Designing Non-Speech Sounds to Support Navigation in Mobile Phone Menus

Grégory Leplâtre and Stephen A. Brewster
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ Scotland
+44 141 339 8855
gregory@dcs.gla.ac.uk
http://www.dcs.gla.ac.uk/∼gregory

ABSTRACT
This paper describes a framework for integrating non-speech audio to hierarchical menu structures where the visual feedback is limited. In the first part of this paper, emphasis is put on how to extract sound design principles from actual navigation problems. These design principles are then applied in the second part, through the design, implementation and evaluation of a set of sounds in a computer-based simulation of the Nokia 6110 mobile phone. The evaluation indicates that non-speech sound improves the performance of navigational tasks in terms of the number of errors made and the number of keypresses taken to complete the given tasks. This study provides both theoretical and practical insights about the design of audio cues intended to support navigation in complex menu structures.
Keywords Telephone-Based Interfaces, Mobile Phones, Navigation, Menus, Sonification.

INTRODUCTION

This paper describes a framework for using non-speech audio to support navigation in menu-based interfaces such as mobile phone interfaces where the visual feedback is limited. Our approach puts emphasis on providing sound design principles to support real-life navigation tasks in a given interface. In previous work, it has been shown that non-speech audio - namely *earcons* [1] - could be used successfully to represent a hierarchical structure such as a menu structure [2, 6]. Nevertheless it has remained unproven whether sound could increase the usability of a complex, real-life menu. Therefore we have implemented a set of sounds into a computer-based simulation of the Nokia 6110 mobile phone to evaluate their benefit as far as managing ecologically valid navigational tasks is concerned.

In this paper, we evaluate the latter sonified interface by comparing the performance of two groups on a set of navigational tasks. The first group (Group 1) used the sonified interface, and the second group (Group 2) used the simulation of the original device to complete the same set of tasks. The performances are measured in terms of the number of keypresses taken to complete the tasks, average time between two successive keypresses, and error made. The subjective workload as well as data related to the performance of two tasks in blind-condition are also presented.

USING NON-SPEECH SOUND TO SUPPORT NAVIGATION

Why Support Navigation?

The telephone is a ubiquitous device and the preferred way for many to complete an increasing number of tasks. It also causes many users frustration. Roberts *et al.* suggest that the problems experienced by users while using a telephone-based interface is not the consequence of a poorly designed system, but result from the nature of the interface itself [9]. The main drawbacks of the device are: Auditory interaction - all information is presented serially, limited input device and wide variability in telephone-based equipment.

Schumacher *et al.* maintain that the narrow channel of interaction resulting from the structure of the interface reduces the usability of telephone-based systems [10]. Apart from interaction problems caused by the physical structure of the interface, the increase of functions available in these systems contributes towards making them difficult to use [7]. Practically, Yankelovich *et al.* argue that one of the major problems with telephone-based interfaces centres on navigation [11].

Previous Work

The possibility that earcons could represent hierarchical structures has been proven successful in previous research [2]. Earcons are structured non-speech sounds that can be combined, transformed, can inherit other earcons properties and constitute an auditory language of representation. These functions can be applied to create hierarchical structures of sounds called *hierarchical earcons*. Brewster *et al.* conducted different experiments that investigated the use of earcons to represent hierarchical structures. These experiments have shown that people could recall a 25-node hierarchy of earcons with good accuracy [2]. Subsequently, we have shown that musically trained as well as non-musically trained subjects could accurately recall a set of 25 sounds using a single instrument relying on syntactic musical patterns [6].

Navigating With Non-Speech Sound

There are many ways to envisage navigation in a menu. This tends to make the term navigation too general for our purposes. One can talk about searching, browsing, scanning and so on, all under the term 'navigation'. Before considering any attempt at sonifying a menu structure, it is essential that we understand what is meant by the term 'navigation'. Norman's book on menu selection provides valuable insight on the issue [8]. The integration of sound into menus will inevitably be based on the *information acquisition process* [8]. This process invariably underlies all different styles or strategies of searching mentioned above. Accordingly, Howes points out three broad aspects of human behaviour that account for the menu acquisition process [5]. Below, we describe how non-speech audio can be integrated into a menu structure in respect to these aspects:

