

Evaluation of a multimodal Virtual Personal Assistant

Glória Branco, Luís Almeida, Nuno Beires, Rui Gomes

Voice Services and Platforms - Portugal Telecom Inovação, Porto, Portugal

[\[gloria, lalmeida, nbeires, shortcut-r-gomes\]@ptinovacao.pt](mailto:[gloria, lalmeida, nbeires, shortcut-r-gomes]@ptinovacao.pt)

Abstract

Nowadays mobility is a reality. The small size and the rising computational power of personal mobile devices enable the access and the exchange of an increasingly greater volume of data and information, anywhere and anytime. Multimodal interfaces, combining voice and visual modes can simplify the interaction but they raise new challenges concerning the usability and acceptability of such interfaces.

In this paper, we present the results of the evaluation of a Virtual Personal Assistant application that provides e-mail management facilities through an intelligent and multimodal interaction with a fixed or a mobile device. The results show that the quality and speed of the system feedback as well as the recognition accuracy of the spoken components are key factors to a better user experience and for the acceptance of multimodal systems.

Key words: Voice-enabled Interfaces, Multimodal Interaction, Usability

1. Introduction

Nowadays mobility is a reality. The small size and the rising computational power of personal mobile devices enable the access and the exchange of an increasingly greater volume of data and information, anywhere and anytime. Applications and services that allow the access and management of this information are required. Multimodal interfaces, combining voice-enabled mode - to override the constraint imposed by the small size of screens - and visual mode can simplify the interaction but they raise new challenges concerning the usability and acceptability of such interfaces. In the last years, several applications have been developed and reported, such as [Almeida *et al* (2002), Hemsén (2003), Pieraccini *et al* (2002)] among others.

If usability is one essential component in guiding the design and service development, the acceptability of such interfaces also needs to be studied. The user acceptance depends on different factors, for instance, their needs, expectancies and privacy, the interaction context and even in social and cultural values. In many aspects, see for instance [Dybkjær *et al* (2005), Pieraccini *et al* (2005)] the usability and acceptability evaluation of multimodal conversational systems is an open research issue.

In the framework of the European Project FASiL (Flexible and Adaptive Spoken Language and Multimodal Interfaces), a Virtual Personal Assistant (VPA) was developed and evaluated. The VPA was a multilingual (English, Swedish and Portuguese), multimodal and multi-device application that provided e-mail management facilities through a conversational interface.

The main goal of the evaluation was to assess the usability and the accessibility of VPA. Representative users were asked to complete a set of typical e-mail tasks and measures were

taken of effectiveness, efficiency and satisfaction. Objective measures of system performance were taken too. The evaluation results can be found in [Alexander *et al* (2004)].

PT Inovação (PTIN) carried out the evaluation of the Portuguese VPA. Since the focus of our activity is the multimodal dialog systems and services, we were particularly interested in the understanding of which factors affect the user experience and the acceptance of multimodal services. Besides the analysis of the overall results we want a deeper insight into two relevant aspects broadly referred in the literature: the concept/recognition accuracy and the use (how and when) of multimodality [James *et al* (2000), Oviatt *et al* (1997), Oviatt (1999)].

This paper reports on the preliminary results of the users trials of the Portuguese VPA. Section 2 briefly describes the system; section 3 describes the evaluation method and, finally, sections 4 and 5 present and discuss the results.

2. The System

The core system was developed by one of the consortium partners, Vox Generation. The VOX platform included the natural language understanding component and the dialog manager component. The platform interacts with the ASR and TTS sub-systems provided by Nuance, formerly ScanSoft. The University of Sheffield developed the categorizer and the summarizer modules, included in the VPA prototype.

