THE IMPACT OF THE SONIFICATION OF A VOCAL SERVER ON ITS USABILITY AND ITS USER-FRIENDLINESS

Valérie Maffiolo *Noël Chateau* *Marc Mersiol*

Human Interactions Direction
France Telecom R&D, 2, avenue Pierre-Marzin, 22307 Lannion Cedex, France
valerie.maffiolo @ francetelecom.com noel.chateau @ francetelecom.com marc.mersiol@ francetelecom.com

ABSTRACT

This paper deals with the evaluation of the impact of the addition of eight sounds in a vocal server on its usability and user-friendliness. Thirty-two subjects tested two versions of a voicemail service through eight scenarios, with and without sounds. Three types of data were collected: behavioral data, declarative data (answers to questionnaires), and galvanic skin responses. The main results show a learning effect from one version to another that is not dependent on the versions tested. No effect of the sonification on the usability and on the perceived user-friendliness of this voicemail server was found.

1. INTRODUCTION

Thanks to the development of software and hardware dedicated to sound generation and control, the study of the sonification of man-machine interfaces has considerably grown this last decade. One of the main fields of investigation has focused on the question of the improvement of the usability of man-machine interfaces by the addition of earcons [1]. Earcons can be defined as "abstract, synthetic tones that can be used in structured combinations to create sound messages to represent parts of an interface" [2]. Brewster and his colleagues have conducted a considerable number of studies in which they have shown the benefits, in terms of efficiency, of the introduction of earcons, mainly in graphic interfaces [3].

Concerning vocal interfaces, Brewster [4, 5] studied the memorization of a tree hierarchy as those that can be found in vocal servers by associating earcons to the nodes of the hierarchy. To each node of the hierarchy corresponds an earcon whose properties are inherited by earcons of higher levels. In these studies, Brewster showed that after a little practice, the memorization performances of tree hierarchies are significantly improved by the addition of earcons. Wolf *et al.* [6] have stressed that the major problem in the usability of vocal servers is that users often get lost in the tree hierarchy. Therefore, if earcons can help them to better memorize the tree hierarchy of a vocal server and if earcons give users useful feedback information (which was demonstrated in graphic interfaces), it might be hypothesized that earcons could help users in an interactive context when they have to navigate through this hierarchy.

Another aspect of earcons that should not be neglected is that as musical sounds, earcons potentially can convey an emotional content that can affect users' activity and their impressions about the user-friendliness of the sonified interface [7, 8].

The studies reviewed here are very promising for telecommunication operators since they suggest that the introduction of earcons in vocal servers might be an efficient way of improving the vocal servers' usability and user-friendliness. These studies allow us to formulate the four following hypotheses:

H1: As earcons can help users to memorize the tree representation of a vocal server, a sonified vocal server should improve the system's efficiency by reducing the number of elementary actions the user has to do to achieve a specific task.

H2: As earcons can provide relevant feedback information on the system's state, a sonified vocal server should improve the system's efficiency by reducing reaction times, the number of errors, and more generally the time required to achieve a specific task.

H3: A sonified vocal server should improve users' expertise by helping them to better navigate in the system (thanks to an improved memorization of the hierarchy) and to better identify the system's states (thanks to audio feedback information).

H4: The presence of earcons in the interface of a vocal server might improve its user-friendliness by conveying some emotional content in the interface.

This paper reports on a study that tested these four hypotheses by introducing eight musical sounds in a DTMF (Dual Tone Modulation Frequency) vocal server (a voicemail) named *Avantys*. Two versions of this server (one sonified, the other not) were submitted to thirty-two naïve users for testing. Behavioral and declarative data, as well as galvanic skin responses were collected in order to obtain the most comprehensive view of the activity, reactions, and feelings of the users.

2. METHODOLOGY

2.1. The vocal server Avantys

When a user connects to Avantys, after a welcoming prompt, he reaches the Main Menu where two choices are proposed: the user can either select the Consultation Menu to hear new or old (stored) voice messages and deal with them, or select the Personalization Menu to configure the service (recording and selection of announcements, secret code management, notification management, *i.e.* management of time slots during which the system can call the user to tell him he has received a new message). Sub-menus exist in each of these three personalization menus but they won't be dealt with here because of lack of space. As a whole, Avantys can be seen as twenty-two boxes: one for the Main Menu, five in the Consultation Menu, fifteen in the Personalization Menu (five in the Sub-Menu Announcements, three in the Sub-Menu Secret Code and seven in the Sub-Menu Notification) and one for the Help Menu.

