

Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences

Markku Turunen, Jaakko Hakulinen and Anssi Kainulainen

Speech-based and Pervasive Interaction group, TAUCHI, Department of Computer Sciences
University of Tampere, Tampere, Finland

{Markku.Turunen,Jaakko.Hakulinen,Anssi.Kainulainen}@cs.uta.fi

Abstract

We present findings from the long-term study of a speech-based bus timetable system. After the deployment of the prototype system we have collected data from real usage for 30 months. In addition, we have conducted usability tests to get subjective ratings of the pilot system. The comparison of these evaluations shows that the results obtained with usability tests differ significantly from those gained from the real usage, and the data of the initial use differs significantly from the data collected after that. For example, the differences in help requests, interruptions, speech recognition rejections, silence timeouts, and repeat requests are highly significant, and in some cases, such as explicit quit requests, enormous (65% versus 3%).

Index Terms: spoken dialog systems, evaluation, pilot studies

1. Introduction

We have developed multiple bus timetable systems in various research projects on top of the common Jaspis system architecture [1]. The lessons learned from previous spoken dialogue timetable systems suggest that open, user-initiative dialogue strategy based on the data collected from human-human interaction fails to provide robust enough interface [2]. Similar experiences have been gained with other studies [3]. In the case of the Stopman system we developed a task-oriented interface that provides the basic functionality in a system initiative manner, while the rest of the functionality is available with a user-initiative interface. We have conducted many usability studies to the system, and collected data in a 30-month public use with real users. In this paper we present the experiences gained from the different evaluations. In particular, we focus on the differences of data collected in usability studies, and data from different periods of real use.

2. The Stopman System

The Stopman system provides timetables for each of the about 1200 bus stops in Tampere City area. The aim of the system is to satisfy most of the callers with the first timetable listing, as demonstrated in Example 1. At the beginning of the call, the system requests the user to give a bus stop name (S_1). The most fundamental information is included in the initial timetable listing, which explains the length of the prompt shown in the example (S_3). After this, the rest of the functions are available. Functionality includes navigation in the timetable, selection of specific bus lines, and specifying a certain time (U_3).

S_1 :	Welcome to Stopman. You can ask help by saying "tell instructions". Please tell the name of the bus stop, for example "Central Square".
U_1 :	"Alexander Church".
S_2 :	Do you want to retrieve timetable for stop "Alexander Church"?
U_2 :	"Yes."
S_3 :	The time is 10:10. Today, the next buses leaving from bus stop "Alexander Church" are to "Vehmainen" number 5, now (list continues)... Please give your command, for example "tell instructions".
U_3 :	"Select line."
S_4 :	Please tell the line number. You can list the lines by saying "list lines".
U_4 :	"Thirteen."
S_5 :	Next number thirteen buses from "Alexander Church" are to Hermia in 1 minute, to "Ikuri", in (list continues)...

Example 1: An example dialogue with the Stopman system.

2.1. System functionality categories

The interaction with the Stopman system consists of 10 types of user inputs (Table 1). The mandatory input is the name or the number of a bus stop. It is not possible to have meaningful interaction without this, and all other input is regarded as optional. The second category of user inputs consists of the two ways to end the call (hang-up, explicit request). The rest of the categories include help and repeat requests, confirmations, advanced functionality (i.e., the functionality other than mandatory), user interruptions, and different error situations. All functionality is available with speech and DTMF inputs, and the system gives help on how to use these modalities.

	Description	Example
1	Mandatory functionality	"Main library"
2	End of call	"Thanks, goodbye!"
3	Help requests	"Tell instructions"
4	Repeat requests	"Repeat the last one"
5	Confirmations	"Yes"
6	Advanced functionality	"Select another day"
7	ASR rejections	<NOT RECOGNIZED>
8	Missing inputs	<SILENCE >
9	Invalid inputs	<INVALID DTMF>
10	User interruptions	<USER INTERRUPT>

Table 1: Stopman functionality categories.

3. Evaluation of the Stopman system

The Stopman system has been publicly available since August 2003. All calls to the system are recorded and logfiles have been analyzed. In addition, various evaluations have been done to the system to make it more efficient and pleasant to use. These range from a formal usability study to experiences collected from the users.

3.1. First version of the system

The Stopman system was tested in several usability tests during summer 2003, and experiences were collected in a three month public usage in August - October 2003. An improved version was released for public in November 2003. The Tampere City Transport Company had a promotion campaign in their web pages and in a newsletter that was delivered to all households in the Tampere area. In addition, an announcement about the system appeared in several local newspapers.