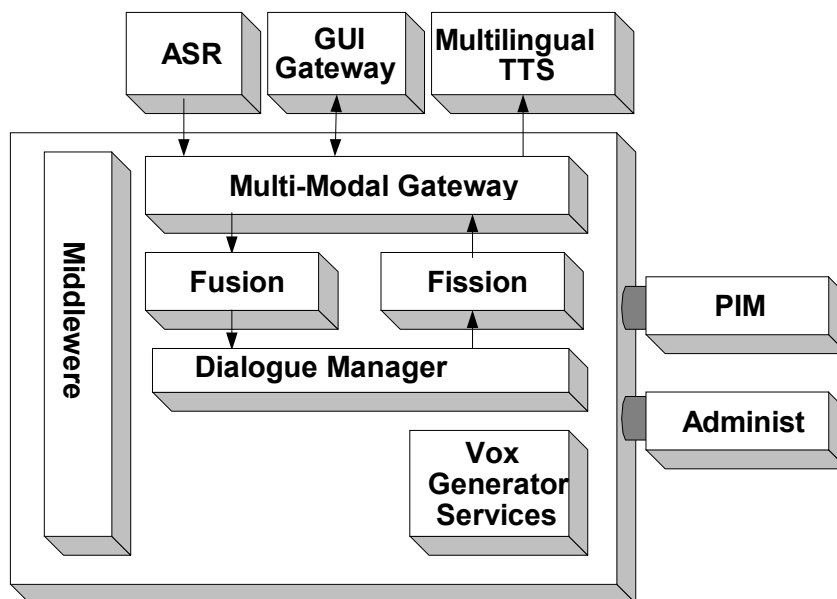


Figure 1: The system architecture

The application could be accessed through various devices such as a mobile or a fixed phone, a PDA or a personal computer. It was possible to interact with the system in a variety of ways: using speech in and out via a telephone (voice user interface); using keyboard, mouse or touch-screen for input and video display out (graphics user interface), using a combination of these such as speech in and display out (multimodal interface).

The VPA prototype covered the basic e-mail functionality. Using natural language the users could search for e-mail by sender, date or category, they could send, reply to or

forward e-mails to multiple recipients, including a written or recorded message and priority level. They could also ask for an e-mail summary.

The dialog strategy followed a mixed-initiative approach and barge-in was supported. The prompts and hints tried guiding the interaction in the desired direction, providing the most common options and using examples in error situations.

U	Search for new e-mails from António Silva.
S	You have 2 messages from António Silva. E-mail 1: from António Silva, subject: documentation; received ...
U	<i>(barge-in)</i> Next
S	E-mail 2: from António Silva with high priority, subject: meeting tomorrow, received ...
U	<i>(barge-in)</i> play me the e-mail
S	The message says: <i>Please confirm</i> ...
U	I want respond to this message.
S	Reply. Say no if wrong.
	<i>(message recording)</i>
S	Do you want to send the e-mail? You can also add recipients, replay the message, ...
U	<i>(barge-in)</i> Send e-mail with copy to Maria Costa ... and blind copy to João Antunes.

Figure 2: An example of dialog (translated from Portuguese)

3. Method

The evaluation design as well as the protocol and the support materials were developed by the Royal National Institute of the Blind (RNIB) who co-ordinated the user evaluations. For the Portuguese evaluation, we translate and adapt the original material.

The evaluation was carried out in a laboratory environment, with 12 native Portuguese speakers interacting with VPA to carry out a set of typical e-mail tasks. The participants received a document with the introductory instructions, the tasks to accomplish and the questionnaires. At the start of the experiment, subjects were given a short explanation of the VPA system and the test objective. They fill in a pre-questionnaire with personal information and they were asked to sign a consent form to record the experiment. Two evaluators observed the participants to assist them and to note the issues found and the user's comments. After each task, users were required to express their opinion through a 5-point Likert scale questionnaire. A post-test questionnaire was designed also to find out the user's overall impression about the system. The questionnaire had a nine 5-point Likert scale statements and a few open questions, to provide feedback on the good and bad aspects of VPA and their experience. It was also asked about the future use of VPA.