2.2. Sonification of Avantys

The first hypothesis we wanted to test is in keeping with the mapping of musical sounds to the tree hierarchy of the vocal server (equivalent to a spatial mapping, as a map could describe the tree hierarchy). The second hypothesis is in keeping with the association of musical sounds to feedback that can inform users about the system's state (equivalent to a temporal mapping, as the system's responses to users' actions could be described by a temporal graph). In order to test jointly these two hypotheses, both "Navigation" musical sounds and "Feedback" musical sounds had to be created.

The Avantys' description shows that the server does not have a very wide tree hierarchy. From the users' perspective, it could best be described by the three main menus: the Consultation Menu, the Personalization Menu, and the Help Menu. Other sub-menus such as the Secret Code and Notification Management might appear less important to users. For this reason, we decided not to construct a set of earcons that would be mapped on the whole tree hierarchy. Moreover, we didn't want to introduce too many sounds in the server since in a previous experience with the introduction of musical sounds in a commercial vocal server, we obtained very negative feedback from users who complained about the presence of too many sounds.

Consequently, four "navigation" musical sounds were constructed with midi instruments (piano and flute) and were associated to 1) the Main Menu (after the Welcoming Prompt), 2) the Consultation Menu, 3) the Personalization Menu, and 4) the Help Menu.

Concerning "Feedback" musical sounds, four sounds were also created for the following feedbacks:

1) background music during the Prompt, inviting users to record a message (the music stops with a beep telling the user to deliver his message),
2) error feedback (user action is not authorized by the system),
3) inactivity feedback (the user does not react while the system is waiting for his action),
4) connection timeout feedback (too many errors have been made or the connection time is over so the system will automatically hang up).

2.3. Test procedure

2.3.1. Creation of the two versions

Two versions of Avantys were tested: one with musical sounds and the other without. To avoid the development of two versions of the service, the sonified version consisted of the original version (without sounds) with sounds added online and live during the tests by an experimenter according to the Wizard of Oz technique.

2.3.2. Experimental design and scenarios

In order to test H1, H2, and H3, both versions of the service had to be compared directly by users, but also had to be tested twice. Consequently, four groups of eight subjects were created as described in table 1 below. The first group tested the sonified version twice and the fourth group tested the original version twice. The second group tested the original version in a first session followed by the sonified version in a second session. The third group tested the sonified version followed by the testing of the original version. Together, the users formed a general-public panel comprised of eighteen women and fourteen men, covering all socio-professional categories, with ages ranging from 17 to 53.

		Session 1	
		Sonified	*Original*
Session 2	*Sonified*	Group SS Sonified-Sonified	Group OS Original-Sonified
	Original	Group SO Sonified-Original	Group OO Original-Original

Table 1. *Experimental design for the four groups*

For each session, users had to realize four different scenarios (1: listen to and store a new message, 2: create a personalized announcement and activate it, 3: modify the time slot for the notification and activate it, and 4: reactivate the system's default standard announcement and delete a message). Prior to these four scenarios, for the first session, each subject ran a training scenario without any specific task in order to become familiar with the service. For this training scenario, subjects were instructed to hang up after 5 minutes of interaction.

2.3.3. Collected data

Three types of data were collected: behavioral data, declarative data, and galvanic skin responses. The analysis of the subjects' activity permits the evaluation of the subjects' performance through their success at accomplishing the task, the time required to perform the task, and the basic actions done to perform it (*e.g.* the number of keys pressed). After each session of the test, questionnaires collect subjective data on the subjects' feelings concerning the usability and style of the vocal server and, in the case of the sonified version, the memorization of sounds and the subject's associated feelings. The third category of data collected is the galvanic skin response of the subjects giving information on their level of stress during the use of the server. The galvanic skin responses are measured with two electrodes put on two fingers (the index finger and the little finger). In order not to modify the position of the electrodes, the subjects are asked to not manipulate the phone handset with their hands, but to use the function "Hand-Free" to call the server.

