USABILITY EVALUATION OF IVR SYSTEMS WITH DTMF AND ASR

Cristina Delogu, Andrea Di Carlo, Paolo Rotundi, Danilo Sartori

Fondazione Ugo Bordoni, v. B. Castiglione 54 Roma, Italy 00142

e-mail: cristina;adicarlo@fub.it

ABSTRACT

The paper presents an usability evaluation of 4 different prototypes of the same IVR application: three of them are automatic speech recognition (ASR) based and the other is Dual Tone Multi Frequency (DTMF) based.

Our work consists of the automation of a service currently in use with an operator who provides information about new facilities offered by Telecom Italia, such as call waiting and call forwarding. The usability of the different prototypes has been evaluated through objective and subjective measures. Objective measures such as task completion and correctness, number of calls for task, transaction time, number of turns, and recognition accuracy have been obtained through the system's logfiles and the recorded speech utterances.

To gather subjective measures the users were asked to fill in a questionnaire about their perception of the quality of the overall interaction, their effort in interacting with the system, and their satisfaction with different features of the system.

In general a good correspondence between objective and subjective measures was found. The latter always confirmed and illustrated the results obtained with the former.

1. INTRODUCTION

In the last ten years simple operator services have been automated using Interactive Voice Response (IVR) systems. To date, speaker-independent ASR technology using small- and medium-sized vocabularies is forcing changes in IVR system design. Free from the constraints of touch-tone technology or the limitations of only digits speech recognizers, we may now begin to envisage a variety of IVR interaction designs, and a greater variety of telephone applications. Hopefully, the removal of the technology-imposed design constraints will also result in applications that appear more natural to the users. To achieve this goal the large-scale use of speech recognition in telephony applications will require considerable application tuning [1]: that is, ASR technology must be evaluated under application-specific conditions in order to reliably estimate its performance in the field.

2. EVALUATION

2.1. Performance evaluation

There are many speech recognition systems available on the market place, with different characteristics. As a consequence of this growth of the market, there exists a growing need to define standards and test procedures to evaluate ASR systems [2]. In order to choose among different ASR systems, it is crucial to know their performance. A performance measure, which is usually given for the evaluation of automatic speech recognition systems, is in terms of recognition accuracy, i.e. the percentage of utterances that are accurately recognized. Generally the average recognition rate given by the manufacturers is around 98-99%, without ever specifying the conditions under which testing was carried out. Anyway, high recognition accuracy is necessary but not sufficient. In fact, it doesn't say anything about other factors that are more important in contributing to overall system performance, such as vocabulary size, speech type, speakers dependency, level and type of environmental noise.

2.2. Usability evaluation

Since a good speech recognition system is not enough to do a good service, the performance evaluation of a system (in terms of percentage of utterances that are correctly recognized) doesn't accurately characterize how well the system will work in an application. This is due to the fact that performance evaluation doesn't consider the inherent complexity of the application and its impact on the user.

As performance evaluation has been crucial for the development of speech technology systems, in the same way usability evaluation will be crucial for the development of services incorporating such systems. Usability evaluation is concerned with the evaluation of the design of the application rather than the ability of a system to perform within that design.
Usability is defined as the effectiveness, efficiency, and satisfaction with which specific users achieve specified goals in particular environments [3]. Measures of effectiveness relate the goals or sub-goals of using the system to the accuracy and completeness with which these goals can be achieved. Measures of efficiency relate the level of effectiveness achieved to the expenditure of resources. The resources may be mental or physical effort, which can be used to give measures of human efficiency, or time, which can be used to give a measure of temporal efficiency, or financial cost, which can be used to give a measure of economic efficiency. From a user's perspective, the time s/he spends carrying out the task, or the effort required to complete the task are the resources s/he uses.

Furthermore, one of the most important features of usability evaluation is user satisfaction. Every test should include a method for contacting people using the system for comment and reactions. Measures of satisfaction describe the perceived usability of the overall system by its users and the acceptability to the system to the people who use it and to other people affected by its use. Measures of satisfaction may relate to specific aspects of the system or may be measures of satisfaction with the overall system. Measures of satisfaction can provide a useful indication of the user's perception of usability, even if it is not possible to obtain measures of effectiveness and efficiency.

