

Subjective Evaluation of Mobile Speech Input: Experiences from Two Case Studies

Markku Turunen¹, Jaakko Hakulinen¹, Tomi Heimonen¹, Alekski Melto¹, Tuuli Laivo¹, Hannu Soronen², Juho Hella¹, and Juha-Pekka Rajaniemi¹

¹ Tampere Unit for Computer Human Interaction,
Department of Computer Sciences,
University of Tampere
FIN-33014 University of Tampere, Finland
+358 3 3551 8559

¹ firstname.lastname@cs.uta.fi

² The Unit of Human-Centered Technology,
Department of Software Systems,
Tampere University of Technology,
FIN-33101 Tampere, Finland
+358 3 3115 3839

² firstname.lastname@tut.fi

ABSTRACT

Thorough understanding of subjective and objective measurements of speech-based interaction, especially of its user experience, is vital for practical application development. We present findings from two case studies where multimodal applications containing speech input were evaluated using a subjective evaluation methodology. Responses were investigated using both correlation analysis and exploratory factor analysis. The results from the correlation analysis suggest that previous experience with multimodal technologies does not overall significantly influence expectations or perceptions of actual use. Results from the factor analysis process aided us in characterizing the nature of user experience and serve as a starting point for further analyses. Finally, our results suggest that there is little correlation between subjective and objective metrics.

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

Speech input, questionnaire, subjective metrics, correlation, exploratory factor analysis, SUXES.

1. INTRODUCTION

Even there are many evaluation metrics for speech-enabled applications, they are mainly targeted for technical and objective evaluation. For example, Möller [10] reports 36 different objective metrics. It is also common to collect subjective metrics with questionnaires, i.e., views and opinions from participants of the tests. Still, little is known on issues such as user expectations and user experience of speech input. In general, it has been claimed that “usability of voice driven services is still poorly

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

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

understood”, so improvement in the understanding of subjective measurement of speech-based user interfaces is vital for development [7]. This is particularly true with emerging application areas, such as mobile and pervasive speech applications.

To better understand the user experience of speech input in different application domains, more detailed studies of subjective metrics should be made. To address this issue, we performed an in-depth analysis to two datasets from evaluations of two different mobile speech-based systems using an adaptation of SERVQUAL questionnaire methodology called SUXES [16].

In the following, we describe the two applications used in the evaluation; a mobile phone controlled multimodal media center application, and a mobile phone-based public transport route guidance application, both featuring speech input. Next, the evaluation methodology is described along with a brief overview of the individual evaluations and their results. We conclude with results from the correlation analysis, exploratory factor analysis, correlation with objective metrics, and discussion of our findings and their implications.

2. SPEECH APPLICATIONS

We have evaluated two rather different multimodal applications using primarily subjective evaluation methods. The application domains are completely different, the only common modality being speech input. Both applications have been evaluated both in large-scale public pilot studies and laboratory experiments. Next, we briefly present the applications focusing on their speech input capabilities.

2.1 Travelman Application

Travelman [14] is a multimodal mobile application providing route guidance for public transport in Finland. It has two main functions: (1) planning a journey and (2) interactive guidance during the journey. In the journey-planning phase, users enter the departure and destination addresses or locations using one of the available input methods. The user evaluation focused 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 and domain-specific predictive text input). The language model for speech recognition consisted of 8896 street names and addresses, 2053 place names, and numbers from 1 to 100, totaling 11049 words. In addition to the GUI, the application includes tightly synchronized speech output, targeted for both regular and visually-impaired users.

2.2 Media Center Application

The Media Center application [17] 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, performing gestures by moving the mobile phone, and using mobile phone keypad. In addition, haptic icons are used to provide tactile feedback, and speech output is used to make the system accessible for blind and visually-impaired users.

The application has different language models for different usage situation and user groups. For general purposes, the speech input features commands for overall navigation in the application (e.g., “Go to program guide”), navigation inside the EPG (“Show Monday afternoon”) and for watching media (“Go to documentary channel”). It is also possible to record multiple episodes with a single utterance (“Record all the Tom the Tractor shows this week”), and highlight programs based on their genre (“Show me all the children programs tomorrow morning”). Speech recognition is implemented with context-free grammars. A grammar of about 900 words was used in the evaluated version.