2.3.4. Material

Two rooms are used for this experience. The test room is a soundproof room with no windows. The user sits at a table and uses a phone handset. Two cameras record the test (Figure 1).

Figure 1. *Plan of the test room*

Two experimenters are present in a control room, located behind the test room. There are two video playbacks, one showing the keyboard of the phone used by the subject, and the other one showing a front view of the subject. These video monitors make it possible to see which keys the subject uses, and to control how he or she "feels" the test. Eventually, they also permit to break in on the test if the subject has major difficulties with the task (for example, if the subject doesn't know what to do after having made a mistake).

Three computers (P.C.) are used in the control room. The first is used to code the subject's activity with software developed in FTR&D known to facilitate the capture of the subjects' actions. The second is used to record in continuum the physiological measures with the software "Biograph" [9], while the third generates the sounds that are sent to the subject through the phone handset with the Wizard of Oz technique. The eight sounds are predefined and each is affected to a key of the computer keyboard with the software "Mixman StudioPro" [10]. The sounds are activated by using the keyboard and sent back through the microphone of the phone in the control room to the phone handset of the subject via the phone network (principle of double listening).

3. RESULTS

3.1. Success of scenarios

Table 2 gives the percentages of success per group and per session ; all scenarios failed. It can be observed that for one given session, G_OS excepted, there is almost no difference between groups. On the other hand, the performances of each group, again G_OS excepted, improve from one session to another. Concerning G_OS, the results of this group are surprisingly high and might be attributed to a specific distribution of the panel (more students who showed less difficulties in using the vocal server). From these first results it can be concluded that the sonified version did not lead to a higher success rate and did not improve the users' expertise.

	Session 1	Session 2
G_SS	43.75	59.38
G_OS	71.88	71.88
G_OO	40.63	56.25
G_SO	46.88	59.38

Table 2. *Percentage of success for the four groups for the four scenarios*

3.2. Efficiency

Figure 2 gives the time required by those subjects who succeeded in the different tasks of the scenarios for the four groups (abscissa), and for the two sessions (parameter). It can be observed that the session had an important effect on the dependent variable (ANOVA: $F(1,55) = 24.28$, $p<0.001$) whereas the group's effect was small (ANOVA: $F(3,55) = 3.09$, $p<0.05$). However, this reaction is not the expression of an effect depending upon the version of the service. An HSD Tuckey test conducted on all pairs of conditions showed that the only significant difference between groups was between G_SS and G_SO for the first session, that is, between two groups which had the same version to test. Consequently, no effect of the version on the time needed to achieve the task can be observed.

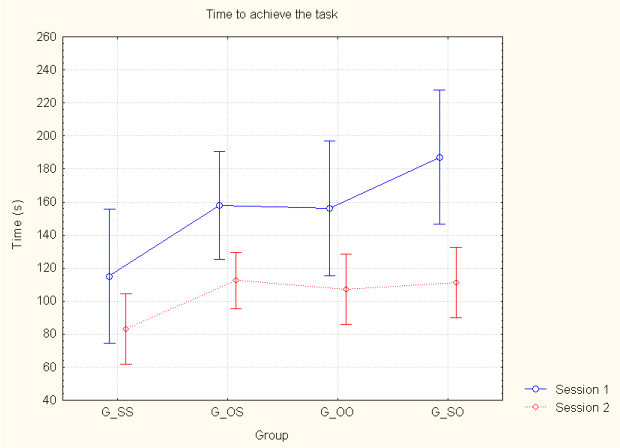

Figure 2. *Time required to achieve the task for the four groups and for the two sessions*

An elementary action done to perform the test is pressing a certain number of keys (Figure 3).