3. THE EXPERIMENT

3.1. The prototypes

The application considered in our work is the automation of a service presently in use with an operator who provides information about new products and facilities offered by Telecom Italia, such as call waiting, call forwarding, three-way calling, and answer call.

We choose this application because it is a typical service generally provided by a DTMF system, and we were interested in making a comparison between touch tone and speech recognition modalities. Furthermore, given the nature of the application, which is extremely simple with simple recognition needs, users would call relatively infrequently and would have little opportunity to learn about the application. What we were interested in knowing was whether the recognition application could provide simple information under these conditions.

We used a PC-based IVR system with a Natural Microsystems telephone board and with a speaker independent isolated word and connected speech recognizer developed by Voice Processing Corp. The application design environment and the technology integration are by Interactive Media Srl.

Three different prototypes of the same application have been developed: they all used speech recognition technology (isolated word recognition and continuous digit recognition), with a menu dialogue that explicitly requires users to respond within the given constraints. VOICE_1 allows users to use only spoken digits; VOICE_2 and VOICE_3 allow users to say command words instead of digits. Furthermore, a DTMF prototype that allows users to enter commands only through the telephone keyboard have been developed, with the purpose of making a comparison between touch tone and speech recognition technologies for telephony services. The ASR prototypes differ among them also in the design of the interaction: VOICE_3 uses prompts with the "audible quoting" technique, in order to better communicate to users which command are allowed, as well as a more sophisticated recovery technique. Audible quoting presents two voices for a single prompt, with one voice presenting the carrier phrase and the second voice stating the possible vocabulary choices [4].

3.2. Experimental Design

First of all, we were interested in evaluating some features of current state-of-the-art speech recognition technology available in the marketplace, i.e., speaker independent vocal access through numbers and words, when it is used in a real-life application.

Second, we wanted to make a comparison between ASR and touch-tone. In fact, there are few available data on when people actually prefer to interact with services via speech recognition rather then touch-tone, especially for Italian services (in Italy touch-tone services are less widespread than in USA) [5], [6].

140 students who were paid for their participation took part to the experiment as users of the service. Each prototype was evaluated with 30 different subjects. Furthermore a Comparison Group of other 20 subjects tested all prototypes, in order to make a comparison among the different systems characteristics.

Users were asked to perform a set of four written scenarios. Two of them required the subjects to activate a service, while the other two required to listen to some information about a service to be able to answer some questions about it.

During each call, the application created log files that captured time-stamped records of every speech and nonspeech related event that occurred in the call flow. These log files were automatically converted into call event data for subsequent statistical analysis. In addition, each of the user's utterances was recorded.

At the end of the experimental session, users were given a questionnaire that asked for ratings of their perception of the quality of the overall interaction, their effort in interacting with the system, and their satisfaction with the different features of the system.

The metrics measured through logfiles and utterances recording, were: task completion and correctness, as well as number of calls per task, task duration time, number of turns per task, number of times subjects spoke before the beep, recognition performance.

The questionnaire measured several metrics in different areas, such as overall judgement, expectations, cognitive load, satisfaction, easiness of use, efficiency, learnability.

3.3. Results

3.3.1 The three ASR systems.

For the three prototypes, we obtained the percentage of 95% of completed tasks where 88% of them were corrected. With respect to the task duration time, we found a significant difference (p=.004) between VOICE_1 and VOICE_3: subjects of VOICE_1 completed their tasks in less time (about 2 minutes per task) than subjects of VOICE_3 (about 3 minutes per task). This difference can be due to a greater complexity in the design of VOICE_3. Prompts of VOICE_3 are longer and more numerous than those of the other prototypes; furthermore a more detailed error recovery procedure is used in VOICE_3. This complexity of VOICE_3 is confirmed by a significant difference (p=.000) also found between VOICE_3 and the other two prototypes in the number of turns. Subjects of VOICE_3 made more turns (an average of 12 turns per task) than subjects of VOICE_1 (an average of 7 turns per task) and those of VOICE_2 (an average of 9 turns per task).