- **People learn devices by exploration**
  Let us consider the following analogy: In real-life, people orientate themselves by using elements of the environment as navigational cues. These elements can have been intentionally placed in the environment for this purpose - road signs - or they can be chosen arbitrarily by individuals - petrol stations, a well-known building, and so on. If we transpose this behaviour to navigation in a menu with limited graphical feedback, it is clear that the same amount of environmental information is not made available to users. As the interaction is mostly temporal, the amount of auditory information has to remain limited. In particular, cues that are artificially placed in the environment - road signs - have to be restricted when presented sonically. Unfortunately, in essence this category of cues tends to be the most effective, as they are designed to support navigational tasks. As a result, sounds should be integrated into the menu environment as transparently as possible; they should seem to fit "naturally" in their environment. Thus, sounds need to be integrated into the menu as part of the menu structure, rather than on the top of it. If we come back to the previous real-life analogy, the environmental cues of a menu are the menu items themselves. The difference between the semantic content of the menu items is the main navigational cue available to users. Indeed, navigation is primarily driven by the semantic relations between menu items. A means of making these semantic differences more obvious involves associating a sound with a menu item so that the semantic content of the sound adds to that of the menu item label.

- **People acquire display-based knowledge**
  People use elements of the graphical display to learn the structure of the menu. It has been shown that when in front of a device, users could effectively perform a succession of menu selections, but away from the device they could not report the names or order of the constituent commands used. In other words, users do not recall of a whole stream of successive actions, but rather build a mental model of the menu structure in which, at each point of the stream, they know what action to take to progress towards their initial goal. In this process, the context in which each action is taken helps recollection of what the next action should be. The visual context is a major factor in the operation. Sound should be used in graphically limited menus to overcome this lack of graphical context.

- **People improve with practice**
  This statement suggests that users' needs of navigational cues decrease over time. Consequently, the amount of sounds provided to support navigation should accommodate to these needs. Different levels of sonification should be available to users.

CASE STUDY: SONIFICATION OF A COMPLEX MOBILE PHONE MENU

This case-study has been realised on the Nokia 6110 mobile phone. The technology did not permit us to implement sounds in the device, therefore we have developed a Java simulation of the device in order to achieve this. The Java Sound API released in the JDK1.3 Beta was used to play audio files. The simulation window is displayed in Figure 1. For the purpose of this experiment, all the functions of the device were not implemented in the simulation: only the menu features which have an incidence on the menu behaviour itself have been implemented. For instance, it was possible to carry out operations such as: personalise a profile, change some of the settings, divert calls, bar calls or check and erase calls. On the other hand, it was not made possible to carry out such operations as: give a call, use the alphanumeric keypad, and more generally type data into the device, or use the phone repertoire.

Five buttons were implemented in the simulation, allowing users to navigate in the menu structure. These buttons are indicated on Figure 1. Buttons 1 and 2 allowed users to performed the action specified just above them on the screen. In the example of Figure 1, Button 1 and 2 allowed users to *Select* the item (i.e., go down one level in the menu hierarchy), and to go *Back* (i.e.,

Figure 1: Java window containing the simulation of the Nokia 6110. On this picture, the size of the window has been scaled down. The simulation has the same size as that of the real device.

go back up one level in the menu hierarchy), respectively. These two buttons consitently allowed users to go up and down in the hierarchy. Buttons 3 and 4 are used to scan lists of items at a given level of the hierarchy. It is important to notice that the interface allowed users to loop over a list of menus. Button 5 allowed users to jump to the top of the menu from any position within the menu.

Sound design principles

The sonification implemented consisted of approximately 150 different sounds which have been integrated to the menu. Each node of the menu hierarchy is associated with a sound that is played when it has been reached. We have based our design on a number of principles described below[1].

- **Distinctiveness of main menu sounds**
  As Brewster *et al.*'s study indicates, the most efficient parameter that can be used to distinguish sounds is timbre [2]. Accordingly, all sounds within a main menu have been designed using the same instrument. Additionally, arpeggios and chords have been alternated as motifs for the top level menu. For instance, the first main menu item (*Messages*) is associated with a piano chord, the second (*Call register*) was associated with a pitched wooden block arpeggio, and the third one (*Profiles*) with a voice-like sound chord. Moreover, the ranges of pitches of chords and arpeggios have been chosen so that the motifs sound even more distinct.