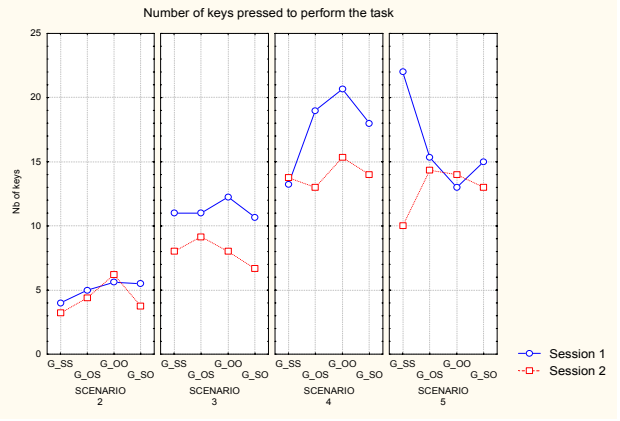

Figure 3. *Number of keys pressed to perform the task for each scenario, the four groups, and the two sessions*

An analysis of the results of the two sessions shows that subjects pressed fewer keys during Session 2 than during Session 1; this confirms the development of the subjects' expertise. This result is not very important for Scenario 2 during which the subjects seem to have optimized the number of pressed keys since the first session. The fact that the curves corresponding to sessions 1 and 2 intercross seems to show that the expertise needs to go further (Scenarios 2, 4, and 5).

An analysis of the results of the different scenarios shows that the number of keys pressed by subjects increases from Scenario 2 to Scenario 5; this indicates the increasing difficulty and duration of the scenarios.

An analysis of the results between groups shows no difference. No effect of the sonification of Avantys on the number of keys pressed to perform the task can be demonstrated.

3.3. Consultation of the Help Menu

The subjects use the Help Menu less in the second session than in the first, because they have fewer difficulties during the second utilization of the vocal server. This finding confirms a learning effect from one version to another, which does not depend upon the sonification of the server.

The subjects' perplexity during the tasks is demonstrated by the number of times they ask for help. When the subjects have understood well what they have to do and know how to do it, they don't consult the Help Menu (see for instance on Figure 4, Scenarios 3 and 4 in Session 2). On the contrary, for Scenario 5, the number of times the Help Menu is consulted is high whatever the session; this brings into focus the subjects' difficulties in performing the required task.

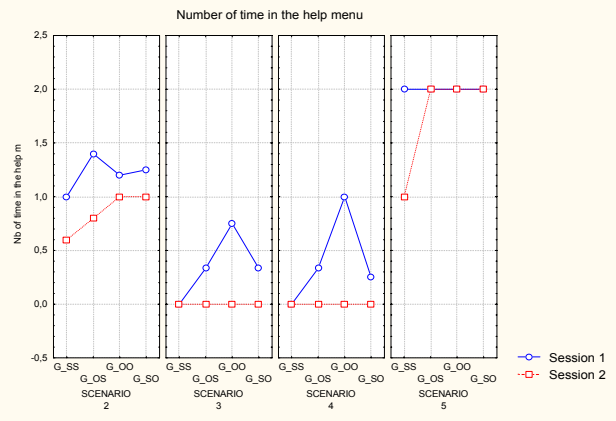

Figure 4. *Number of times subjects performing the tasks use the Help Menu*

For all the groups of subjects, the time spent in the Help Menu is 50% longer in Session 1 than in Session 2. The periods in the Help Menu seem to become progressively longer between Scenario 2 and Scenario 5 (Figure 5).

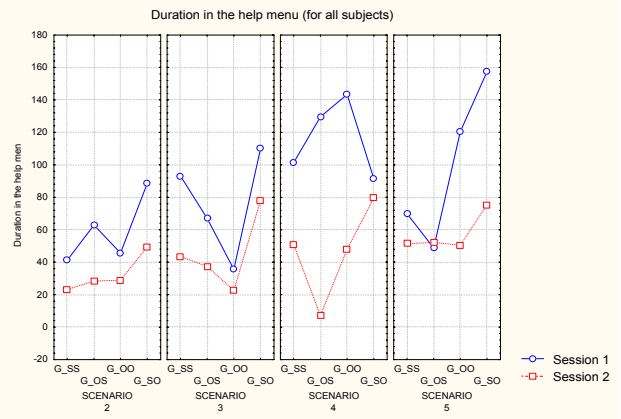

Figure 5. *Time spent in the Help Menu for all subjects*

This tendency, particularly clear during the first session, tends to fade during the second session. The time spent in the Help Menu during the first sessions varies from one group to another and from one scenario to another. During the second sessions, those durations became more homogeneous between groups.