3. EVALUATION METHODOLOGY

We used the SUXES evaluation methodology [16] for subjective evaluation of the applications. SUXES is based on the SERVQUAL method, developed by marketing academics in the seventies [11], which we have adapted first for spoken dialogue systems [4], and then for multimodal applications [15]. In this method, the main aim of the evaluation is to capture user expectations and user experience of different interaction techniques, as well as the whole application. To achieve this, sets of questionnaires are used to collect responses before and after the use of the applications in question. In this way, a measure of the gap between the pre-test expectations and the post-test perceptions (experiences) can be calculated.

3.1 Procedure

The evaluation procedure is divided into eight steps. In the first step, the aim of the evaluation is introduced, either using a Web-based wizard or a human moderator. In the second step, participants fill in a background demographic questionnaire. In addition, participants’ level of experience with the application domain, the devices used to interact with the application, and the methods used for interaction are queried. In the third step, the participant makes a reservation for the actual test. In the fourth step, the Web-based 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. In the fifth step, user expectations are gathered with the SUXES questionnaires asking both acceptable and desired levels of quality. Participants fill in the

questionnaire based on the introduction of the application they received in step four. Sixth step consists of the actual user experiment. The Web-based wizard presents the tasks one by one, giving exact task descriptions. In the seventh step, participants fill in the SUXES experience questionnaire based on the actual use of the system. The questionnaire consists of the same statements used in step 5. This time the participants give only one value to indicate their perceived experience. In the eighth and final step, the participants fill in feedback questionnaire, including general questions related to the application and the test situation, as well as open-ended questions for feedback and comments.

3.2 Questionnaires

The expectations and the experience questionnaires in SUXES 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 various dimensions of the experience: speed, pleasantness, clearness, error free use, robustness, learning curve, naturalness, usefulness, and future use. For example, one statement related to the speed of a modality 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, the participants mark two values, an acceptable level and a desired level of quality for each statement. As its name implies, the acceptable level means the lowest acceptable quality level, while the desired level is the uppermost level, i.e., there is no point to go beyond it. We have found a 7-point scale to work well for all statements. After the user experiment, the participants mark the perceived levels for each statement to the experience questionnaire using exactly the same statements as in the expectations questionnaire. This time, however, they give a single value for each statement according to their actual perceptions of the use.

Figure 1 illustrates user expectations and perceptions. In this example the participant has marked 3 as accepted level, 6 as the desired level, and 5 as the perceived level. 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 called the Zone of Tolerance. For the example in Figure 1, the Zone of Tolerance is <3, 6>, with the perceived user experience value of 5 falling within the Zone of Tolerance.

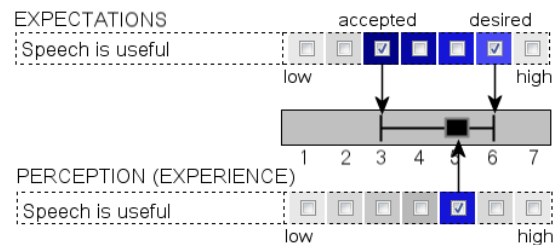


Figure 1: User expectations and perception questionnaire.

4. USER EVALUATIONS

The SUXES evaluation procedure was almost the same for both applications. In both cases, the participants were recruited from

the local university, and they received extra credit towards the completion of an undergraduate course as compensation for participating in the study. The participants were informed that the tests were regular usability tests to discover problems in the applications. In general, both participant groups can be considered similar, and they represent one of the most likely user groups of the applications in question.

4.1 Travelman Evaluation

For Travelman evaluation, 38 participants (27 male, 11 female) were recruited. Their age ranged from 18 to 45 years (mean = 23.7 years, SD = 5.7) 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 tasks were complete route planning sessions, i.e., users gave departure and destination addresses with the different input modalities, and reviewed 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 modality presentation was counterbalanced. The task set to 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 (based on the minimum error-free key presses required to enter the address).

4.2 Media Center Evaluation

For the Media Center evaluation, 26 participants (10 male, 16 female) were recruited. Their age ranged from 19 to 33 years (mean = 22.6 years, SD = 3.0). 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 changing channels in the electronic program guide. The order of task presentation was the same for each participant. We did not compare the input modalities against each other, but instead wanted to see how modalities are used. The participants were free to use any of the input modalities to complete the task.

5. RESULTS

Next, we summarize the main results from the two evaluations, focusing on subjective evaluation of speech input. More comprehensive results, including other modalities and factors, are reported elsewhere ([15] and [18]).