A difference has been observed also in number of times subjects spoke before the beep between VOICE_3 and VOICE_1 (p=.015) and VOICE_2 (p=.000). Again, this can be due to higher naturalness of VOICE_3 prompts that caused subjects to exhibit a more natural behaviour.

Anyway, only few subjects spoke before the beep, suggesting that they were not disturbed by the system's request of speaking after the beep. This is confirmed by subjective data that showed that 72% of subjects of the three prototypes found natural to speak after the beep and only 11% of them found it unnatural.

The analysis of the subjective data showed a high level of satisfaction with the three ASR prototypes: 81% of subjects of the three prototypes scored the interaction with the service as very positive; 89% of them found the service easy to use. With respect to pleasantness, subjects of VOICE_2 and VOICE_3 judged the interaction with the system agreeable more than subjects of VOICE_1. 86% of subjects of the three prototypes judged the system's prompts to be very clear.

With respect to performance evaluation, we observed a significant difference (p=.0005) in the isolated word recognition (for command recognition), between VOICE_1 and the other two prototypes. VOICE_1 obtained 96.3% of correct recognitions, while VOICE_2 and VOICE_3 obtained respectively 91.4% and 91.6%. This difference could have been easily predicted: in fact VOICE_2 and VOICE_3 used vocabularies of specific application words realized at our lab, while VOICE_1 used a digit vocabulary provided by the manufacturer. We observed a less great discrepancy between recognition performance (in the laboratory test) and application performance (in the usability evaluation) obtained by VOICE_2 and VOICE_3. In fact, in the laboratory test the recognition performance was about 95%, i.e. about 4 points better than that obtained in the usability evaluation.

The most interesting result is that observed in the performance of the continuous digit recognition, used to insert the telephone number in the activation procedure. Although the three prototypes used the same vocabulary of digits, error rate gradually increases from VOICE_1 to VOICE_3, and between these two systems the difference is statistically significant (p=.025). VOICE_1 obtained 98.3% of correct recognitions, while VOICE_2 and VOICE_3 obtained respectively 94.8% and 93.1% of correct recognitions.

This result can be due to the increased naturalness in VOICE_2 and VOICE_3 that could have driven subjects to say the telephone number in a more natural way, therefore less appropriate for the automatic recognition. This difference couldn't have been foreseen because the three prototypes used the same vocabulary of digits (provided by the manufacturer).

These objective data on the performance are confirmed by the answers to the questionnaire. In fact there is a significant difference between VOICE_1 and VOICE_2 (p=.007) and VOICE_3 (p=.012) with respect to the perception of system's errors. Subjects answers are related to the objective system's performance: only 10% of subjects of VOICE_1 affirmed that the system made some errors, against 35% of subjects of VOICE_2 and 34% of subjects of VOICE_3.

The same occurred for the errors that subjects perceived to have made: only 28% of subjects of VOICE_1 against 32% of VOICE_2 and 47% of VOICE_3 affirmed to have made some errors during their interaction with the service.

One of the differences between VOICE_3 and the other two prototypes lied in the recovery procedure, that in VOICE_3 was better organized. The recovery procedure was able to solve most of the problematic situations occurring both with the isolated word recognition and with the continuous speech recognition: we recovered 93 situation out of 123. This result is confirmed by the subjective evaluation. In fact, a difference has been observed with respect to the utility of the recovery messages: 92% of subjects of VOICE_3 judged them as quite useful with respect to 76% of subjects of VOICE_1 and 65% of VOICE_2.

3.3.2 Differences between touch-tone and speech input modalities.

In general the DTMF prototype showed a performance very close to that obtained by VOICE_1, in terms of task completion time, number of turns per task.

With respect to DTMF performance, all subjects stated that the system never failed during the interaction.

The more interesting differences between touch-tone and speech input modalities are in task duration time and barge-in.

