Assessment of Spoken and Multimodal Applications: Lessons Learned from Laboratory and Field Studies

Markku Turunen, Jaakko Hakulinen, Tomi Heimonen

Department of Computer Sciences, University of Tampere, Finland
(markku.turunen,jaakko.hakulinen,tomi.heimonen}@cs.uta.fi

Abstract

In this paper, we present the key lessons learnt from numerous evaluations conducted to measure the quality of spoken and multimodal applications. The issues we address include the relation of laboratory and field studies, long-term and pilot evaluations, unimodality and multimodality, objective and subjective metrics, and user expectations and experiences. We present concrete case studies to discuss the above issues. For example, there are major differences in evaluating speech-only and multimodal systems. Similarly, there are major differences between laboratory and field studies, which need to be considered in successful evaluations.

Index Terms: spoken dialogue system, evaluation, service quality, multimodal applications.

1. Introduction

Several different metrics have been used in the evaluation of speech-based and multimodal speech applications. Quite often the focus has been on technical evaluation or traditional usability evaluation. In traditional evaluation approaches, focus has mostly been on objective evaluation metrics. When subjective metrics are used, they focus on practical usability aspects, such as efficiency and effectiveness, although there are other crucial success factors for speech-enabled services, such as hedonistic user experience. Only recently these have gained some notice, although the need for new evaluation solutions was identified in the community already some time ago (see e.g., [4] and [6]).

One of the most important aspects of evaluation is to understand the nature of speech applications. Most of the practical speech systems are services, and they should be evaluated as such, meaning that the overall service quality should be the primary measure of concern. In the future, it will be even more important to address the different dimensions of service quality, since traditional usability-oriented measures for task-based system are not able to address all relevant user experience factors in emerging application areas.

In this paper, we present several case studies from the perspective of service quality. Our evaluations are based primarily on the use of subjective user experience method called SUXES [9], which is based on a service quality approach originally developed for real world services, and later applied to speech interfaces [3]. It produces a subjective measure of the gap between the pre-test user expectations and the post-test perceptions (user experience), which has allowed us to study different quality factors in many ways.

Here, we present our findings from laboratory and field studies, experiences related to ecological validity of studies, long term evaluations, user expectation and experiences, evaluation of speech-only and multimodal systems, and the relationship between subjective and objective metrics. First, we introduce the applications and their evaluations that we base our analysis on, and present the key findings related to the above issues. We conclude with ideas for future work.

2. Case Studies

The evaluations discussed in the paper were conducted between the years 2003 – 2009 and involved three applications. The applications and the related evaluations are presented in brief to provide the necessary background to understand the evaluation challenges discussed in Section 3.

2.1. Stopman - Spoken Timetables

Stopman is a telephone based spoken dialogue system, which provides timetables for each of the about 1200 bus stops in the Tampere city area [11]. The aim of the system is to satisfy most of the callers with a minimum amount of interaction. At the beginning of a call, the system asks for a bus stop name and lists next busses for the stop, after which the user can browse the timetables. The Stopman application consists of 10 types of user inputs, with all functionality being available through speech and DTMF inputs. Additionally, the system provides assistance on how to use these modalities.

The Stopman system was evaluated in several usability tests and public pilots between 2003 and 2005. The system remained in public use after that, but no additional data was collected. All calls to the system during the period were recorded and logs were analyzed. There were different versions of the system available, since we iteratively developed Stopman based on the results from the evaluations. We analyzed the system based on a corpus of 1855 dialogues [11]. The collected data was divided into six categories based on the evaluation time and version. This allowed us to study the differences in the data collected during the initial use of the system, the long-term use, and the data collected in usability tests and from the deployed version of the system. Since we gained large amount of data from the public pilot, in addition to comparing the different data sets, we also addressed the issue of automatic log-file analysis.

2.2. Travelman – Multimodal Route Guidance

Travelman is a multimodal application for smart phones, which provides route guidance for public transport in Finland, such as metro, tram, and municipal bus services in cities and long-distance services in the rest of the country. It has two main functions: (1) providing support for planning a journey, and (2) giving interactive guidance during the journey. In the journey-planning phase, the user enters the departure and destination addresses or locations. Next, the system computes a set of applicable routes, allowing the user to browse them. After selecting a route, the user can listen how the journey progresses in real time, or navigate in the route description manually. Travelman has three input methods for providing the addresses: speech input and two variations of text input; regular multi-tap, and a predictive text input optimized for the relevant address data.

