INTERSPEECH 2009 BRIGHTON

SUXES - User Experience Evaluation Method for Spoken and Multimodal Interaction

Markku Turunen, Jaakko Hakulinen, Aleksi Melto, Tomi Heimonen, Tuuli Laivo and Juho Hella

Department of Computer Sciences, University of Tampere, Finland

firstname.lastname@cs.uta.fi

Abstract

Much work remains to be done with subjective evaluations of speech-based and multimodal systems. In particular, user experience is still hard to evaluate. SUXES is an evaluation method for collecting subjective metrics with user experiments. It captures both user expectations and user experiences, making it possible to analyze the state of the application and its interaction methods, and compare results. We present the SUXES method with examples of user experiments with different applications and modalities.

Index Terms: evaluation, subjective metrics, user experience

1. Introduction

Numerous different metrics have been used in evaluations of spoken dialogue systems and other speech-based and speech-enabled applications. Traditionally, the evaluation of technical components has been more extensive than usability evaluation. Larsen [1] claimed that "usability of voice driven services is still poorly understood" and much work still remains to be done. Usability can be based on certain objective measures but is also often measured with subjective evaluations.

There has been some work on questionnaires for collecting the subjective evaluations of speech and multimodal systems. Larsen [1] identified only two questionnaires that have been developed specifically for speech-based user interfaces and systematically address validity and reliability. These are BT-CCIR and SASSI [2]. Wechsung and Naumann [3] used SASSI alongside more generic SUMI, SAS and AttrakDiff questionnaires in an evaluation of multimodal systems. The results differed greatly between the questionnaires, questioning their reliability. In addition to working towards more reliable questionnaires, recently there has been interest towards understanding user experience in wider sense [4].

In this paper, we present SUXES, a method we have used and gradually improved over years to collect subjective data from users about speech-based and multimodal systems. It provides insights into what are user expectations and experiences of these technologies. Next, we present the SUXES methodology. This is followed by example studies from two recent user experiments carried out with the method. The paper ends with conclusions.

2. SUXES Evaluation Methodology

We have evaluated spoken [5] and multimodal [6] applications using subjective evaluation techniques based on modifications of a service quality metric called SERVQUAL. To apply the original method which was developed by marketing academics for real world services, we have adapted it for evaluation of novel interactive applications and different modalities in multimodal applications. We call the result SUXES. The purpose of SUXES evaluations is to capture both user expectations and user experience of different interaction techniques, or modalities, and the whole application. A set of questionnaires are used before and after the use of the application. In this way, we can measure the gap between the pre-test expectations and the post-test perceptions (experiences).

2.1. Procedure

The SUXES evaluation procedure is divided into four phases, and eight steps, as illustrated in Figure 1. The steps can be performed at different times, and the whole process is guided by a web-based wizard, which makes the procedure semi-automatic.

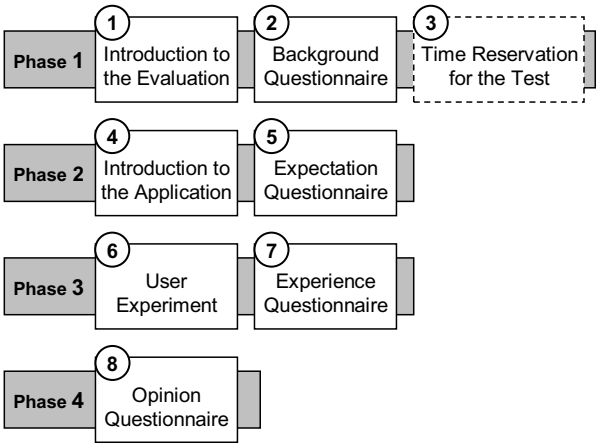

Figure 1: SUXES evaluation procedure.

2.1.1. First phase: Background information

In the first phase the necessary background information is collected from and given to participants with a web wizard. This can be done remotely or in the usability laboratory where the actual test takes place. We have used both approaches in different evaluations.

In the first step, the wizard introduces the aim of the evaluation following best practices in usability evaluation. This step can be replaced with a human introduction.

In the second step, the participant fills in a background questionnaire with items such as age and previous usage history related to the application domain.

In the optional third step, the participant makes a reservation for the actual test with the wizard. If phase 2 takes place at the same time as phase 1, this step can be performed after step 5.

2.1.2. Second phase: User expectations

In the fourth step, the web wizard introduces the application and its input and output modalities to the participant. The main features of the application are presented, but the actual usage instructions are not revealed at that point. When available, we have used the actual web-pages of the application to be evaluated. Multimedia introductions with videos have also been used. In all cases, it is important to give a realistic view of the application, but not expose too much detail. The latter one is important so that we can capture expectations accurately.