Overall, both applications were considered useful and the participants were willing to use them in the future as well, based on their exposure to the applications during the study. This is important for the interpretation of the results from individual modalities, since it is not meaningful to evaluate poorly implemented applications in the first place. If the overall user experience is poor, then it can be difficult to identify the potentially positive aspects of different modalities.

The following results concern only speech input modality, since giving commands by speech was the only common input modality for both systems. The speech recognizer, the mobile device, and the statements in SUXES questionnaires were exactly the same for speech input in both evaluations, enabling us to compare the results from the two studies. Figure 2 illustrates the results, showing the Zones of Tolerance across the dimensions of speech input and the perceived values (black dot/square), with all values reported as medians across all participants.

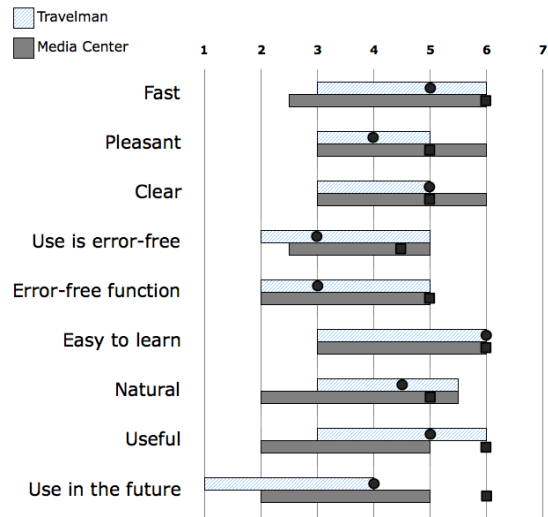


Figure 2: User expectations and perceptions of speech input for Travelman and Media Center applications.

5.1 User Expectations of Speech Input

As illustrated in Figure 2, user expectations of speech input for both applications are quite similar with some minor differences. The most notable differences concern usefulness and future use: user expectations concerning usefulness are higher for Travelman, while the situation is opposite for predicted future use.

In comparison to other modalities, the expectations for speech input are lower than for other modalities in both experiments, being in line with our large-scale (N = 1004) telephone survey, whereby there seems to be distrust towards using speech input in home environments [12] and public places.

5.2 User Experience of Speech Input

Overall, speech input was favorably received by participants in both experiments, as evidenced by the high scores given for usefulness and likelihood of using the modality in the future, in particular when compared to the expectations. While in the Travelman application the users were not too willing to use speech input in the future, in terms of absolute values, in the Media Center application they were, also far exceeding expectations. There are many potential explanations to the results. First, speech input was the most efficient input modality in the Travelman application, but it did not offer any clear advantage over text input, which was the most preferred input modality in this application. In the Media Center application, speech input provided clear benefits to the users. Second, in the Travelman application the participants experienced more speech input errors and system functionality related errors than with the Media Center. Still, the perceived values are clearly within the Zone of Tolerance.

5.3 Recognition Accuracy and User Experiences

In the Travelman application, speech recognition rates varied greatly between different users, ranging from 45% to 100%, with an overall recognition rate of 70%. In overall, these recognition rates and the high variance could be considered problematic. However, 97% of recognition task cases were completed successfully within three attempts, and speech input was superior

to text input even with this accuracy and slow response times [15].

In the Media Center application, speech recognition accuracy was higher, and there was not so much variation. Speech recognition accuracy varied between 80% and 100%, with an overall recognition accuracy of 93%. In general, this can be considered quite high, especially if OOV words and sentences are removed, which raises the overall recognition accuracy to 97%.

When comparing the objective metrics and subjective ratings, we could not find any meaningful correlations in either case. In the Media Center case, none of the subjective questions correlated with the actual recognition accuracy. In the Travelman case, there was a weak correlation between the recognition accuracy and the perceived accuracy of speech input. However, there was no correlation between recognition accuracy and the perceived usefulness of speech input in neither case. Finally, it does not make difference if people are divided in groups based on the recognition accuracy, as we did not observe a significant difference in the subjective ratings between users with high and low recognition accuracy. As an example, users with 100% recognition accuracy used the full scale in their subjective ratings.

Based on these two studies, we can conclude that speech recognition accuracy does not significantly affect the perceived experiences of users, provided that the recognizer functions at an overall acceptable level of accuracy. Even the recognition accuracy was significantly higher in the Media Center case, we believe, based on the data and interviews from both studies, other factors than recognition accuracy to explain most of the user experience ratings. Otherwise, the many users with 100% recognition accuracy in the Travelman case should have had different user experience than those with the significantly lower recognition accuracy.