3.4. Subjects' Impressions

For the first session, 50% of the subjects who tested the sonified version (G_SS & G_SO) find that it is rather hard to situate themselves in the server. This percentage reaches 71% for subjects who tested the original version (G_OS & G_OO). On the contrary, surprisingly, 63% of the subjects of G_SS & G_SO feel that they were rather well guided in the server, whereas 77% of the subjects of G_OS & G_OO have the same feeling. The percentage of subjects who find the service rather user-friendly is equivalent for both (G_SS & G_SO) groups and (G_OS & G_OO) groups: 69% *vs.* 65%.

The subjects of G_SS and G_SO found it hard to give the right number of sounds used by the server. This is probably due to the fact that they concentrated mainly on the verbal message in order to use the service in the best way. They often identify two or three sounds. No subject gives the right number of sounds, perhaps because they actually have not heard all the sounds while using the server. The exact positions of sounds in the server are hard to identify; the places best identified are those at the beginning of the service, during access to the different menus, at the beginning of a message recording, and for the validation of an action. Concerning the usefulness of the sounds, 50% of the subjects of G_SS & G_SO think that sounds are useful, while 62% think they are pleasant. For the subjects, sounds can be used to indicate the moment they start using a functionality and/or for having an auditory confirmation of the functions opened, for distinguishing between the different menus, or for being informed of their position in the architecture of the service. A few sounds (such as the welcoming music) are felt to be suitable for improving conviviality. In fact, the majority of subjects are skeptical as to the possibility of a sound helping navigation in a service.

For the second session, results are rather similar. The interesting thing is that for all groups, more than 65% prefer the second version to the first one. This means that subjects' judgments were influenced more by training than by the differences between the versions (sonified *vs.* original). They prefer the second version since they have fewer difficulties in using it. The subjects of group G_SS hear more sounds because they are more attentive. For the subjects of groups G_SS and G_OS, sounds do not help navigation in the service (69%). A few subjects consider them as stressful and disturbing because they "disturb comprehension" (56%). In general, it seems that "there are moments when the music fits well, and others when it is surprising and annoying".

Globally, the users are not very enthusiastic about Avantys. Its usability is not very good even if the subjects are conscious that they will be able to use it once they have had a little practice. For this server, the sounds do not seem to help users, but the subjects consider the sonification idea as a good one. The sonification as realized in this study, even if it improves conviviality, does not improve the usability of the server. A different sonification using shorter or environmental sounds is proposed by a few subjects.

3.5. Skin Conductance

Four indicators of the variation of the emotional activity of the subjects during the test are defined through four variables issued from the primary signal of each subject:

- the relative mean is the temporal mean within the value of the temporal mean obtained for the first scenario;
- the temporal variance makes it possible to evaluate the variation of the electric conductance during the test;
- the percentage of time passed over the starting value, this value being calculated as the mean of data on the first five seconds of the test;
- and the number of peaks by minute.

For each group of subjects, the galvanic skin response (GSR) is analyzed according to the success or the failure in performing the different tasks. For the first session, the results show an effect on the relative means concerning the factor success/failing ($F=4,67$; $p<0,033$): the mean of the values of the GSR is higher in a success situation than in a failure situation. It seems that a success generates a greater affect than a failure. For the second session, there is an effect of the success/failure factor observed with the number of peaks ($F=9,8$; $p<0,002$). This result indicates a nervous irritation of subjects in a failure situation.

Figure 6 shows that the number of peaks increases between Session 1 and Session 2; this indicates an increasing reactivity of the subjects during the test.

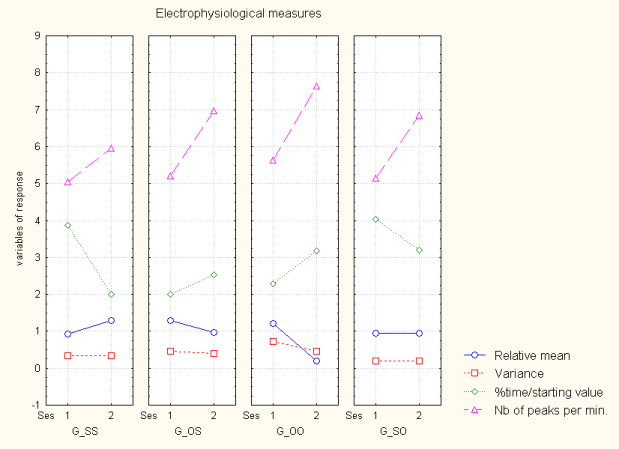

Figure 6. *Effect of the group and the session on the GSR*

Globally, the analysis of the galvanic skin responses allows neither to reinforce nor to invalidate the data obtained by questionnaires and analysis of the activity. The richness of such an analysis lays in the real-time analysis of the measured data, as with the increasing of the GSR when the subject is stressed.