In order to study the different input and output methods, we arranged a controlled laboratory study with 38 participants.
We collected objective and subjective metrics in order to analyze the interactions and elicit feedback from the participants. We used the SUXES evaluation method to collect subjective ratings. The evaluation is presented in detail in [5].

2.3. Multimodal Media Center

The multimodal Media Center application [10] provides control over digital television content, including a novel full high-definition electronic program guide (EPG). It uses a mobile phone as a remote control and users are able to control the system with speech input, gestures performed by moving the phone, and by using the mobile phone keypad. In addition, haptic icons are used to provide tactile feedback, and speech output is utilized to make the system accessible for blind and visually impaired users. Finally, a physical touch interface was developed to study how near field communication (NFC) technology can be used in such a scenario [8].

The application has different speech recognition language models for different usage situations and user groups. For general-purpose use, the speech input includes commands for overall navigation in the application (e.g., “Go to program guide”), navigation inside the EPG (“Show Monday afternoon”) and for watching live TV (“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’s programs tomorrow morning”).

We have run several evaluations on the Media Center application. In the beginning, we evaluated the application in a public pilot study in a local museum over the duration of nine months. Museum visitors were free to use the system, and could provide feedback using a web questionnaire. Several user studies were conducted with museum visitors, as well. In order to evaluate the system in a controlled setting, we arranged two laboratory studies where the test participants conducted tasks representative of typical usage scenarios, e.g., selecting a recorded program, setting up recordings, and switching channels in the electronic program guide. In the second study, we used a modified version of the system, which included a different set of gestures, and physical pointing as an additional modality. Finally, we ran several long-term field evaluations in the homes of physically impaired users and users with different types of visual impairments.

3. Evaluation Challenges

The multitude of evaluations conducted with the three systems have provided us not only data about the usability and users’ experiences with the systems, but also general understanding of what we can gain from different types of evaluations. This includes, for example, differences between laboratory and field studies, evaluations of different durations, and the ecological validity of different evaluation approaches.

3.1. User expectations and user experiences

Subjective user expectations and user experiences play an important role in overall service acceptance. In most of our evaluations people had relatively low, and in some cases even skeptical, expectations towards speech input and output. As a first task in the development of the Media Center application we conducted a large-scale telephone survey (N = 1004) focusing on Finnish users’ concerns and ideas regarding speech input in home environments [7]. Our results showed major reservations towards its functionality and pleasantness, whereas people rated speech output extremely high. Similarly, people have shown distrust towards speech input in our different evaluations of the Media Center application, as well as in other application areas, such as in the case of the Travelman system. It appears this distrust is specific to speech as a modality, since people have lower expectations towards it than other novel user interface modalities, such as gestures, haptic feedback, or physical pointing. This has been especially evident in our studies on the Media Center application.

However, the actual user experiences show major changes in users’ attitudes after having used speech-based applications. Post-test experiences of using speech have been very positive in all of our evaluations. It is noteworthy that the situation can also be quite the opposite when novel interaction methods are introduced. For example, in the Media Center case we failed to initially meet the high user expectations for haptic feedback and gestures. This resulted in low user experience, both in absolute terms, and in particular when compared to the expectations. The situation changed completely in later evaluations after we redesigned the gesture and haptic interfaces based on the findings of the initial user study.

3.2. Objective and subjective metrics

The positive relationship between objective and subjective metrics is considered one of the cornerstones of evaluation of service quality. However, in the numerous studies with the Travelman and Media Center applications, we have not been able to find meaningful correlations between objective and subjective metrics. In the Media Center case, for example, none of the subjective metrics correlated with the recognition accuracy. In the Travelman case, we found weak correlations between recognition accuracy and the perceived usefulness of speech input. However, there was no correlation between recognition accuracy and the perceived usefulness of speech input in either case, nor did we observe a significant difference in the subjective ratings between users with high and low recognition accuracy. Even users with 100% recognition accuracy used the full seven-step scale in their subjective ratings.

Based on these two studies, we can conclude that speech recognition accuracy does not significantly affect the users’ perceived experience, provided that the recognizer functions at a reasonable level of accuracy. Even though recognition accuracy was significantly higher in the Media Center case than in Travelman, the data and interviews from both studies suggest that factors other than recognition accuracy explain most of the differences in user experience ratings. Otherwise, the many users with 100% recognition accuracy in the Travelman case should have had different user experience than those with significantly lower (e.g., < 50%) accuracy.