- **Homogeneity of the sonification**
  It is particularly important that the streams of sounds which result from navigation in the menu, sound as homogeneous as possible. While scanning the main menu items, the instrument changes each time a sound is played. To maintain cohesiveness among these sounds, the 9 top level menu motifs were designed over a harmonic movement that resolves on the first menu item. This also allowed users to notice more easily when they have completed a loop through all the top menu items. As a result, each main menu has a particular tonality, which is preserved in all its submenus.

- **Distinctiveness and gradation of each menu level**
  Users tend to loose track of their position when they perform many horizontal movements - scanning a list of items - and vertical movements - selecting menu items. It is therefore important to help users keep track of the level they are at in the hierarchy. Simultaneous to each sound, a brief percussive sound was played. This sound changed at each level. Such cues have a much lower intensity than the motif they accompany, as it is only important that the users notice that the cue has changed, once they have moved up or down a level. In addition, the duration of the sounds decreases as we go deeper

[1] The sonified simulation is available from: http://www.dcs.gla.ac.uk/∼gregory

in the hierarchy. The motifs also become less and less complex as the depth of the hierarchy increases. These last two aspects are closely related to the last point of this section.

- **Use of brief sounds**
  Time is a critical factor in the interaction; the integration of sounds in a menu should not slow users down. Therefore, sounds should be brief. Or, more accurately, the effective part of each sound should be brief, but the sound itself could feasibly last longer. For example, a chord associated with a top level menu item can last 2 seconds, but only the first hundred milliseconds need to be heard by the user. As the sound envelope decreases rapidly, the user can skip to another menu item whose sound will be played before the previous one is finished. However, no information will be missed by the user, but, regarding the careful choice of amplitude envelopes, the transition should happen smoothly, even if the user were to navigate quickly.

- **Distribution of sound *weights***
  According to the usual rules of construction of hierarchical earcons, the sounds of each level of the hierarchy inherit their properties from the sounds of the level above them, plus supplementary properties from their own level. As a result, sounds become more complex as the depth of the hierarchy increases. This factor makes the previous design principle difficult to implement in large hierarchical structures. Moreover, this design does not correspond with real-life navigational tasks as discussed in the first section of this paper. Therefore we have designed the sounds so that their complexity decreases as the depth of the hierarchy increases. The *weight* of each sound corresponds to the importance of the menu item it is related to. This notion of weight would take a lot of space to be defined properly, but is easy to undestand on a simple example: Figure 2 shows the hierarchical structure of one of the submenus of the *Profiles* menu. The right part of this menu clearly contains more information that the left part. Accordingly, we have designed the sounds of the two children of the root menu item as follows: the left one is made of one note from the chord associated with its father whereas the right one is made of the three remaining notes of its father chord (see Figure 3). This provides users with information about what is available below their position in the menu.

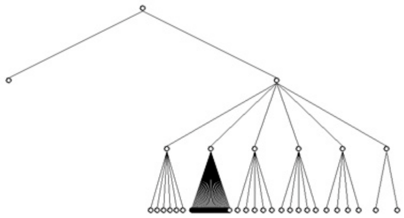
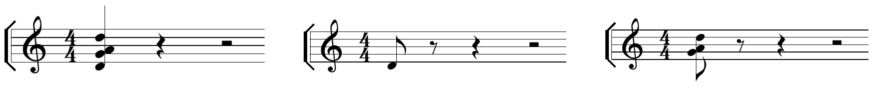


Figure 2: Hierarchical structure of the *Meeting* profile, which is the third submenu of the *Profiles* menu. The five profiles of the *Profiles* menu have the same hierarchical structure.



(a) root (b) left child (c) right child

Figure 3: Deconstruction of a menu item sound into its two children according to the amount of information available underneath them.

EVALUATION

Subjects

Twenty four subjects from the University of Glasgow, two groups of twelve, were used for this experiment. All the subjects had no experience of using the Nokia 6110 or any other Nokia mobile phone. In addition, to limit the influence of transfer from mobile phones of other brands, we tried to recruit subjects who would have no experience of mobile phones at all. We managed to find ten of them for each group, the remaining two in each group had a limited experience of using a mobile phone menu. Because they were all computer literate, the subjects were all familiar with navigation in hierarchical menus.

Experimental design

The experiment consisted of 56 tasks dedicated to finding a particular menu item, or execute an action at different times of the experiment. The first seven tasks were considered as practice tasks to make sure that all the subjects understood how to navigate in the menu in the Java simulation of the Nokia 6110. The remaining tasks were spread over the experiemnt as shown in Tables 1 and 2.

