User Expectations and Real Experience on a Multimodal Interactive System

Kristiina Jokinen and Topi Hurtig

University of Tampere and University of Helsinki, Finland

firstname.lastname@helsinki.fi

10.21437/Interspeech.2006-156

Abstract

We present evaluation results of a multimodal route navigation system that allows interaction using speech and tactile/visual modes. Various functional aspects of the system were studied, related especially to the IO-modalities and their use as means of communication. We compared the users' expectations before the evaluation with their actual experience of the system, and found significant differences among various user groups.

Index Terms: evaluation, multimodal dialogue systems, mobile map interface

1. Introduction

Multimodal systems are usually considered advantageous over unimodal systems as they bring flexibility and naturalness to interaction. Users can choose the modality that best suits to their particular situation or preferences, and they may also use similar interaction strategies that they have learnt in human-human communication, so interaction is expected to become easier and more enjoyable. Multimodality also has synergy: interpretation accuracy can increase since information is encoded in redundant or complementary modalities (e.g. in noisy environments it is beneficial to combine speech recognition and lip-reading), and different modalities bring in different benefits (e.g. it is easier to point to an object than refer to it by speaking).

A common method for evaluating dialogue systems is to measure system performance and interview users to find out their subjective view of the usability of the system. E.g. in the Paradise framework [3], the overall objective is to maximize user satisfaction by maximizing task success and minimizing dialogue cost. Practical systems should also measure the quality of the service, i.e. there is a need to quantify the value of the system for the users [2]. Evaluation does not only deal with the system's performance as it is perceived by the users, but also with what the users desire or expect from the system.

The problem with standard evaluations is that objective and subjective criteria do not necessarily match, i.e. task success and user satisfaction may show opposite values. Paradoxically, users can tolerate problems and difficulties such as long waiting times and mere errors, if only the system is interesting and the users motivated at using it. The main issue then is to select appropriate design features and quality measures for the user's perception of the system. These deal with the understanding of the users and the underlying task, but also with the recognition of communicative principles to support easy, natural interaction and trust on the system's reliability to provide truthful information. Quality features are difficult to determine for spoken dialogue systems, however, and even harder for multimodal systems: it is not clear how to measure the impact of individual multimodal channels on the user experience as a whole.

The objective of this paper is to study various challenges related to the evaluation of interactive multimodal dialogue systems, and to report on the evaluation results of one particular system, MUMS [1]. We set to investigate the users' preferences over speech and tactile interface, how their expectations differed from their actual experience with the system, and if the users' age and gender influenced the expectations. The MUMS system is a mobile PDA-based route navigation system which allows the user to query public transportation information using spoken language commands and pen-pointing gestures on a map, and which provides route information in speech and graphical output. The system is described in more detail in [1].

2. Evaluation set-up

We had 10 male and 7 female test users, of different professions and aged between 23 and 61. They were familiar with computers in everyday tasks, but had varying levels of experience with speech and tactile interfaces. All were given a 15-minute crash-course on the system functionalities before the actual evaluation, and also a short hands-on training on the use of the system. All sessions were video-taped, and system logs were recorded both on the system server and the client application side. The tests were conducted indoors to minimize background noise.

The users were divided into two groups: the speech-group was instructed to interact with a speech interface which also has a tactile input option, while the tactile-group was told they will interact with a tactile system which has spoken dialogue capabilities. We expected a priming effect on the users' behaviour and expectations: prior knowledge of the system would have impact on the evaluation.

The users were given 7 scenario-based tasks to find suitable public transportation using MUMS. The tasks were designed to favour speech or tactile input (e.g. scrolling of the map vs. exact address) or to be modality-neutral, but the users were free to choose any combination of input modalities. The same set of tasks was given to both groups but in different order: the first task favoured the group's chosen modality, then a combination, then the other modality; the 4 last ones were modality-neutral. A sample task is given below:

> *Task 2: You car breaks down on the Länsiväylä Bridge just as you left Lauttasaari. The tow truck has arrived; find a way to Lehtisaari to pick up your children from school.*

The expected and observed system performance was measured by asking the users to fill in the same evaluation form twice. The users were asked to describe their expectations of the system right after the crash-course before they had any real experience of the system, and then assess the observed performance with the same form after the actual 7 tasks. The evaluation form contained 37 questions, organized into six groups concerning the user's perception of the system's speech and graphical inter-

face (1-11), the system's functionality and consistency (12-17), appropriateness of the responses (18-23), and the usability of the system, trust, consideration and easiness to complete the task (24-29). There were also questions of the user's eagerness to use the system in the future (30-35), and the overall assessment of the system (36-37). The answers had a scale from 1 to 5, with 5 signifying the highest mark.

