same experimental setup and a similar mobile application [5] shows that even if the expectations for speech input are mostly the same, the user experience of speech input observed in this study is superior, particularly when future use is considered. One reason for this is that in this scenario users found the speech input to provide additional benefit to the interaction that the other modalities could not easily offer, especially when issuing more complex commands. This suggests that mobile speech interfaces could also be highly desirable in other home environment scenarios, especially if the user is given the opportunity to issue logical spoken commands for tasks that normally require several key presses or the use multiple interfaces (*e.g.*, home automation).

Regarding other modalities, there were a lot of expectations for gesture input and haptic feedback, but both modalities were perceived unfavorably. In this setting gestures were designed as an alternative to keypad-based navigation. During use, participants often switched from the gesture-based navigation to the keypad navigation. The results indicate that major factors affecting gestures were errors, clarity of the commands and overall speed and pleasantness of use. This suggests that when gestures are used, they should be simple enough to be recognized robustly and add significant value to offset the cost they require over keypad use. Therefore, they should be used for more complex tasks than basic navigation, for example, to support macro level navigation from the current day to the next or current set of channels to the next. In the case of haptic feedback, we can conclude from the user feedback that even the fairly limited set of five types of haptic patterns is too much for this type of setting without training. Users were not able to associate them properly with the interaction, and they became too complex and annoying. Haptic feedback should be kept simple and not overused, unless clear mappings to other modalities, tasks or content can be found.

In our future work, the system will be expanded for other areas such as music, video, photo management, and environmental control. We hope to find new uses for the modalities. For example, gestures and haptic feedback might be better suited for tasks such as browsing photograph and music collections, where they can be mapped more efficiently to the content.

6. ACKNOWLEDGMENTS

This work was supported by the Finnish Funding Agency for Technology and Innovation (TEKES) under the Ubicom-programme in the "Ambient Intelligence Based on Sound, Speech and Multisensor Interaction"-project (TÄPLÄ, grant 40223/07).

7. REFERENCES

[1] Ferscha, A., Vogl, S., Emsenhuber, B., and Wally, B. 2007. Physical shortcuts for media remote controls. In Proceedings of the 2nd international Conference on intelligent Technologies For interactive Entertainment. ICST, Brussels, Belgium, 1-8.

[2] Hartikainen, M., Salonen, E.-P., Turunen, M., 2004. Subjective evaluation of spoken dialogue systems using SERVQUAL method. In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'04), 2273-2276.

[3] Hoggan, E. and Brewster, S.A. 2007. Designing Audio and Tactile Crossmodal Icons for Mobile Devices. In Proceedings of ACM International Conference on Multimodal Interfaces (ICMI'07). ACM Press, 162-169.

[4] Ibrahim, A., and Johansson, P. 2003. Multimodal Dialogue Systems: A Case Study for Interactive TV. Carbonell, Noelle; Stephanidis, Constantine (Eds.) Universal Access. Theoretical Perspectives, Practice, and Experience, 7th ERCIM International Workshop on User Interfaces for All, Revised Papers. Springer, LNCS, Vol. 2615. 209-218.

[5] Melto, A., Turunen, M., Kainulainen, A., Hakulinen, J., Heimonen, T., and Antila, V. 2008. Evaluation of predictive text and speech inputs in a multimodal mobile route guidance application. In Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '08). ACM, New York, NY, 355-358.

[6] Schlömer, T., Poppinga, B., Henze, N., and Boll, S. 2008. Gesture recognition with a Wii controller. In Proceedings of the 2nd International Conference on Tangible and Embedded Interaction (TEI '08). ACM, New York, NY, 11-14.

[7] Soronen, H., Turunen, M., and Hakulinen, J. 2008. Voice Commands in Home Environment - a Consumer Survey. In Proceedings of Interspeech 2008, 2078-2081.

[8] Turunen, M., Hakulinen, J., Melto, A., Heimonen, T., Laivo, T., and Hella, J. 2009. SUXES – User Experience Evaluation Method for Spoken and Multimodal Interaction. In Proceedings of Interspeech 2009.

[9] Turunen, M., Hakulinen, J., Melto, A., Hella, J., Rajaniemi, J.-P., Mäkinen, E., Rantala, J., Heimonen, T., Laivo, T., Soronen, H., Hansen, M., Valkama, P., Miettinen, T., Raisamo, R. 2009. Speech-based and Multimodal Media Center for Different User Groups. In Proceedings of Interspeech 2009.

[10] Wittenburg, K., Lanning, T., Schwenke, D., Shubin, H., and Vetro, A. 2006. The prospects for unrestricted speech input for TV content search. In Proceedings of the Working Conference on Advanced Visual interfaces (AVI '06). ACM, New York, NY, 352-359.