5.4 Background Variables, User Expectations and Experiences

When evaluating applications based on relatively novel technologies such as speech input, it is also necessary to consider the effect of the participants' background variables to the expectations and experiences. For example, previous experience with speech input on mobile phones might affect expectations towards speech-controlled mobile applications such as Travelman.

In both evaluations we found individual significant correlations between the background variables and user expectations of speech input. For example, in the Travelman application, users with previous S60 experience had more positive expectations for pleasantness and clearness of speech input, and users with previous experience on other smart phones than S60 had higher expectations for speech input in general (see [15] for further details). In the Media Center application, users with S60 experience had more positive expectations for usefulness of speech input.

However, previous user experience does not seem to correlate with perceived experiences. In neither case there are significant correlations between the background variables and user perceptions.

To summarize, we found several correlations between background variables and user expectations when previous smart phone usage and speech input are considered. However, these correlations

seem to differ from application to application. For practical purposes, it is nevertheless useful to analyze these correlations, if only to be better informed about the whole context of the evaluation, and be able to properly frame the results from individual subjective metrics.

5.5 Exploratory Factor Analysis of User Experiences

In the factor analysis stage the perceived user experiences on different statements were examined in order to reduce the number of variables, i.e. to find higher-level construct factors that could explain the user experiences on a more general level compared to the individual statements.

The factors were extracted using the principal components method. Varimax rotation was used to maximize the sum of the variance of the loading vectors. Initial factorability was checked with Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. Both showed adequacy for factor analysis (KMO = 0.792/0.710, and Bartlett's test sig. = 0.000). For Travelman, all items were retained in the analysis based on the communalities, while in the Media Center case the variable "fast" was eliminated from the analysis due to low communality score (.481).

Table 1 shows the resulting rotated component matrix (rotation converged in 3 iterations) for Travelman, while Table 2 shows the same for the Media Center application. Loadings below .50 are hidden.

Table 1: Rotated component matrix for Travelman application.

	Factor	
	1	2
Fast	.891	
Pleasant	.649	.583
Clear		.827
Use is error-free	.601	.513
Error-free function	.714	
Easy to learn		.732
Natural	.654	
Useful	.704	
Use in the future	.837	

Travelman factor analysis resulted in two factors that explain 67.0% of the total variation: *efficiency and satisfaction of use and learning curve*. The former was loaded with the variables fast, pleasant, error-free use and function, natural, useful, and future use. The latter was loaded with clear and easy to learn.

Table 2: Rotated component matrix for Media Center application.

	Factor	
	1	2
Pleasant	.884	
Clear	.879	
Use is error-free		.915
Error-free function		.943
Easy to learn	.660	
Natural	.929	
Useful	.788	
Use in the future	.832	

Media Center factor analysis resulted in two factors as well, explaining 84.4% of the total variation: *error robustness* and *satisfaction of use*. The former was loaded with the variables error-free use and function. The latter was loaded with pleasant, clear, easy to learn, natural, useful, and use in the future.

As can be seen from the tables, factor analysis resulted in two factor structures in both cases, but the structures were different. We believe that fundamentally the changes in factor structures are due to the differences in the application domains and functionality, and the affordances of speech input as an interaction method in the two contexts.

In the Travelman experiment, the speech commands were more ambiguous, which could explain why the learning curve factor emerges as a distinct factor. Two variables, pleasant and error-free use are loaded into both factors, suggesting that they play a role in the ease of learning as well as the overall efficiency and satisfaction with use. On the other hand, in the Media Center evaluation the use of speech input was much less error prone than the other modalities, as evidenced also by higher recognition levels. Hence the error robustness factor encapsulates the two variables related to error-free use.

6. CONCLUSIONS AND DISCUSSION

We have presented two case studies with subjective evaluation of speech input in mobile devices. The results from the studies show major differences in terms of the suitability of speech input for the applications in question. In the Travelman application, speech was not the most preferred input method, even if it was superior in terms of efficiency. In the Media Center case, speech was perceived very positively, and it even surpassed the upper limit of user expectations. This was the first time during our more than ten years of experience from speech-based application development and evaluation when speech input was received extremely well after the use of the application. This shows that speech input has a huge potential when properly used and playing to the strengths of the application domain. To paraphrase Bill Buxton, in addition to getting the design of a speech-based application right, it is also needful to design the right speech-based application.