3.1 Participants

A total of 12 native speakers without accent were selected from PTIN staff to perform the tests; 8 were males and 4 females, aged from 19 to 46 (mean 30,6 years, std dev 6,4). About 75% of the participants had an academic education, 16,7 % mid-level education and basic-level education the remaining. Three of them had experience with spoken-driven applications and 2 had little experience. The others don't have experience with this kind of services. None of the participants was familiar with the system. The participants were representative of the

target VPA users since all them were professionals from the ICT domain, with a high mobility level and they were experienced computer and e-mail users.

3.2 Experimental setup

A mailbox was set up specifically for the evaluations and it was populated with e-mails related with the tasks to carry out. A set of 5 tasks was designed to cover the functionalities of the system: login and browsing mailbox, search for and reply to an e-mail; search and forward; administer and manage the recipient list and finding and deleting an e-mail. The complexity increased from task 1 to task 5: from browsing the mailbox till a more complex scenario where the user needs to search for an e-mail by sender and to execute the instructions from that e-mail. The tasks were described as open scenarios to allow the users to take the initiative and to experiment the system.

For the users trials the graphic interface was a web-based page, simulating a mobile phone. The users used a desktop PC with Internet access to interact with the GUI and a fixed phone to convey voice to the system.

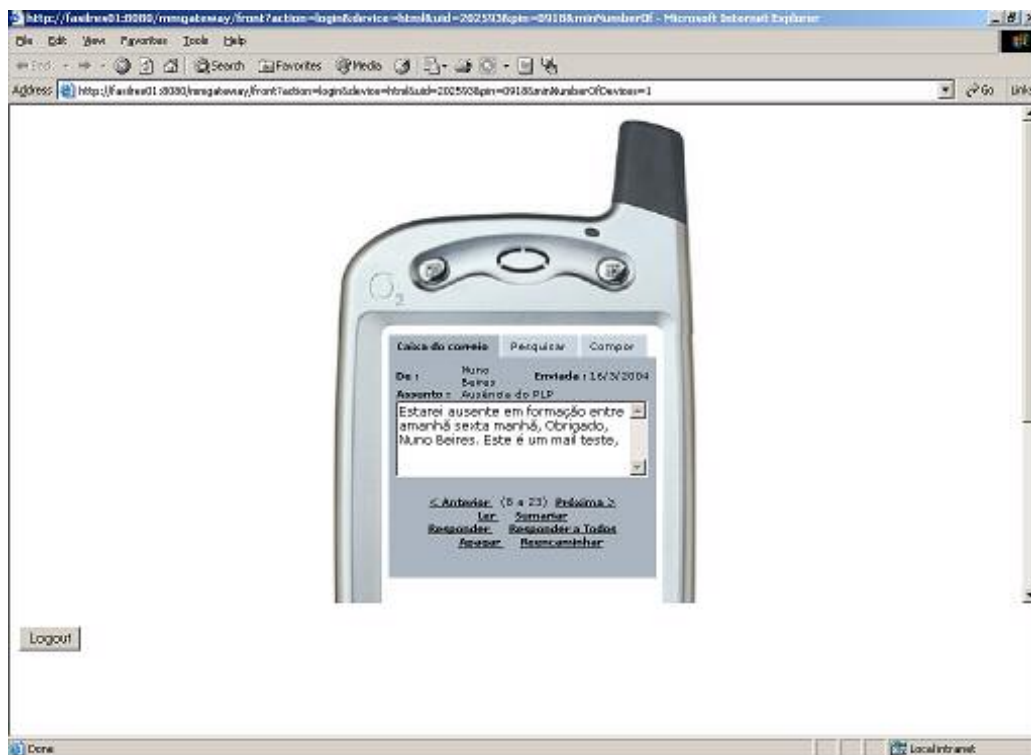


Figure 3: The graphical interface

“InoVox”, an IVR platform developed by PTIN, recorded the user utterances in full. In addition to the system log files, a tool, developed by VOX, allowed the analysis of the spoken interaction between the user and the system.