The first version was in public use for fifteen months from November 2003 to January 2005. The number of calls to the system was pretty stable after the first three months. The exceptions were July (the holiday month in Finland), and October and March, when the system was used by usability course participants. With this version of the system we collected 1062 dialogues (including the usability studies). The average number of calls per month was 124 for the first three months, and 52 for the rest of the months.

3.2. Second version of the system

The second version of the system has been in public use since February 2005. In this version a possibility to use names of the stops in addition to stop numbers was included. Otherwise, this version was similar to the first version. We have included to this analysis 793 dialogues from fourteen months between February 2005 and March 2006 (including the usability studies). The number of calls to the system was similar to the first version, on average there were 52 calls per month again, July being an exception.

3.3. Usability course tests

The system was tested with the participants of the course "Introduction to Interactive Technology" in October 2004 and October 2005. In 2004 the participants were asked to call to the system to accomplish given tasks. Each task required one call to the system. The tasks were rather simple. The main purpose was to get feedback from the use of pauses in system prompts. The calls were made to the same publicly available system. This means that most of the calls made in October 2004 are test calls. During March 2004, the system was used as an introduction to speech applications in a study of another spoken dialogue system [4]. In the October 2005 study the participants performed a bit more complex tasks and data was collected with a copy of the system running on a separate server.

3.4. Description of the data

In total, we collected 1855 dialogues with the system. The collected data is divided into six categories. In the first category there is the data from the first month of usage (November 2003). The second category consists of the data

from the second month (December 2003). The third category includes the data from the rest of the months of the first version (January 2004 – January 2005), excluding the months with usability studies (March 2004 and October 2004). The Fourth category contains the data from the months with mixed data from the real usage and the usability studies (March and October 2004). The fifth category includes all data from the real use of the second version (February 2005 – March 2006). Finally, the sixth category contains the data from the October 2005 usability study. The categories are presented in Table 2.

	Description	Dialogues	Time
1	First month	127	11/2003
2	Second month	87	12/2003
3	First version	630	1/2004-1/2005
4	Mixed data	218	3/2004 & 10/2004
5	Second version	725	2/2005 – 3/2006
6	Usability study	68	10/2005

Table 2: Data categories.

4. Results

The results of the pilot usage and usability tests are presented according to the functionality of the system (Table 1) and the data categories (Table 2). All figures present the amount of dialogues where the functionality in question is present. In overall, there are highly significant differences between the data categories in each of the system functionality with one exception: there are no significant differences in the case of invalid system inputs. Next we present the differences in more detail. All differences mentioned are either *highly significant* ($p < 0.001$) or *significant* ($p < 0.01$) by Chi Square test, unless otherwise noted.

4.1. Ending the call

There are *highly significant* differences concerning how the call was ended. During the first month, 17% of the calls contained an explicit request to end the call (e.g., "thank you and goodbye"), as illustrated in Figure 1, while in the other calls the users simply hang up. This decreased to 5% during the first version, and 9% during the second version (this is one of the few *significant* differences between the system versions).

There are *highly significant* differences between the real use and usability tests: during the months when there was mixed data from pilot usage and usability studies, the amount was 15%, and during the October 2005 usability study the amount was 65%. The amount of confirmations follows this pattern, since the end of the call was confirmed explicitly.

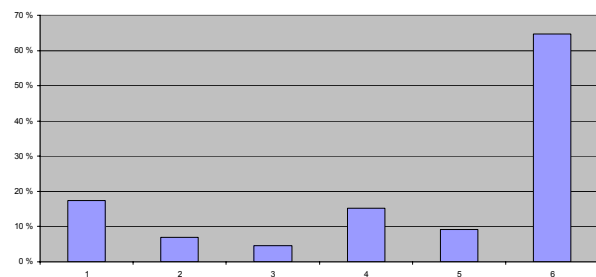


Figure 1: Explicit quit requests.

4.2. Mandatory functionality

The proportion of calls not containing mandatory functionality is illustrated in Figure 2. There are significant differences between the first month and other categories ($p < 0.05$), except between the first month and mixed usage. As illustrated, during the first month almost 20% of the calls did not contain mandatory functionality, i.e., users did not get any timetables, while in the second month only 6% of calls were such.

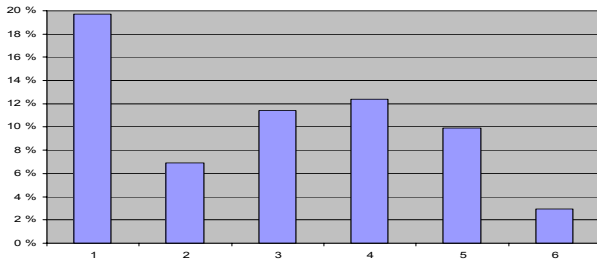


Figure 2: Calls without mandatory functionality.