3. Results

We investigated how age, gender, and prior knowledge of the system affect the user's expectations and actual perception of the system performance. We looked at the differences among the various user groups concerning the perceived functionality as well as in the overall change between expectations and perception, i.e. the measurement for how big the users' positive surprise or negative disappointment was as compared to their expectations. The differences between the groups were measured by T-Tests and One-way ANOVA, and the results are presented on the 0.05 significance level.

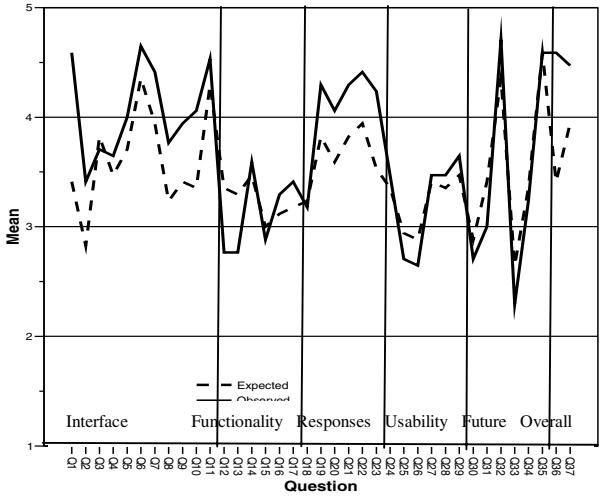

Figure 1 Expected and observed means for all test users.

Fig. 1 shows the mean values of the expected and observed features for all test users. The general tendency in the results is clear: the users' expectations are fulfilled, and the system seems to have offered a genuinely positive experience. The combination of speech and tactile gestures is considered very natural (questions 1-11), and the users especially enjoyed the combination of synthetic speech and graphical map as a means to receive route information (questions 20-21). As a whole, the users felt it was easy to learn to use the system, and were very enthusiastic about using it again in the future (questions 36-37).

The biggest disappointment is experienced with the system's speed and accurate indication of how quickly it serves the user (questions 12-13 in Fig. 1). This is corroborated by the fact that the average time elapsed from the end of user input to receiving the system's acknowledgement is about 8 seconds, and a further 25 seconds is needed to wait for the system's final response. It is possible to speed up the system somewhat e.g. by providing acknowledgement in parallel with the sending of the input to the server for analysis, but it is, however, difficult to shorten the time needed for input analysis and especially data transfer over the mobile GPRS network. Another negative experience concerns the system's ability to take the user into consideration (questions 25-26): although this is only a slight disappointment, it still shows that the users expect a more personalized and "intelligent" system.

When comparing unimodal use (questions 30-31 and 33-34) with multimodal use (questions 32 and 35), we notice that the users had been more positive in their expectations of the use of a unimodal system (speech or tactile) than what they experienced with the actual system. It is interesting that the users had not been overly optimistic of an unimodal system in the first place (average 3 in the scale) but their experience with the multimodal system confirms that even moderate expectations were too high.

3.1. Age

How about the actual differences? When studying the user's expectations in different age groups, we find that the middle-aged group (ages 33-48) stood clearly out in several cases (dotted line in Fig. 2): their expectations of the system performance and of the usability of the system's multimodal aspects are low. The young (ages 0-32) and the old (ages 49+) score generally higher, and have very similar expectations. When comparing expectations with the actual performance evaluation, no obvious disappointments or surprises were experienced in the groups, but interestingly enough, the scores for the middle-age group rose to the same level as the other groups, i.e. their actual experience of the system was rather positive.