4. Results

Metrics collected per session consisted of objective metrics extracted from the logging and subjective metrics collected via the questionnaires. In addition, the evaluators annotated by hand all the events that occur during the tasks execution. The questionnaires were used to

obtain the user satisfaction by asking the user to specify the degree to which they agreed with the set of statements on a 5-point Likert scale, from ‘Very satisfied’ to ‘Very unsatisfied’.

Even if too few users participated to allow for statistical manipulation of results, we think that there are interesting findings, which could be confirmed in subsequent larger scale studies.

4.1 Task analysis

For each task, several objective measures were taken during experiments. The evaluators annotated the time elapsed to complete the task, the actual task completion and the number of Graphical User Interfaces (GUI) interactions. From the logs of the spoken interaction we also collected the concept accuracy (correct answers to users request), the user utterances without response, the number of misunderstandings and the incorrect answers from the system.

The Table 1 present the values collected by task: the time in minutes to complete the task (median and range), the task completion rate, the number of spoken and graphical interactions (median and range) and the system performance values (in percentage).

Table 1: Task summary

Task	Time (range)	Comp. %	Interactions		Spoken interaction (%)			
			VUI	GUI.	Correct responses	No Responses	Misunderst andings	Incorrect responses
T1	10 (7-18)	83,3	14,0 (9-49)	6,5 (0-22)	54,3	13,7	6,8	25,2
T2	7 (5-12)	75	11,5 (5-33)	4,5 (0-16)	60,2	11,4	8,5	19,9
T3	5 (3-16)	100	14,5 (1-26)	1,5 (0-20)	67	9,8	9,8	13,3
T4	6 (2-10)	50	16,5 (4-41)	2 (0-6)	55,84	7,4	18,6	18,2
T5	10 (4-18)	75	32 (9-66)	4 (0-11)	59,4	9,4	9,4	21,8

Task one was an exploratory task to allow the user to become familiarized with the system and with the different interaction modes. In the remaining tasks, the user goal was to accomplish the task objective. The majority of the incomplete tasks were result of system stability problems.

4.2 Questionnaire Overall Results

The post-test questionnaire had a nine 5-point Likert scale statements and a few open questions, to provide feedback on the good and bad aspects of VPA and their experience. It was also asked about the future use of VPA.

The users were asked to express their opinion from “very satisfied” to “very unsatisfied” with a neutral point. The questionnaire questions asked for:

- how intuitive was the interface,
- how easy was work with the system,
- how confident users were using the system,
- how satisfied they were in general with the system and the experience,
- how satisfied they were with the interaction control (where they were and what to do),

- how satisfied were with the interaction quality, particularly with the recognition errors,
- how appreciate the TTS voice for prompts,
- how appreciate the TTS voice for e-mail reading,
- how satisfied were with the dialog with the system.