4.3. Help requests

The amount of help requests is illustrated in Figure 3. The use of this functionality had a *highly significant* decrease after the first two months. In the first month, 17% of the calls contained explicit help requests, while after two months only 6% of calls contained help requests. In the categories with usability studies help was requested in more than 25% of the dialogues. It is noteworthy, that this is the only case with a *significant* difference between the second month and the rest of the use.

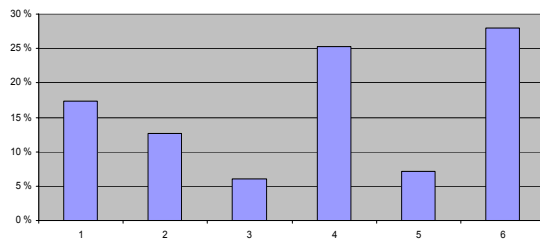


Figure 3: Help requests.

4.4. ASR rejections

There are no differences in speech recognizer rejections between the initial usage and the rest of the months. However, there are *highly significant* differences between the usability test categories and real usage, as illustrated in Figure 4.

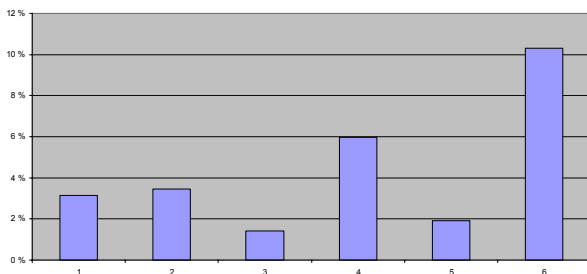


Figure 4: ASR rejections.

4.5. User interruptions

The amount of user interruptions decreased *highly significantly* after the first month, as illustrated in Figure 5. During the first month, 28% of the calls contained user interruptions, while after the second month it was only 7%. This was against our assumptions, since we assumed that experienced users interrupt the system more often. This behavior was encountered again when there were calls from the participants of usability studies: the proportion of calls with interruptions (26%) resembled the first month. In this case the data from the second system version is not comparable with the first version since user interruptions were not logged anymore, and thus the second version is not considered here.

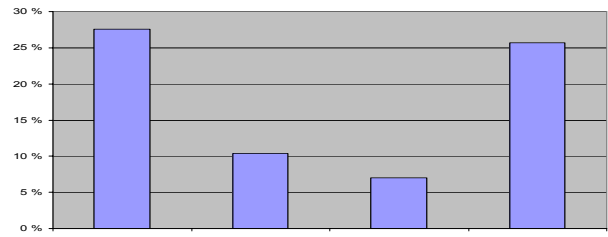


Figure 5: User interruptions.

4.6. Silence timeouts

There were *highly significant* differences in silence timeouts, i.e., when no speech was detected during the user turns in a given time. This is illustrated in Figure 6. The amount of calls with silence timeouts was very high (24%) in the first month, but decreased to 3% within two months. This is the first of two cases with a *highly significant* difference between the two versions. Again, the usability tests show major differences: there were silence timeouts in 21% of the calls during the mixed usage, and 37% during the October 2005 usability test.

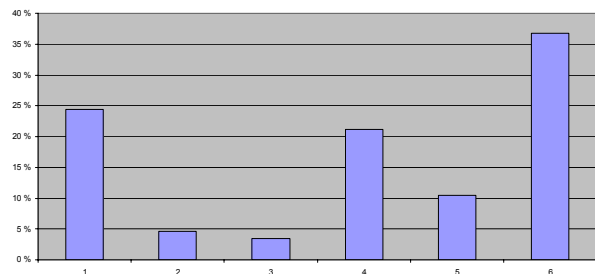


Figure 6: Silence timeouts.

4.7. Advanced functionality

There were no significant differences in the use of advanced functionality between the real usage conditions. On average, only 11% of the calls contained advanced functionality. This means that almost 90% of the callers never used the system's capabilities beyond the mandatory functionality. There were *highly significant* differences between the real use and the usability tests. The amount of advanced functionality in usability tests is not surprising, since in most cases this was needed to accomplish the given tasks. Still, the tasks were anything but artificial. For example, in October 2005 the users needed to select another time instead of the present moment.

4.8. Repeat requests

There were no significant differences in the proportion of calls with repeat requests during the real usage. However, there were *highly significant* differences between the usability studies and real usage, as illustrated in Figure 7. To generalize, real users do not ask to repeat the last system utterance, while this is quite common in usability studies.