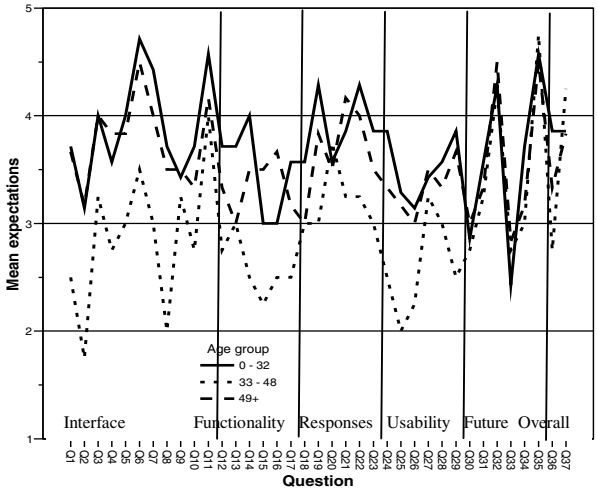

Figure 2 Expectations for age groups 0-32, 33-48 and 49+.

If the test subjects are divided into two groups (0-35 and 36+), statistically significant differences within the groups are found between expectations and perceived performance evaluation concerning speech-only interaction, the system's speaking voice, tactile interface, and the interest value of the system. E.g. the younger users (age < 35 years) were disappointed in their expectations about the system interest value, i.e. their evaluation of the perceived system was lower than what they had expected. For the older users, however, the system fulfilled their expectations and turned out to be more interesting than what they had expected. The situation is opposite when considering the future use of a speech-only system: expectations of the younger users are fulfilled whereas those of the older users decrease. It seems like the older users had a positive experience of the system as a whole even though an individual modality (speech) was dis-

appointing, whereas the younger ones were surprised at the level of technology of a single modality, but critical of the whole. The younger users also seem more content with using the tactile interface: they find it natural, responsive and usable (questions 19-23 in Fig. 3). In addition, they are willing to use the system unimodally, whereas the older group seems to enjoy multimodal aspects of the system (questions 30, 31, 33 and 34 in Fig. 3).

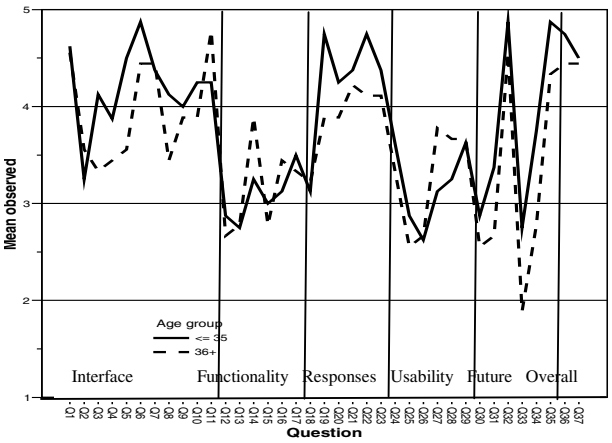

Figure 3. Observed means for age groups 0-35 and 36+.

3.2. Gender

We found only a few differences in the evaluations between the male and female groups. Gender does not affect the differences between expectations and perceived qualities, except in cases which concern the system's interaction capabilities (questions 24-28 in Fig. 4): female users seem to perceive the system more understanding and considerate than what they expected, i.e. they positively feel that the system understands what they say, takes their individual needs into account, etc. On the other hand, they are also very disappointed at their expectations concerning unimodal use of the map input and speech output.

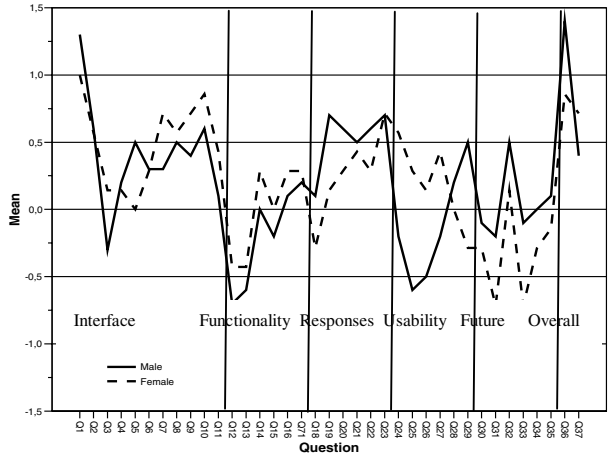

Figure 4. Observed-expected difference means for gender groups.