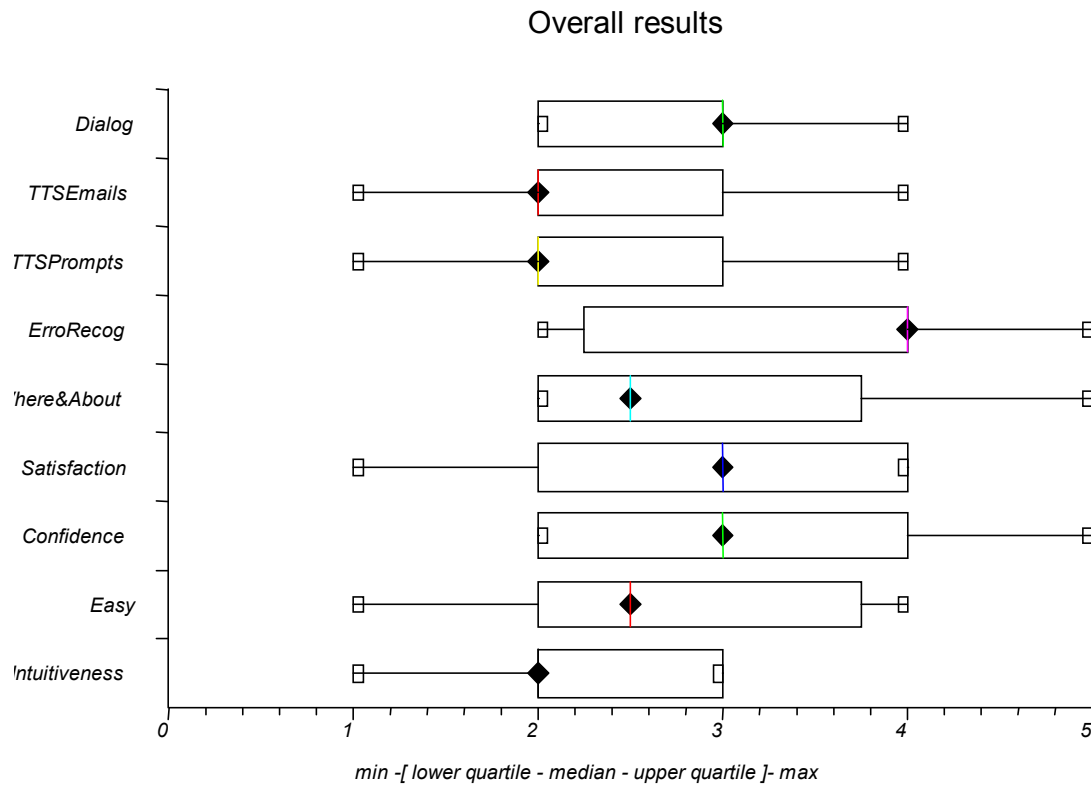


Figure 4: Post-test questionnaire: median and inter-quartile range.

The results of the overall appreciation (for the individual components of the post-test questionnaire) are shown in the box plot in Figure 4. The central box represents the distance between the first and third quartiles with the median between them marked with a diamond; the minimum and the maximum values are marked in the lines extremes (1 stands for “very satisfied” and 5 for “very unsatisfied”). The VPA was seen as easy to use and intuitive (only one user used the system Help). The interaction control (where&about) was appreciated too. The overall satisfaction had a neutral appreciation (median=3; mode=4). The question regarding the interaction quality (related with error recognition) was severely ranked by the users: median= 4 (unsatisfied). The quality of dialog and the confidence had a neutral appreciation, as well as the TTS quality.

The study of the relationships between the overall satisfaction and the remaining statements using Spearman’s correlation coefficient indicate a significant correlation with easy of use ($\rho=0,74$); confidence ($\rho=0,79$), dialog ($\rho=0,87$); interaction control ($\rho=0,73$) and error recognition ($\rho=0,69$). The values for intuitiveness and the TTS quality of prompts and e-mails reading were not significant.

The results of Mann-Whitney test did not show evidence for differences between females and males. Regarding the experimented or naïve users, the Kruskal-Wallis test results lack of statistical significance too.

To find the relationship between user overall satisfaction (subjective) and the concept accuracy (objective) we calculate the correlation between them. Not surprising, the result of Spearman correlation was statistically significant: $\rho=0,85$, $p=0,0004$. However, no statistical relationship was found between the concept accuracy values and the interaction quality, the subjective evaluation of error recognition.

5. Discussion

Table 1 shows that the preferred modality was speech. The users followed a similar approach to interact with the system: they used speech input to convey the commands and use the graphical interface to read the messages and to scroll quickly through the contacts list. A more intensive use of the GUI occurred when the system presented problems, in order to overcome the recognition or understanding errors and the slowness of the system response.

The users tried to use natural language, using short phrases but with complex commands. Only one user adopted a very pragmatic style since the beginning and for all the tasks: short command utterances and GUI confirmations/commands to speed up task completion. In contrast, another user used a real natural language approach, with out-of-context utterances, hesitations and so on. As a curiosity, the overall concept accuracy for the first user was 71,4% and for the second was 52,4%.