In order to find out what are the factors of subjective metrics, we used exploratory factor analysis to analyze the subjective responses to speech input modality in both applications. We were able to find factors that not only make intuitive sense, given our observations of application use during evaluations, but also capture the characteristics what we believe to be the salient features of user experience in these cases. For Travelman the key characteristics of use were learning curve and overall satisfaction and efficiency of use. We interpret these findings to mean that in terms of learning, the use of speech input was relatively easy from the users' perspective due to the clarity and simplicity of the commands. The other aspect, efficiency and satisfaction are highlighted by speed and naturalness of speech input. However, the factors are not clearly separated when it comes to error robustness and pleasantness. This indicates that there may be a negative interaction with the perceived satisfaction and number of speech input errors experienced, which affects both general use as well as learning. In the Media Center case the factors are clearly separated. Error robustness shows that users were able to use the application with a minimum of interruptions. Satisfaction of use indicates an overall high level of agreement with speech input as an interaction method in the media center context. Interestingly,

speed of interaction was not loaded into either factor, suggesting that it did not have a considerable effect on either facet. While the results from the factor analysis do not directly inform design, unlike the individual SUXES dimensions, we believe that it is a worthwhile step towards better understanding and characterizing the overall user experience in mobile speech applications.

Traditionally, recognition accuracy is considered one of the major factors for user satisfaction in speech-based interaction evaluation methods [19]. However, as the results here show, we could not find meaningful correlations between recognition accuracy and user satisfaction, or objective and subjective metrics in general. This suggests that unlike in speech-only applications, where accuracy is a significant usefulness factor, accuracy does not play significant role in multimodal applications, where the use of the application is not dependent on any single modality.

The findings from our studies are in general in line with other studies. For example, faster and less error-prone methods are not always the preferred ones, as has been shown in studies where users prefer system initiative dialogues over more efficient user initiative dialogues in spoken dialogue systems [20]. Some researchers even claim that user perceptions do not correlate with objective measures at all [5], and our results support this notion to some extent. The relationship between efficiency and user satisfaction is quite complex, it is unlikely that these attributes correlate with one another universally [2].

From a methodological standpoint, our results raise two interesting questions. First, the ecological validity of the evaluation conditions seems to play a significant role. In general, there are several challenges for evaluating speech input when user expectations are considered [9]. In comparisons of laboratory experiments and real usage of speech applications major differences have been found in earlier studies ([13] and [1]). The presented laboratory study of Travelman resembles a situation where a user is planning a route at home before starting the journey, but fails to provide any strong motivation for the use of speech input. In mobile situations, the hands- and eyes-free interaction may favor speech and limit the usefulness of other methods. For example, pen-based soft-key text input has been shown to be slower while walking [8]. Furthermore, in our laboratory conditions, speech input did not provide any added value. For future experiments, it will be a major challenge to assess the performance and user experience of multimodal speech applications in situations where a user is on the move in varying social and physical conditions.

However, the laboratory study results of the Media Center evaluation were comparable with our previous experiences with in situ evaluations. The Media Center application was available in a local Media museum in a living room environment for all the guests through May 2008 to March 2009. In the summer 2008, 21 user tests were held in the museum. In this study, the relation between expectations and experiences was also evaluated. The overall satisfaction with the speech interface was positive similarly to the findings from the laboratory experiment, and in contrast to the more skeptical expectations.

The second question relates to the perceived quality of use and the application in question. For example, Hassenzahl identifies two independent dimensions of product quality: pragmatic quality and hedonic quality [3]. Jetter and Gerken [6] have further developed Hassenzahl's model and introduce the user-product-relationship,

which includes traditional usability, functionality, hedonic quality and underlying user values. In the Travelman and Media Center applications, it is obvious that for the most part the users benefit from the system if it is functional: they can enter addresses more quickly or control the television more fluently than when using phone or remote control keypads. So, functionality and “cognitive” usability still override hedonic factors. Yet we have noticed that e.g. with the Media Center application, once the system is stable enough for basic use and more functions are introduced, the use becomes also hedonically motivated: since the speech interface allows e.g. physically disabled users to live more independently, it improves their abilities and feelings of equality and creates pleasure through success. Evaluating and differentiating between pragmatic quality and hedonic quality is difficult using the existing methodology, especially in laboratory experiments. Our future research will focus on examining these issues and augmenting the SUXES framework to better account specifically for the hedonic quality and user values.