In the fifth step, user expectations are gathered with a questionnaire. The participant fills in the questionnaire based on the introduction they received in step four.

2.1.3. Third phase: experiment and user experience

Sixth step is the experiment where the participant uses the actual system to accomplish a set of tasks. The web wizard presents tasks descriptions to the participant one by one. If the participant has a question related to the task, a test conductor can provide answers as necessary. When ready, the participant clicks "proceed" on the wizard to continue to the next task. All our experiments have been arranged in a dedicated usability laboratory. The web-wizard has been running on a dedicated machine, and there has been an experienced test conductor present all the time. The test procedure, however, supports remote and mobile experiments, and this will be one of the key focus areas for us in the future.

In the seventh step, the participant fills in an experience questionnaire based on the actual use of the system. The questionnaire consists of the same statements as in step 5. This time the participant gives only one value to indicate their perceived experience.

2.1.4. Fourth phase: Feedback

In the eighth and final step, the participant fills in a feedback questionnaire. This questionnaire is usually designed separately for each experiment. There can be different questions related to the application and the test situation, as well as a possibility to provide feedback and comments. An interview can also take place at this point.

2.2. Questionnaires

Questionnaires are the key elements to capture user expectations and user experience. The SUXES method contains three main questionnaires and optional feedback questionnaire.

2.2.1. Background questionnaire

In the background questionnaire, basic demographic information such as age and gender are asked. In addition, participant's level of experience with the application domain, the devices used to interact with the application, and the methods used for interaction are asked. In these questions, we have asked the frequency of use, for example "*How often do you use speech recognition applications with mobile devices*". Five options, "Daily", "Weekly", "Monthly", "Yearly", and "Never" seems to capture the frequency of use adequately. With these options, we have found strong correlations between the previous experience and user expectations [6].

2.2.2. Expectations and experience questionnaires

The expectations and the experience questionnaires contain various statements about the quality of the application and each of the modalities used. Based on the original SERVQUAL questions and our experiences with evaluation of interactive multimodal applications, we have defined a set of nine statements for each item (application or modality) to be evaluated. The statements relate to speed, pleasantness, clearness, error free use, robustness, learning curve, naturalness, usefulness, and future use. For example, one statement is "Speech input is quick to use". It is noteworthy, that the same statements can be used for the overall application and for different input and output modalities.

In the expectations questionnaire, participants mark two values, an acceptable level and a desired level of quality for each statement. The acceptable level means the lowest acceptable quality level, while the desired level is the uppermost level, i.e., a participant considers there is no point to go beyond it. We have found a seven step scale to work well for all statements. The two resulting values form the Zone of Tolerance, where the user experience is expected to be in most cases.

After the user experiment, the participants mark the perceived levels for each statement to the experience questionnaire. The statements in the experience questionnaire are exactly the same as in the expectations questionnaire. This time, however, they give a single value for each statement according their actual perceptions of the use. This single experience value, perceived level of quality, can be compared to expectations values, the acceptable and desired level of quality.

Figure 2 illustrates user expectations and perceptions. In this example the user has marked 3 as accepted level, 6 as the desired level, and 5 as the perceived level.

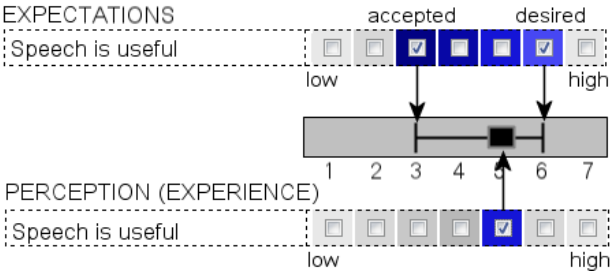

Figure 2: Interpreting expectations and perceptions.

2.2.3. Feedback questionnaire

The purpose of the feedback questionnaire is to get additional opinions and comments from participants. We have used, for example, Likert-scale questions and check-boxes to indicate the most preferred interaction methods etc. These are used, for example, to measure the reliability of the experience questionnaire.

2.3. Analysis measures

As presented previously, the expectations and the experience questionnaires produce three values for each statement in the questionnaires. Based on these values, there will be a gap between expectations and experiences. The gap can be expressed using two dis-confirmation measures, the Measure of Service Superiority (MSS) and the Measure of Service Adequacy (MSA). MSS is 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 (ie, the Zone of Tolerance). MSS values are negative and MSA values are positive.

For the example in Figure 2, the Zone of Tolerance is <3, 6>, MSS is -1, and MSA is 2, meaning the perceived user experience is within the Zone of Tolerance, matching user expectations rather nicely.