Additional differences between the male and female groups are found in the after-test comments and impressions: besides the system taking an individual user into account, the female group was surprised about the system's multimodal usability (questions 7-11 and 14 in Fig. 5), considered the system very easy to learn, and asserted a very clear enthusiasm about using the system in the future (questions 36 and 37 in Fig. 4).

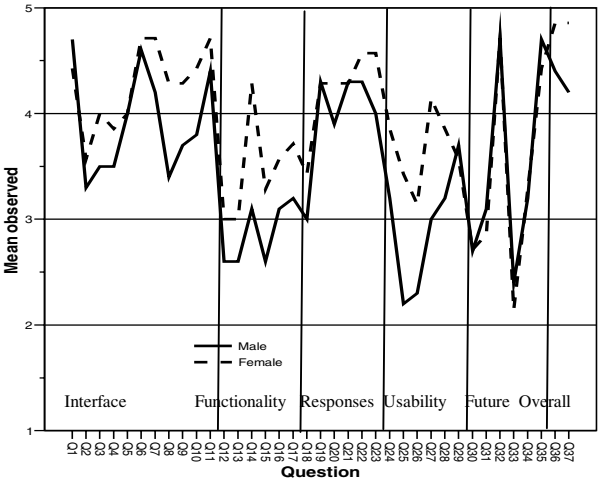

Figure 5. Observed means for gender groups.

3.3. Speech vs. tactile group

As expected, differences in evaluations of the speech and tactile group were found. The speech group gave very positive reviews for the system's multimodal aspects: they enjoyed using the tactile interface (questions 4-6 in Fig. 6), and felt that the system's speech and graphical representation contribute to the intelligibility of the system's output (questions 22-23). In addition, the speech group is more willing to use a tactile interface unimodally than the tactile group.

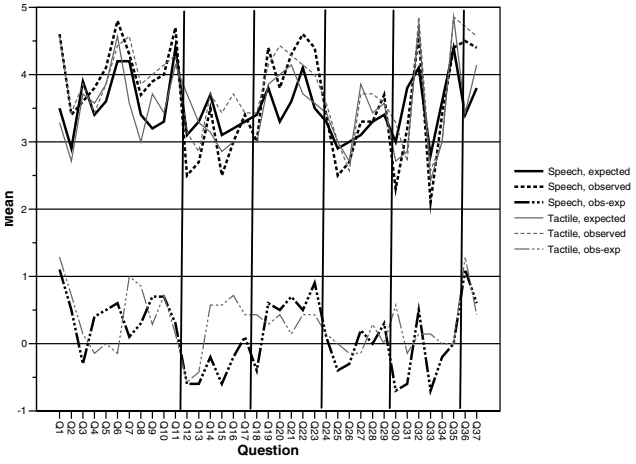

Figure 6. Observed, expected and observed-expected difference means for speech and tactile groups.

Also the tactile group felt that the possibility to use several modalities makes interaction flexible and the system easy to use. However, the tactile group is happier than the speech group with the system's performance (questions 12-18 in Fig. 6), especially with the rate of how often the system interprets user input correctly. There is also evidence that the tactile group is more

willing to use a unimodal speech system (question 30), although the results are not statistically significant. Interestingly, but also not statistically significant, the speech group seems to feel, more than the tactile group, that the system is slow, even though the response time of the system is not affected by the form of user input (question 12).

It is interesting that the priming effect came to play a role in the evaluation. The speech group perceived the use of the map more positively than the tactile group (it was pleasant to look at, it was intuitive to use), whereas the tactile group was rather critical at the map qualities, and in fact the difference between their expectations and the perceived system qualities was in absolute terms negative. Analogously, the tactile group was slightly more positive at the use of the speech input and output compared with what they expected, whereas the speech group was disappointed with the use of speech only system.