Similarly, when we compared background variables with user expectations and user experiences in the Travelman and Media Center studies, it was clear that some background variables influence user expectations. A representative example is familiarity with technology (such as smart phones). However, we could not find any correlations between background variables and user experiences. Furthermore, even the correlations between background variables and user expectations differ from application to application. These findings suggest that it would be extremely challenging, if not impossible, to measure and predict subjective system quality based on objective metrics or background information alone.

3.3. Speech-only vs. multimodality

In multimodal systems, the correlation between objective and subjective metrics can be particularly challenging to measure. In speech-only systems, the user is restricted to using speech. Here, traditional evaluation metrics, such as PARADISE, have
been successful in finding correlations between objective and subjective metrics. In multimodal systems the situation is remarkably different. As demonstrated in many of our studies, e.g., in the case of the multimodal Media Center, people switch between modalities quite rapidly if one modality fails or does not meet their expectations. Therefore, it is hard to provide equal exposure to different modalities in multimodal systems unless people are forced to try out each modality, which in turn creates its own problems. Furthermore, when we analyzed the results of our multimodal system evaluations on a per modality basis, in almost all evaluations the preferred modality dominates the results. This means that even if we can get very detailed and meaningful results for different service quality factors for the preferred modality (e.g., people consider it to be fast to use), it is hard to tease out those same factors for other, less preferred modalities: people are simply unable to differentiate between different aspects of the experience.

Finally, we have found major differences in how people use different modalities in multimodal applications in laboratory studies and field settings. In the evaluation of the Stopman timetable system, all functionality was available either with speech or touchtone (DTMF) inputs. When we compared how these modalities were used in usability studies versus their real world usage, we found nearly an order of magnitude difference. In usability studies only 6% of the calls contained touchtone inputs, while in the field study logs it varied from 30% to 55%. In further analysis of these differences, we found that people used touchtone input in fewer calls during the first month than the rest of the months. This suggests that especially when studying multimodal speech application, longer-term studies are likely to provide more accurate findings on modality use than short field studies or lab based studies.

3.4. Laboratory vs. field studies

The relationship between laboratory and field studies is getting more relevant with new more and more new applications targeted for mobile and pervasive use. We studied this relationship in the evaluation of the Stopman timetable system, and found numerous significant differences in the data collected during the initial use of the system, the data collected after the first month of service availability, and the data collected in usability tests. Most importantly, there were highly significant differences between real usage data and usability studies in almost all aspects. Furthermore, there were highly significant differences between the initial usage (first month) and the rest of the pilot usage in almost all aspects of the system use. These statistically (highly) significant differences include how people end the call (speaking or just hanging-up), the extent of functionality used, amount of help requests, ASR rejections, user interruptions, silence timeouts, the use of advanced functionality, the amount of repeat requests, and different modalities used during a call. For example, usability test participants tended to end the call in a polite way (“thanks, goodbye”), while in the real world they just hang-up after they have received the necessary information. Similarly, people request more help in user studies compared to real life situations (even in the initial usage stage). On the other hand, we were surprised to find that in the real world novice users interrupt the system more often than experienced users, similarly to laboratory studies.

The only significant difference in the data collected after the first two months of system use and subsequent use, however, was in the amount of help requests. This suggests that while usage patterns become stable quite soon, users still require more help during the first months after adoption.

There are many likely reasons for these results. For example, the users quickly learn to use the system effectively, and after a while hoax calls, where people are clearly just testing the system and toying with it, disappear. In usability studies, the tasks do not have real meaning for the participants, and sometimes they are even too accommodating. Finally, the contexts of use are quite different (e.g., people running to catch a bus behave differently compared to test participants sitting in a lab), which leads to the question of ecological validity, which is further discussed in the next section.

As a conclusion of this study, usability test data is more similar to data from initial usage sessions than that of long-term usage. Still, there are highly significant differences in behaviour between usability tests and the initial use of the system. This suggests that while usability tests are appropriate methods to validate the initial, baseline usability of an application, one should carefully consider whether to use the results when modelling interaction, for example. Measuring service quality based on laboratory tests or initial usage provides a totally different picture than experiences from long-term usage. These differences are due to the differing contexts of use and evolving usage patterns brought about by learning and experience in using the system.

Interestingly, our results correspond quite well with those of a later study with a very similar application and evaluation approach [1]. However, even if the main message is the same, i.e., laboratory results differ from field rest results, the details can differ, even to the extent that the actual findings can be opposite. For example, in [1] real world users used more touchtone input than test participants, similarly to our study. However, in their study the test participants requested less help, which contradicts our findings. Clearly, results depend on the application and other contextual factors and comparing findings between studies is challenging even if similar evaluation approaches are used.