The majority of the users used and appreciated a mixed initiative dialogue. However, the system initiative several times annoyed the users because it was most often inappropriate or out of context. For instance, the system prompt warning the user that he was trying to send a message without recipients was welcomed; but, when the user composed the message with the recipients and said “send” and the system answer is “who would you like to send it to?”, the user felt frustration. Besides the system stability problems one possible explanation is the over-confirmation strategy implemented.

From the post-test questionnaire results, we can conclude that the overall user satisfaction is influenced by the ease of use of the system, the control that the user had upon the system actions and the quality of the interaction. The use of the TTS Portuguese voice was well accepted by the users.

The lack of correlation between the concept accuracy (objective values) and the interaction quality (subjective evaluation of error recognition), suggests that the user perception of the recognition accuracy is influenced by other factors, for instance the system response time. However, this aspect needs a more detailed study.

The user opinions ranged from positive to very negative but for the majority of the users the overall opinion was positive. The multimodal VPA concept was attractive to all users and was seen as a key enabler supporting the growing user mobile attitude.

When asked about the future use of the system 33% of the users said that they would use it, one user didn't answer and 58% said that they wouldn't use the system. The main reasons for such negative answers were the system's significant latency (slow response time) and the recognition problems.

6. Conclusions

We presented the preliminary results concerning the user evaluation of a Virtual Assistant prototype developed in the context of the FASiL project.

The conversational and multimodal approach to this kind of applications was very well accepted and supported by the users.

We can conclude that the quality and speed of the system feedback as well as the recognition accuracy of the spoken components are key factors to a better user experience and for the acceptance and use of voice-enabled systems. The GUI approach adopted as a mean to overcome the slowness of the system response and the recognition problems suggest that multimodal interfaces can overcome the weaknesses of each modality and exploit the full strengths of combined modes.

However, improvements are needed to increase the recognition accuracy of the spoken components, providing a better user experience and open the door to the acceptance and divulgation of multimodal systems.

7. Acknowledgements

We would like to thank all the participants, who kindly participated in the User trial. We would like to thank also to all FASiL partners: CapGemini (Sweden), MIT Media Lab Europe (Ireland), ScanSoft, Inc (USA), The Royal National Institute for the Blind (UK), The Royal National Institute for Deaf People (UK), University of Sheffield (UK), Vox Generation Ltd (UK), for their contributions.

This research was supported by the EU Grant FASiL IST 2001-38685.

References

- [1] Alexander, T., Dixon, E. Usability Evaluation Report for VPA 2 (Public Version). FASiL deliverable D.3.3.3, 2004.
- [2] Almeida, L., Amdal, I., Beires, N., Boualem, M., Boves, L., den Os, E., Filoche, P., Gomes, R., Knudsen, J.E., Kvale, K., Rugelbak, J., Tallec, C., Warakagoda, N. Implementing and evaluating a multimodal and multilingual tourist guide. Proc. CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems (2002).
- [3] Dybkjær, L., Bernsen, N.O., Dybkjær, H. Usability Evaluation Issues in Commercial and Research Systems. COST Workshop (2005).
- [4] Hensen, H. Designing a multimodal dialogue system for mobile phones. Proc. of the 1st Nordic Symposium on Multimodal Communication (2003).
- [5] James, F., Rayner, M., Hockey, B.A. Accuracy, Coverage, and Speed: What Do They Mean to Users? CHI Workshop on Natural Language Interfaces (2000).
- [6] Oviatt, S, De Angeli, Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. Proc. CHI 1997.
- [7] Oviatt, S. Ten myths of multimodal interaction. Communications of the ACM (1999), 75-81.
- [8] Pieraccini, R., Carpenter, B., Woudenberg, E., Caskey, S., Springer, S., Bloom, J., Phillips, M. Multimodal spoken dialog with wireless devices. Proc. of the ISCA Tutorial and Research Workshop (2002).
- [9] Pieraccini, R., Huerta, J. M. Where do we go from here? Research and Commercial Spoken Dialog Systems. Proc. of 6th SIGdialWorkshop on Discourse and Dialogue (2005).