Subjects that used the DTMF prototype employed less time to perform their tasks with respect to subjects of VOICE_2 and VOICE_3. There is a significant difference in task duration time between the DTMF prototype and VOICE_2 (p=.005) and VOICE_3 (p=.000). We think that this difference is due not to a higher difficulty in interacting with these two ASR prototypes, but rather to the naturalness of the interaction

Subjects who used the DTMF prototype interrupted the system's prompts at least once in each task, while subjects of the ASR prototypes almost never did it. Indeed, only for the DTMF subjects we can affirm that they made barge-in, that is they talked over the system's prompts; while subjects of the ASR systems simply spoke before the beep.

Interesting results have been obtained in the analysis of the questionnaire of the Comparison Group, i.e. the group of subjects who interacted with the DTMF prototype as well as with VOICE_1 and VOICE_3.

VOICE_3 obtained higher scores in being judged as the system which would be better accepted in the market place: 56% of preferences vs. 38% for the DTMF and only 6% for VOICE_1.

Interestingly, 75% subjects judged VOICE_3 as the most enjoyable prototype.

When asked with which of the three prototypes they would like to use for a real service, most of subjects chosen VOICE_3.

4. CONCLUSIONS

With respect to the ASR prototypes, the most interesting finding is that greater naturalness in the interaction makes the subject overcome the poorer recognition performance. In fact, VOICE_3 obtained the highest scores regarding user satisfaction.

The comparison between the DTMF prototype and the ASR prototypes suggest that even if the interaction by touch tone was more accurate and rapid than the spoken interaction, the subjects of our experiment didn't prefer the touch tone modality.

In general we found that a good correspondence between objective and subjective measures. These latter always confirmed and clarified the results of the former.

With respect to methodological issues on usability evaluation, we think it should be said that usability evaluation requires a high number of subjects, who have to be recruited and paid for their participation. Not always all the contacted subjects completed the experimental sessions: for this reason it is better to contact more people than those really needed. A substantial number of subjects (58 out of 160) decided to abandon the experiment or were eliminated because of their poor performance.

Also, the available mathematical-statistical tools do not seem to be entirely appropriate for usability evaluation. The general approach of hypothesis testing is not always the best one for interpreting the results. Appropriate nonparametric tests are not always indicated in the literature. The literature is not always explicit about sample sizing or sample size varies from a few dozens subjects to thousands of subjects without an explicit discussion of the reasons and consequences of using these very different resources.

Furthermore, most vendors do not provide tools for an application's performance evaluation. In our research much effort was spent in designing and developing useful tools for collecting our objective measures.

ACKNOWLEDGMENTS

The authors would like to express their thanks to Raffaele Poppa, Claudio Spallancini, and Roberto Valente of Interactive Media Srl. for their kind and useful contribution.

5. REFERENCES

1. R. Rosinski, E. Roskos, J. Delong, D. Brown, R. Goldberg, R. Sachs, "Speech Recognizer Accuracy and Application Performance", in Proceedings of COST Workshop on Speech Technology in the Public telephone Network: Where are we today?, 1997, pp.11-14.

2. D. Gibbon, R. Moore, R. Winski (Eds.), *Handbook of Standards and Resources for Spoken Language Systems,* Berlin: Mouton de Gruyter, 1997.

3. ETSI Technical Report, "Human Factors (HF), Guide for usability evaluations", ETR 095, 1993.

4. S. Basson, S. Springer, C. Fong, H. Leung, E. Man, M. Olson, J. Pitrelli, R. Singh, S. Wong, "User participation and compliance in speech automated telecommunications applications", in Proceedings of ICSLP96, Philadelphia, 1996, pp. 1680-1683.

5. D. Karis, "Speech Recognition Systems: Performance, Preference, and Design", in Proceedings of HFT'97, 1997, pp. 65-72.

6. D. Albesano, P. Baggia, M. Danieli, R. Gemello, E. Gerbino, C. Rullent, "A Robust System for Human-Machine Dialogue in Telephony-Based Applications", International Journal of Speech Technology, vol. 2, n. 2, 1997, pp.101-111.