Occurrences	*Call Divert*	*Profiles*	*Call register*	*Call barrings*
1	18	8	23	28
2	20	9	34	30
3	32	11	35	43
4	37	12	36	51
5	38	14	42	55
6	44	16	57	
7	56	24		
8		33		
9		40		
10		50		

Table 1: Occurences of tasks related to three main menus: *Call Diverts*, *Profiles*, *Call Register*, and to the *Call barrings* menu.

Occurrences	*Welcome Note*	*Time*	*Language*	*Any Key Answers*	*Own Number Sending*	*Speed Dialling*
1	21	22	27	25	26	31
2	52	39	41			53
3		48				

Table 2: Occurences of tasks related to *isolated* menu items in the *Settings* menu.

As mentioned in the introduction, Group 1 performed the tasks using a sonified simulation of the Nokia 6110, whereas Group 2 used an exact simulation of the device. After the experiment, both groups were requested to fill a standard NASA-TLX workload form [3]. An *annoyance* entry was added to the form for Group 1, to have some feedback about how annoying they found the sounds.

Additional tasks for Group 1

In order to gather additional information concerning how the sounds were understood by the subjects, we made group 1 perform two tasks in blind conditions. After they had completed the 56 main tasks and then filled in a standard NASA-TLX workload form, we asked them to perform two more tasks. At this point, the mobile phone simulation screen went blank, and the only feedback available was auditory. The two tasks were:

1. Press the red button to go back to the top level. Then try and activate the last profile in the 'Profiles' menu.
2. Try and cancel all call divert (you may use the red button to go back to the top level menu first). If you do not know how to do that, just try and select the 'Call divert' menu.

Hypotheses

To compare the performance of both groups, we have looked at the keypress numbers, the average time between successive keypresses and the error made. Globally, we expect non-speech sound to help users build a better mental model of the menu structure; A better global mental representation of the hierarchy through sounds should help Group 1 completing the tasks more effectively. In addition, assuming that sound helps encode and recall the menu items more easily, the performance of Group 1 should improve more throughout the experiment than that of Group 2. This should be particularly noticeable for the tasks involving what we define as *isolated*[2] menu items.

1. **Number of keypresses to complete the tasks**
   (a) **Overall difference between groups**
   Globally, we assume that Group 1 will use less keypresses to perform the tasks.

[2] By isolated, we mean that the items have no clear semantic connection to the menu they belong to. On the contrary, for example, *Divert when busy* is strongly connected to its father: *Call divert*.

(b) **Evolution of the difference between groups**
In addition, we hypothesise that the difference of performance of the tasks between the two groups in terms of number of keypresses will increase. In other words, Group 1 are expected to perform the tasks increasingly more efficiently than Group 2.

2. **Keypress time**
Given the care taken to minimise the lengths of the sounds, we do not expect the average time between two successive keypresses to be significantly different for both groups.

3. **Errors**
Some of the tasks being fairly demanding, we can assume that some subjects will fail them. Subjects may either complete a task incorrectly, or give up completing a task. We expect subjects from Group 1 to complete more tasks successfully than subjects from Group 2.

4. **Difference of performance between groups for tasks related to isolated menu items**
Given that the sounds have been designed to increase the semantic content of menu items, we expect the difference of performances between the groups obtained on the tasks related to isolated items displayed in Table 2, to increase over the experiment.

RESULTS

Number of keypresses

Overall difference between groups

Firstly, we have looked at the total number of keypresses used by each participant for each task over the whole experiment This analysis provided us with a primary indication of the differences of performances between the groups. Subjects from Group 1 averaged 1162 keypresses to complete the experiment, whereas subjects from Group 2 took 17% more i.e., 1362 keypresses. There was a noticeable difference between the variances of the results within the groups: 34074 for Group 1 and 108038 for Group 2. An F-test on both samples showed that, the total number of keypresses of Group 2 subjects were significantly more dispersed than these of Group 1 ($F = 0.31, p = 0.034$). Overall the variance was quite high, which was to be expected due to the limited amount of subjects and the nature of the tasks proposed. Nevertheless, despite the high variance of the results, the difference of performance in terms of the number of keypresses approached significance ($T_{11} = -1.83, p = 0.084$).