7. ACKNOWLEDGEMENTS

This work has been supported by the Finnish Funding Agency for Technology and Innovation (TEKES) in the following projects: “Ambient Intelligence Based on Sound, Speech and Multisensor Interaction” (TÄPLÄ) and “New Methods and Applications of Speech technology” (PUMS).

8. REFERENCES

- [1] Ai, H., Raux, A., Bohus, D., Eskenazi, M., and Litman, D. Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users. In Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, 2007.
- [2] Frøkjær, E., Hertzum, M. and Hornbæk, K. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? Proc. CHI 2000, ACM Press (2000), New York, NY, 345-352.
- [3] Hassenzahl, M. 2004. The thing and I: understanding the relationship between user and product. In *Funology: From Usability To Enjoyment*, M. A. Blythe, K. Overbeeke, A. F. Monk, and P. C. Wright, Eds. Kluwer Academic Publishers, Norwell, MA, 31-42.
- [4] Hartikainen, M., Salonen, E.-P. & Turunen, M. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. Proceedings of ICSLP 2004: 2273-2276.
- [5] Hornbæk, K. and Law, E.L. Meta-analysis of correlations among usability measures. Proc. CHI 2007, ACM Press (2007), New York, NY, 617-626.
- [6] Jetter, H-C; Gerken, J. 2006. A Simplified Model of user Experience for Practical Application. NordiCHI 2006, Oslo: The 2ne COST294-MAUSE International Open Workshop "User eXperience - Towards a unified view".
- [7] Larsen, L. B. Assessment of spoken dialogue system usability – what are we really measuring? Proceedings of 8th European Conference on Speech Communication and Technology, Eurospeech 2003. ISCA: 1945–1948.
- [8] Mizobuchi, S, Chignell, M., and Newton, D., Mobile Text Entry: Relationship between Walking Speed and Text Input Task Difficulty. Proceedings of MobileHCI 2005, 2005.
- [9] Moore, R. K. Research Challenges in the Automation of Spoken Language Interaction. In Proceedings of Workshop on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005), 2005.
- [10] Möller, S. Parameters for quantifying the interaction with spoken dialogue telephone services. Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. ACL, 2005, pp. 166–177.
- [11] Parasuraman, A., Zeithaml, V.A. and Berry, L.L., “SERVQUAL: A multiple-item scale for measuring consumer perceptions.
- [12] Soronen, H., Turunen, M., Hakulinen, J. Voice Commands in Home Environment - a Consumer Survey. In Proceedings of Interspeech 2008: 2078-2081, 2008.
- [13] Turunen, M. Hakulinen, J., and Kainulainen, A., Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences. Proceedings of Interspeech 2006: 1057-1060, 2006.
- [14] Turunen, M. Hakulinen, J., Kainulainen, A., Melto, A., and Hurtig, T. Design of a Rich Multimodal Interface for Mobile Spoken Route Guidance. Proceedings of Interspeech 2007 - Eurospeech: 2193-2196, 2007.
- [15] Turunen, M., Melto, A., Hakulinen, J., Kainulainen, A., and Heimonen, T. User Expectations, User Experiences and Objective Metrics in a Multimodal Mobile Application. Proceedings of the Third Workshop on Speech in Mobile and Pervasive Environments (SIMPE), 2008.
- [16] Turunen, M., Hakulinen, J., Melto, A., Heimonen, T., Laivo, T., and Hella, J. SUXES – User Experience Evaluation Method for Spoken and Multimodal Interaction. Proceedings of Interspeech 2009.
- [17] 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. Proceedings of Interspeech 2009.
- [18] Turunen, M., Melto, A., Hello, J., Heimonen, T., Hakulinen, J., Mäkinen, E., Laivo, T., Soronen, H. User Expectations and User Experience with Different Modalities in a Mobile Phone Controlled Home Entertainment System. Proceedings of MobileHCI 2009 (to appear).
- [19] Walker, M., Litman, D., Kamm, C., and Abella, A. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97): 271–280.
- [20] Walker, M.A., Fromer, J., Di Fabbrizio, G., Mestel, C., and Hindle, D., “What can I say?: evaluating a spoken language interface to Email”, SIGCHI Conference on Human Factors in Computing Systems Proc., 1998.