The two groups behave in the opposite way when evaluating the combined effect of speech and tactile output: the tactile group is significantly more positively surprised at the system capabilities than the speech group. The tactile group also has positive experiences concerning the system's performance: the system seems to show it is helpful, considerate, and cooperative, it functions in a consistent way and understands what the user says, and the interaction seems to succeed at the first try. Since the system was the same for both groups, and especially, the speech recognizer worked in a similar fashion for both groups, the differences in the user experience cannot be simply regarded as stemming from the inadequate performance of the speech recognizer. Rather, the differences seem related to the users' predisposition and their prior knowledge of the system, as set in the instructions and initial presentation of the system. If the user expects the main interaction modality to be speech, this brings in tacit assumptions of what a fluent human communication is like, and what can be expected from the system. Evaluation of a speech-based system thus suffers from the users' expectations of the system's understanding capability, even though speech may be only one of many modalities in the interaction. On the other hand, tactile systems have the tactile/visual modality as their primary mode of communication, and so speech can be enjoyed as an additional, interesting feature of the system functionality, without high expectations of spoken communication.

4. Discussion and conclusions

The user's perception of an interactive system depends on the system's communicative capabilities related to the task: natural intuitive interaction vs. quick and simple prompts. In this paper we investigated the users' experience of the multimodal route navigation system MUMS: the use of speech and tactile/visual modalities, their mixture, and the system in general. The users' expectations were fulfilled, and some system properties such as understandability and pleasantness of the output very positively experienced. However, the evaluations also showed that speech adds an extra difficult aspect to evaluation, since the users easily expect the system to possess more fluent spoken language capabilities than what is technologically possible. Tactile systems, on the other hand, can benefit from speech in quite a different way: their main interaction mode is not regarded as language-oriented communication, so speech can provide an additional value for the tactile interface users. Although the users' experience with tactile interfaces varied largely, it did not seem to have effect on the results. Of course, the users' prior experience with tactile interfaces is difficult to measure as it can range from common touch-screen cash-machines to fairly rare specific experience on pen-pointing PDA devices. All the users, however, were unanimous that a system with both speech and tactile/visual IO-possibilities is preferable to a unimodal one.

Evaluation focused especially on the user experience and on comparison of the users' expectations with their real experience of the system. We have shown that the users' predisposition towards the system and prior information about its capabilities affect evaluation, although the individual differences may not be generalized. In general, prior information makes the users more critical: expectations of the impeccable functioning of the highlighted properties are big, quite unlike if they are introduced as additional features. The newness factor plays a part, too: novice users are fascinated by the novel aspects of the system and tend not to pay so much attention to the practical usability.

Concerning the users' age, we found statistically significant differences in that the actual system seems to fulfil expectations of the older users better than those of the younger users, although the younger users were pleasantly surprised at the individual modality technology. Interestingly, the age group between years 33 and 48 had very low expectations about the system's performance and usability, but their experience with the system brought them to the same level with the other groups regarding the system's observed performance. As for the gender differences, female users tend to regard the system's "softer" characteristics more positively than the male users, and here the differences were statistically significant.

Our studies shows how individual users regard practical multimodal and interactive systems differently. Although the differences may not be always pinpointed down to prior knowledge, predisposition, age, or gender differences, it is important to notice that the goal of building one single practical system that would suit most users is not reasonable. Rather, there is a need for adaptive systems that allow the users to use different modalities when interacting with the system, and which can also tailor their responses according to the user preferences.

5. Acknowledgements

The main part of the work was done while the authors were employed by the Dept of Phonetics University of Helsinki. We would also like to thank the project partners in Tampere for their collaboration in the system development and the test users, especially the group from Helsinki City Transportation Authority for helpful comments and participation in the evaluation.

6. References

[1] Hurtig, T. and Jokinen, K., "On Multimodal Route Navigation in PDAs", Procs of 2nd Baltic Conf on Human Language Tech, pp. 261–266, Tallinn, Estonia, 2005.

[2] Möller, S., "A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems", In: Jokinen, K. and S. McRoy (eds.) Procs of 3rd SIGDial Workshop on Discourse and Dialogue, pp. 142-153. Philadelphia, U.S, 2002.

[3] Walker, M., Litman, D., Kamm, C. and Abella, A., "PARADISE: A framework for evaluating spoken dialogue agents", Procs of 35th ACL. pp. 271–280, Madrid, 1997.