The number of keypresses is a good indicator of how efficiently the tasks have been performed. However, several issues arise whilst running statistical tests on these raw data: Firstly, the number of keypresses taken to complete a task does not take into account whether a task has been completed successfully or not. In addition, given the nature of the tasks, the total number of keypresses tends to be dispersed. Running statistics on these values for a relatively limited number of subjects is not representative of the real differences between the group performances. Moreover, the distribution of the total number of keypresses is quite heavily tailed, especially for Group 2. As mentioned above, the variances of the groups are highly heterogeneous. Finally, the numbers of keypresses for each task are heterogeneous.

A common practice to accomodate the data is to perform classical transformations of the values [4]. However, logarithmic or square root transformations would not be helpful since the data of Group 2 are heavily tailed in both low and high ends. An alternative method might involve using trimmed samples, but given the relatively small amount of data, reducing the size of the samples does not seem an appropriate solution either. Besides, all these transformations do not address the issue of tasks completed successfully or not. In the present case, the most adequate transformation of the data appears to be normalising the number of keypresses for each task and subject. This method also allows us to take into account success or failure. To normalise the data, we divided the number of keypresses to complete a task, by the maximum number of keypresses over the 24 subjects for this task.

The distributions of the normalised number of keypresses (NNKP) for each group was far less skewed and the variances of the two groups are in a similar range. The variance of the total NNKP per group across the 42 non-trial task was 2.46 for Group 1 and 2.31 for Group 2. A t-test on these 2 groups of 42 values showed that the difference of means (4.07 for Group 1 versus 4.70 for Group 2) approached significance ($T_{41} = -1.84, p = 0.069$). Similarly, the variances of the total NNKP per task accross the 12 subjects of each group were quite similar (3.43 for Group 1 versus 8.40 for Group 2). A t-test on these 2 groups of 12 values showed a significance difference between the means (14.25 for Group 1 versus 16.42 for Group 2, $T_{11} = -2.19, p = 0.040$).

Regardless of whether the tasks have been completed successfully or not, these results tend to prove hypothesis 1a. However, even though most of the tasks have been completed successfully, it seems unfair to discard this information in the preceding statistics. We propose a means of taking this information into account whilst looking at keypress numbers.

There are two main concerns with unsuccessful tasks. On the one hand, a task can be completed unsuccessfully because - after a long searc - a subject may have given up looking for the requested menu item. In this case, where the subject is lost, the number of keypresses for this subject and task would be substantially high. On the other hand, the re-occurrence of a task that has not been completed successfully by a subject at previous attempts night result in the subject not attempting it again. In this case, the total number of keypresses, for this task and the group which the subect belongs, would be unfairly low Depending on the context, unsuccessful tasks may then result in an unfair deviation of the data.

The most adequate treatment of the data ties in replacing the number of keypresses for given task and subject, by the maximum number of keypresses taken by any of the 24 subjects to complete that task successfully. The same tests as previously have been performed on these data. Again, the variance of the total NNKP accross the 42 non-trial tasks were similar for both groups (3.48 fr Group 1 and 4.04 for Group 2). A t-test showed that there was a significant difference between the means of total NNKP per task of the 2 groups ($T_{41} = -2.26, p = 0.027$). Similarly, a t-test showed that there was a significant difference between the means of total NNKP per subject for the 2 groups ($T_{11} = -2.24, p = 0.040$). These results show that taking into account success or failure to complete a task increases the differences between Group 1 and Group 2 as far as numbers of keypresses are concerned. This meets our hypotheses.

Evolution of the difference between groups

The second hypothesis we have formulated regarding keypresses was related to the evolution of the data in time. As for the previous analysis, we have looked at the original data, and at treated data to show the influence of success or failure in the account of keypresses. The method used to compare how the difference between groups evolve consisted in running a linear regression on the variable: $T(task) = T_1(task)/T_2(task)$, for the 42 non-trial tasks, where $T_1(task)$ and $T_2(task)$ are the keypress totals for $task$, for groups 1 and 2 respectively.

Without adjustment, the slope of the regression line was almost flat, which indicated that the difference of performance of the 2 groups was constant throughout the experiment. However, this measure is highly biased by the fact that the number of failed tasks increased more in Group 2 than in Group 1 throughout the experiment (cf. Errors section, below): Most of the late experiment tasks failed have not been attempted at all (The number of keypresses for such task and subject is null). Therefore, the analysis performed on the raw data does not seem to be satisfactorily meaningful.