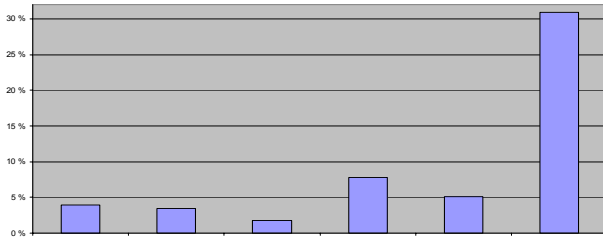


Figure 7: Repeat requests.

4.9. Modalities

Figure 8 depicts how different modalities, i.e., speech and touchtone (DTMF) inputs were used for other purposes than interrupting the system prompts. There are several *highly significant* differences. To summarize, people used touchtones in fewer calls during the first month than the rest of the months, while speech was used more often during the first month. Between usability studies and real usage this difference is even greater. In the October 2005 usability study only 6% of the calls contained touchtone inputs. This is the second case with a *highly significant* difference between the versions.

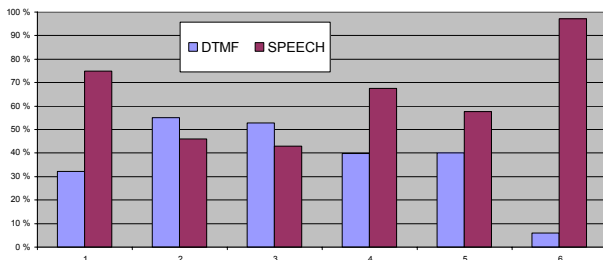


Figure 8: Modalities used.

5. Discussion and Conclusions

We have presented results from a 30-month public usage of a spoken dialogue system. We found significant differences in the data collected during the initial use of the system, the data collected after the first two months, and the data collected in usability tests. There are highly significant differences between the first month and the rest of the pilot usage in almost all aspects of the system use. The only significant difference between the second month and the rest of usage, however, was in the amount of help requests. This suggests that the usage becomes stable quite soon, but the users still require more help during the first months. We found no difference between the first month of the second version of the system and the rest of its usage. The differences between the system versions were quite few, while the differences between real use and usability studies were extremely high in almost all aspects.

In this study all calls to the system were included. There were, however, calls that contained no successful user interaction, as perceived by the system. When these calls are removed, *the differences are still significant in all categories*, even in the categories concerning mandatory functionality and silence timeouts: their differences are smaller, but still significant.

There are many likely reasons for the results. For example, the users learn to use the system more effectively, and after a while hoax calls disappear [5]. In usability studies, the tasks may not have real meaning for the users, and the users are sometimes even too co-operative. According to our studies, the usability test data is more similar to data from initial usage than that of later months. Still, there are highly significant differences between the usability tests and the initial usage. This suggests that usability tests are appropriate methods to improve the initial usability of an application, but it should be considered carefully when the results can be used to model interaction, for example, with machine learning methods. In overall, the evaluation of spoken dialogue systems is quite challenging, as pointed out by many researchers (e.g., [6] and [7]).

6. Acknowledgements

This work is supported by the Ministry of Transport and Communications and the Tampere City Transport Company (HEILI-programme, the 1st version), and the Technology Development Agency of Finland (FENIX-programme, PUMS-project, the 2nd version). We like to thank Leena Helin, Esa-Pekka Salonen and Natalie Jhaveri for their contribution in the implementation and the user tests.

7. References

- [1] Turunen, M., Hakulinen, J., Rähkä, K.-J., Salonen, E.-P., Kainulainen, A., Prusi, P. An architecture and applications for speech-based accessibility systems. *IBM Systems Journal*, Vol. 44, No. 3: 485-504, 2005.
- [2] Turunen, M., Hakulinen, J., Salonen, E.-P., Kainulainen, A., Helin, L. Spoken and Multimodal Bus Timetable Systems: Design, Development and Evaluation. *Proceedings of 10th International Conference on Speech and Computer (SPECOM 2005)*: 389-392, 2005.
- [3] Raux, A., Langner, B., Bohus, D., Black, A. W., Eskenazi, M. Let's Go Public! Taking a Spoken Dialogue System to the Real World. *Proceedings of Interspeech 2005*: 885-888, 2005.
- [4] Hartikainen, M., Salonen, E.-P., Turunen, M. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. *Proceedings of ICSLP 2004*: 2273-2276, 2004.
- [5] Johnsen, M. H., Svendsen, T., Amble, T., Holter, T., Harborg, E. TABOR - a norwegian spoken dialogue system for bus travel information. *Proceedings of ICSLP 2000*, 2000.
- [6] Larsen, L. B. Assessment of spoken dialogue system usability - what are we really measuring? *Proceedings of Eurospeech 2003*: 1945-1948, 2003.
- [7] Möller, S. Towards generic quality prediction models for spoken dialogue systems - a case study. *Proceedings of Interspeech 2005*: 2489-2492, 2005.