3. Case Studies

As an example of the SUXES method, we briefly describe our recent evaluations of two different multimodal systems. Their domains are completely different, the only common modality being speech input.

3.1. Travelman Application

Travelman [7] is a multimodal mobile application providing route guidance for public transport in Finland. It has two main functions: planning a journey and interactive guidance during the journey. In the journey planning phase, a user enters the departure and destination locations using one of the available input methods. The user evaluation was focusing on this functionality of the application.

For entering departure and destination locations, Travelman contains three input methods: speech input and two variations of text input (multi-tap, predictive text input). The language model for speech recognition and the domain-specific predictive text input consists of 11049 words. In addition to the GUI, the application includes tightly synchronized speech output. The application was running on a Nokia N95 mobile phone in the user experiment.

3.2. Media Center Application

The Media Center application [8] allows users to watch and record television broadcasts, listen to music, and view photographs. In the evaluated version the application provides full control over digital television content, including a novel full high-definition resolution electronic program guide (EPG). Users are able to control the media center with speech input, 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 tactile feedback. In the user evaluation, a Nokia N95 mobile phone was used to interact with the application.

3.3. Participants

38 people (27 male, 11 female) participated in the Travelman evaluation, while 26 people (10 male, 16 female) participated in the Media Center evaluation. Their age ranged from 18 to 45 years (mean = 23.7, SD = 5.7) and 19 to 33 years (mean = 22.6, SD = 3.0), respectively. In both cases, the participants were recruited from the local university, and received an extra credit towards the completion of an undergraduate course as compensation. The participant groups can be considered very similar, and they represent one of the most likely user groups of both applications.

3.4. Procedure

The evaluation procedure was almost the same for both cases, following the general SUXES methodology. There were minor exceptions related to the nature of applications and the aims of the evaluations. In the Media Center evaluation, the test took place right after the collection of user expectations, i.e., phases 1 and 2 were carried out at the same time. In the Travelman evaluation, the test took place one to two weeks after the introduction and the expectations questionnaire. We did not encounter any problems or differences in user behavior in the test situation, so a couple of weeks between the evaluation phases appear not to make major difference.

Naturally, the experiment tasks were different for both applications. In the Travelman evaluation, the participants were given four exercise tasks and 21 evaluation tasks, seven for each of the three input modalities (speech input, multi-tap text input, predictive text input). The participants entered departure and destination addresses with the input modalities mentioned, and checked the suggested routes. The evaluation was organized as a within subject study. The three tasks sets were the same for all participants and the order of modalities was counterbalanced. The task set-modality pairings depended on the group. The tasks were always presented in the same order within modality, and the addresses in each set were selected to keep task sets comparable.

In the Media Center evaluation, each participant was given three exercise tasks and 11 evaluation tasks. The tasks reflect typical usage scenarios, e.g., selecting a recorded program, setting up recordings, and switching channels in the EPG. The order of tasks was the same for each participant. Since we did not compare input modalities in the same way as in the Travelman application, but instead wanted to see how modalities are used, the participants were free to use any of the input modalities to complete the task.

In both tests, the participants were not informed that the test was related to input methods until in the end, but instead that it was a regular usability test to discover problems in the software. In both cases, all user actions and relevant application data were gathered with internal logging systems. This provided objective metrics, whose correlations with the subjective metrics from SUXES method we analyzed.

4. Results

We calculated the expected values, the perceived value, the Zones of Tolerance, and MSA and MSS values for each statement for both applications and each input/output method (multi-tap text input, predictive text input, speech input, speech output, gestures, and haptic feedback). Next, we summarize the main results focusing on two issues: (i) the overall user experience of the applications and their different modalities, and (ii) user experience of speech input (the only common modality in both evaluations). For more complete results, including explanation of statistical significances, see the corresponding evaluation reports ([6] and [8]).

4.1. Overall User Experience

In order to measure the overall user experience, we focus on the predicted use of the applications and their different modalities. This dimension forms a kind of "summarization" of the overall user experience. In our studies, it is the only dimension that can be explained by the other dimensions [6]. Figure 3 illustrates the Zones of Tolerance across the dimensions using the median values for the acceptable level (the lower bound of the grey area) and desired level (the upper bound of the grey area), and perceived level (black circles).

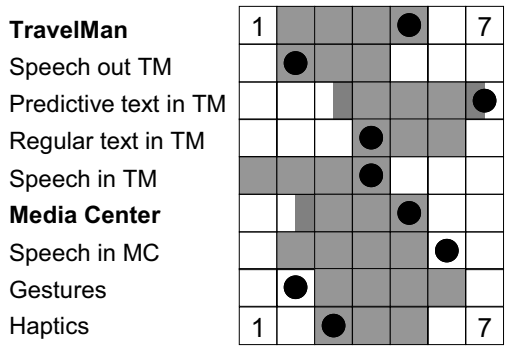

Figure 3: User expectations (grey cells) and user experiences (black circles) of the two applications and their different interaction modalities.