With adjustment[3], the regression on the 42 tasks showed a big increase of the difference of performance of the 2 groups throughout the experiment. The initial value of the regression line ($task = 1$) is 0.94, versus 0.72 for the final value ($task = 42$). In other words, group 1 took 6% less keypresses than group 2 at the beginning of the experiment, and 28% less at the end. This validates hypothesis 1b.

Keypress time

The main concern with the use of non-speech sound is that it tends to increase the time spent to complete tasks. Indeed, as auditory interaction is temporal, content added to the audio channel adds to the time required to process it. However, in the present context, the interaction is not exclusively auditory. Sounds is used to add to the visual mode and have been designed to fit in temporally in parallel with the time spent by a user to visualise a menu item. The comparison of the average keypress times for each group indicates whether the design has been successful or not in this department.

The time spent between two successive keypresses has been collected for each task and participant. Through the whole experiment (56 tasks), the mean time spent on a menu item was $1.14s$ for Group 1 versus $1.08s$ for Group 2. A t-test on the average keypress time per subject showed that this very small difference was far from significant ($T_{11} = 0.66, p = 0.51$). A t-test on the average keypress time per task confirmed this result ($T_{55} = 74, p = 0.46$). These results validate hypothesis 2.

Errors

As part of the experiment's instructions, no request was made about success in completing a task in order to be able to proceed to the next one. Therefore, it was to be expected that some subjects would give up after a while spent unsuccessfully attempting a task. There were two different categories of tasks:

1. Tasks for which the subjects had to *find* a menu item.
2. Tasks that required the subjects to *complete* an action.

For the first category, a task was considered successful when a subject had selected the required item. For the second category, a task was considered successful when the action was completed. Practically, this scenario corresponded with the selection of a required menu item followed by the selection of the completion item: *DONE*.

[3] Let us recall that the adjustment is the operation that allowed us to take success and failure of tasks in the account of the keypress numbers

Group 1	41	41	41	42	42	42	38	41	42	41	42	42
Group 2	31	42	39	40	40	41	42	36	40	34	34	42

Table 3: Totals of tasks completed successfully by the 12 subjects of Group 1 and Group 2. The maximum score is 42.

Table 3 shows the scores achieved by both groups. The average score for Group 1 was 41.25 versus 38.42 for Group 2 (respectively 98.2 and 91.5 tasks completed successfully in average). A t-test performed on the data displayed in Table 3 indicates that this difference was significant ($T_{12} = 2.52, p = 0.026$). These statistics validate hypothesis 3.

We have mentioned above that the number of errors increased more for Group 2 than for Group 1. To ascertain this, we have performed a linear regression on the differences of tasks completed successfully in each group. The correlation between X and Y is close to average ($r = 0.40$). The initial value of the regression line ($task = 1$) is 0.05, versus -1.67 for the final value ($task = 42$). An ANOVA analysis showed a high significance of the regression ($F = 6.60, p = 0.014$). These results show that the evolution of the difference of performance between the groups in terms of errors made is largely in favour of Group 1.

Difference of performance between groups for tasks related to isolated menu items

To measure this difference, we have performed a linear regression through the differences of normalised keypress number for the eleven tasks presented in Table 2. Figure 4 shows the result of the regression analysis. The graph clearly indicates a positive slope of the line. The initial value of the line ($task = 1$) is 0.15, versus 1.47 for the final value ($task = 11$). This difference is relatively large given that the average of NNKP for Group 1 and Group 2 are, 4.06 and 5.16 respectively. The correlation between X and Y is fairly high ($r = 0.66$), and an ANOVA analysis showed a high significance of the regression ($F = 7.04, p = 0.026$). These results validate hypothesis 4.

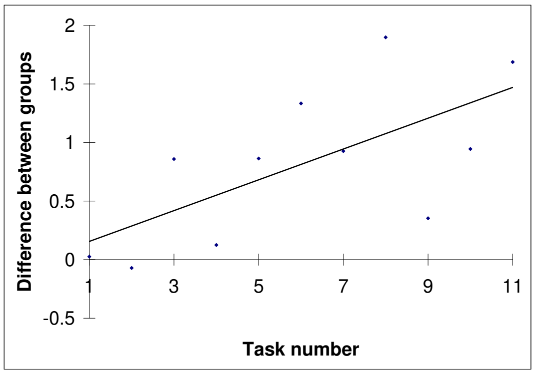


Figure 4: Regression analysis on the difference of performance of tasks related to isolated menu items.