4. DISCUSSION

Our results could not validate our four hypotheses: the sonified version of the server did not improve users' performances during the first utilization, nor during the second. Nor did it satisfy the users more in terms of pleasantness and user-friendliness. The main explanation for this might come from the observation that the vocal server was rather complex to use, with long prompts pronounced at a rather rapid rhythm. Users had to really concentrate on what was being said, and additional sounds were hardly noticed. This was confirmed when subjects who had experienced a sonified version were asked if they had noticed musical sounds and if so, how many. More than 80% answered 'around two', although they all had actually heard at least four sounds.

The interaction with a vocal server requires the full mobilization of the cognitive resources of users since, as Rosson [11] emphasizes, speech prompts must perform two tasks: navigation (procedural information), and information (that information of interest for the user). This dual role of speech is further complicated by the sequential delivering of the information, which means that users can only get *a posteriori* information about their actions (whereas a real-time monitoring of actions is possible in graphic interfaces using visual feedback). This particularity of vocal interfaces forces users to concentrate on the verbal content of the messages to accomplish their tasks and makes them feel that non-verbal messages are more decorative than useful (although users sometimes expressed that some audio feedback might have helped them).

Although Brewster's and his colleagues' results on graphic interfaces and on tree hierarchies were promising, it seems difficult to apply them to vocal servers in an interactive context. However, our results might be highly dependent on the type of sounds we created and on the function we attributed to them. Maybe a different sonification would have yielded a more positive impact of the insertion of musical sounds in our vocal server. Further research is needed to test these hypotheses. Maybe a design based on a sound-oriented approach could be considered, where the sounds would not be added as added-on components but would be considered as the basis of the vocal server.

5. REFERENCES

[1] Blattner, M., Sumikawa, D. & Greenberg, R. "Earcons and Icons: Their Structure and Common Design Principles", *Human Computer Interaction*, Vol. 4(1), pp. 11-44, 1989.

[2] Brewster, S. A., Wright, P. C. and Edwards, A. D. N. "Experimentally Derived Guidelines for the Creation of Earcons", in *Adjunct Proc. of HCI'95*, Huddersfield, UK. Microsoft, 1995.

[3] Brewster, S. A. "Using Non-Speech Sound to Overcome Information Overload", *Displays, Special issue on Multimedia displays*, Vol. 17, pp 179-189, 1997.

[4] Brewster, S. A. "Navigating Telephone-Based Interfaces with Earcons", in *Proc. of BCS HCI'97* (Bristol, UK), Springer Verlag, pp. 39-56, 1997.

[5] Brewster, S. A. "Using Non-speech Sounds to Provide Navigation Cues", in *ACM Transactions on Computer-Human Interaction*, Vol. 5(2), pp. 224-259, 1998.

[6] Wolf, C., Koved, L. and Kunzinger, E. "Ubiquitous Mail: Speech and Graphical Interfaces to an Integrated Voice/Email Mailbox", in Nordby, Helmersen, Gilmore and Arnesen (Eds.) *Proc. of IFIP Interact'95*, Lillehammer, Norway: Chapman & Hall, pp. 247-252, 1995.

[7] Gaver, W. W. "Auditory interfaces", in *Handbook of Human-Computer Interaction,* 2nd edition, M.G. Helander, T.K. Landauer and P. Prabhu (Editors), Elsevier Science, The Netherlands, Amsterdam, 1997.

[8] Leplâtre, G. and Brewster S. A. "An Investigation of Using Music to Provide Navigation Cues" in *Proc. of ICAD'98* (Glasgow, UK), British Computer Society, 1998.

[9] See the Web site: http://www.thoughttechnology.com/

[10] See the Web site: http://www.mixman.com

[11] Rosson, M. B. "Using Synthetic Speech for Remote Access to Information," *Behaviour Research Methods, Instruments and Computers*, Vol. 17, pp. 250-252, 1985.