3.5. Ecological validity and context of use

As demonstrated in the Stopman case and other similar evaluations, there can be significant differences in results between the laboratory and usability studies and real use of the system. This can be in part explained by the test setup. In the following evaluations we have focussed on the ecological validity of the evaluation setting, and its relation to user acceptance of the system and its overall service quality.

In the Travelman evaluation, our laboratory study was staged as a situation where the participant is planning a route at home before starting the journey, so in this case the situation was representative of real world use. However, after a post-mortem analysis, we concluded the setup failed to provide any significant motivation for the participants to use speech input and output. Taking this into account, the results are not surprising: even if speech input was the most efficient input method, people favoured text input, since it was familiar, and speech input did not provide any added value in laboratory conditions. This might change, however, when conditions change. In mobile situations, the need for hands and eyes free interaction may favour speech and limit the usefulness of other methods. For example, pen-based soft-key text input has been shown to be slower while walking.

Taking these experiences into account, we believe that it is unlikely that the same – or similar - findings can be found in laboratory studies when compared to field studies in mobile applications, excluding simple usability defects. In our experience, there are measurable differences in how a mobile system is used in any non-trivial field scenario versus laboratory conditions.
3.6. Public and long-term pilots

Since ecological validity and evaluation parameters play a major role in how well the evaluation results correspond to real world usage, it is beneficial to try to achieve as close as possible realistic conditions for evaluation. Public piloting of systems is one of the techniques we have used to this end.

When we compared the public pilot study of the Media Center application to the laboratory studies, we could not find any major differences. This led to an interesting question regarding ecological validity: are the results of the pilot study representative of real world usage; or is the only way to acquire realistic results to conduct long-term field studies, as indicated by the results of the Stopman evaluation?

In order to answer these questions, we conducted a number of longer-term field studies in the homes of users. Although the Media Center setup was quite different to the one used in the pilot study (for example, for physically disabled users the only input method was speech via wireless microphones mounted on their wheelchairs), the speech interface was essentially the same. The results confirmed the findings from the laboratory and public pilots. Since our users engaged in what can be considered to be real use (some of the pilot users even changed completely their television consumption habits during the pilot, favouring the Media Center application), we can conclude that if we succeed in creating ecologically valid conditions for the laboratory and public pilot studies, we can achieve results comparable to real world usage of the system.

4. Conclusions and Future Work

The evaluations presented in this paper have provided us not only with interesting data on the applications themselves but also insights into more general user reactions and attitudes, and an understanding of what we can learn from evaluations.

One surprising detail that came up more than once in our evaluations was that speech recognition accuracy, when adequate, does not significantly affect the users’ perceived experience. This is in conflict with the belief that recognition accuracy is a major factor of usability. In particular, with multimodal applications, the traditional approaches cannot capture all relevant aspects of user experience. Multimodality overall complicates evaluations, since to properly evaluate all modalities, participants must have motivation to use each of them to receive equal exposure and be able to make informed judgements. This is very hard to achieve in the laboratory and during short evaluations. It is our experience that participants tend to simply use the combination of modalities best suited for the evaluations setup.

While laboratory evaluations can be useful, especially if the setup and tasks match real usage conditions, only longer field evaluations can provide solid data on actual user needs and usage patterns. In our long-term studies, the usage patterns stabilized after the first month, while the first month data resembled patterns captured in laboratory evaluations with predefined tasks. One of the challenges in long-term evaluations is their labour intensiveness. We believe that automation of most of the data analysis will be one of the key factors for successful assessment of spoken system quality.

Since most of the practical speech-based systems are services, they should be evaluated as such. Especially since new application areas are increasingly about entertainment and sometimes even include social aspects such as companionship, evaluation methods and metrics should address this. For example, in our studies with multimodal entertainment applications we have found hedonistic values to form quite an important part of the overall user experience [2].

Our approach to evaluate applications as services is to use the SUXES method, which produces a subjective measure of the gap between the pre-test user expectations and the post-test user experience. This approach has shown that currently people have relatively low expectations towards speech input and output. However, the actual user experiences show major changes in users’ attitudes when they actually use speech-based applications. Post-test experiences of using speech have been very positive in all of our evaluations.

5. Acknowledgements

This work summarizes a number of evaluations, which have been conducted during several years across different projects. We thank all the people involved in these studies.

6. References