As illustrated in Figure 3, there are clear differences both in user expectations and experiences. The only "neutral" modality is multi-tap text input, which can be considered as a baseline for other modalities. Otherwise, it can be seen that domain-specific predictive text input is superior compared to other modalities in the Travelman application. This is noteworthy, since speech input outperformed it when performance was considered, i.e., it shows that objective and subjective metrics do not correlate [6]. The situation was different in the Media Center application, in which speech was considered by far the most potential future use modality, and it in fact did exceed user expectations. There are intuitive reasons for the results. In the Travelman application, speech input did not bring any real advantage over text input, even it was more efficient. In the Media Center application, high-level spoken commands provided a clear advantage over other input modalities (keypad and gestures).

Regarding other modalities, speech output, gestures, and haptic feedback rank very low and barely meet the lowest acceptable level (gestures, in fact, are below it). Again, the results are intuitive. For the regular test participants, these modalities did not offer any clear benefit. For visually and physically-impaired users, who are among our target groups, the situation might be different since these modalities are designed to support their usage, as reported elsewhere ([6] and [8]).

Finally, both applications were considered useful, and users are willing to use them in the future. This is important for the interpretation of the modality results, since it is not meaningful to evaluate poorly implemented applications.

5. Conclusions

SUXES, like the original SERVQUAL method, is particularly suitable for iterative development and prototyping, since it has been designed to provide information for further development efforts. Most importantly, it indicates what the strong features of the application are, and where further development efforts are needed. For example, in the case studies presented, clear conclusions can be made what modalities are useful in the current state, and where development efforts are needed.

One strong side of the SUXES method has also been that it is rather efficient, providing interesting data with reasonable effort. As described, we have used web-based questionnaires and instructions when possible. This minimizes manual labor in data processing and analysis. The use of web-based instructions during the evaluations further reduces the required human effort. While the method as it is used today still requires a test conductor, it is possible to run evaluations with just one person running the user experiments. Furthermore, with some further developments, we hope to enable even more automated and possibly remove human assisted testing. While not always applicable, such automation can significantly increase the amount of data that can be collected.

However, the most important feature of SUXES is the collection of user expectations. It provides context for the interpretation of user experiences. The expectations can show how significant the different factors are and by themselves already provide some insights into how people perceive new types of interactive systems. Thus, SUXES provides one method to better understand user experience. We have also been able to perceived changes in people's attitudes. Namely, attitudes towards speech input have improved in our experiments after users have had a change to use them [5].

In our future work, we focus on experiments with different modalities (e.g., physical browsing), further dimensions (e.g., joyfulness), experiments with different user groups, and further automated operation of the method. More information about SUXES can be found from its homepage: http://www.cs.uta.fi/hci/spi/SUXES/.

6. Acknowledgements

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

7. References

[1] Larsen, L. B. (ed.) Evaluation methodologies for spoken and multi modal dialogue systems. COST278 WG2 and WG3 Report, 2003.

[2] Hone, K. S. & Graham, R. Towards a tool for the subjective assessment of speech system interfaces (SASSI). Natural Language Engineering, Best Practice in Spoken Language Dialogue System Engineering, Special Issue, Vol. 6, Parts 3 & 4, September 2000.

[3] Wechsung, I. and Naumann, A. B. Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires. Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems (LNAI 5078), Springer-Verlag, 2008.

[4] Forlizzi, J. and Ford, S. The building blocks of experience: an early framework for interaction designers. In Proceedings of Symposium on Designing Interactive Systems, 2000, pp. 419 – 423.

[5] Hartikainen, M, Salonen E-P, Turunen M. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. Proceedings of ICSLP 2004: 2273–2276.

[6] Turunen, M., Melto, A., Hakulinen, J., Kainulainen, A., Heimonen, T. User Expectations, User Experiences and Objective Metrics in a Multimodal Mobile Application. In Proceedings of the Third Workshop on Speech in Mobile and Pervasive Environments, 2008.

[7] Turunen, M. Hakulinen, J., Kainulainen, A., Melto, A., and Hurtig, T. Design of a Rich Multimodal Interface for Mobile Spoken Route Guidance. In Proceedings of Interspeech 2007 - Eurospeech: 2193-2196, 2007.

[8] 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. Speech-based and Multimodal Media Center for Different User Groups. In Proceedings of Interspeech 2009.