Additional results

In addition to the data analysed above, we have also collected more data about which we did not formulate any hypothesis. This section presents two of the main additional results.

Tasks in blind condition for Group 1

As mentioned earlier in this chapter, Group 1 were requested to perform two tasks in blind condition at the end of the experiment. For technical reasons, the first subject of the group did not perform these additional tasks. Therefore, only 11 data were collected for these two tasks.

This part of the experiment has been developed for a purely exploratory purpose, therefore, more than the amount of right or wrong answers, it was the types or errors made in which we were interested. Out of the 11 subjects investigated, only 2 managed to complete the first task accurately. 5 managed to complete the second task perferctly. Interesting details arised while looking closer at the data. To perform task 1, subjects had to perform the following actions:

1. Click Button 5 to go to the top of the menu. The feedback to this action is the sound of the top menu item.
2. Click the Button 1 to go down one level i.e., to the first of the main menu items. The feedback for this action is a piano chord, which is the sound for the *Messages* menu.
3. Use Buttons 3 or 4 to scan the list of main menu items until the user thinks the current item is *Profiles*. The feedback is the sound of each main menu item whilst browsed.

4. Click Button 1 to select that menu. The feedback is the sound of the first submenu item.

5. Use Buttons 3 or 4 to reach the last submenu of the list

6. Press Button 1 to select the last submenu, which is *Pager* if the user has commited no mistake so far. the feedback is the sound of the first submenu of the *Pager* submenu, i.e., *activate*.

7. Press Button 1 to select *Activate*. The feedback for this action is the completion sound being played.

Seven out of eleven subjects completed up to step 6 successfully. two of them completed step 7 as well, but the remaining five stopped after completing step 6. It is likely that they believed the profile would be activated from selecting the profile. Therefore, these 5 subjects did not notice the absence of the completion sound to realise they had not completed the requested action. In the remaining 4 subjects, 3 failed at step 3. Two selected the *Call divert* menu and then performed the next steps successfully, and 1 selected the *Settings* menu. These 3 individuals failed at identifying the characteristic sound of the *Profiles* menu. The last subject failed at step 5, and selected the wrong profile.

The second task could be deconstructed in a similar fashion as the first one, without the last step. Additionally, to complete step 6 users were supposed to remember that *Cancel all* was the last of *Call divert* submenus. Five out of eleven participants managed to complete the task successfully. 4 subjects completed the task up to step 4 and then, did not try to go further (2 od them) or got lost in the *call divert* menu before they gave up (the other 2). This indicates that these subjects did not recall that the *Cancel all* submenu was the last of the *Call divert* submenus. The remaining 2 subjects of the group did not manage to select the *Call divert* menu.

One should be careful when interpretating these results. What these tasks evaluate is the understanding of the sounds outside of their context and thus away from the context for which they have been designed. One should remember that the sounds have been designed to increase the semantic content of menu items, not to replace it. Consequently, navigating with sounds as the only feedback can be considered highly demanding.

Overall, 77.3% managed to select the correct main menu with audio feedback only. A level below, 54.5% mangaged to select the correct submenu: $7/11$ when the position of the item was specified (first task), and $5/11$ when the position was not specified (second task). However, 5 subjects failed to notice the absence of the completion sound in the first task, which would have indicated that the task was complete.

Workload

The results of the workload questionnaire are compiled in Table 4. The mental demand and effort expended have been rated higher by Group 1 than by Group 2, which tend to indicate that the better performance achieved by Group 1 required them a greater effort. In the light of these results, the difference of performance observed in the previous sections seem to show that Group 1 have learned more about the structure of the menu. This has a cost in terms of mental effort. On the other hand, The frustration experienced was rated slightly higher in average by Group 2. An interesting point concerns the time pressure. There was no time constraint imposed on the subjects in this experiment. Therefore it was expected to involve values as low as those of Physical Demand. However, this component was rated quite high, especially by Group 2. A plausible explanation for this point is that people rated the time pressure high when they felt they had spent more time on the experiment, or at least on some of the tasks, than they should have. In this respect, Group 1 seemed to be happier with the time they spent performing the experiment than Group 2. As for the performance level achieved, the difference between the group is marginal. Looking closer at the data, it occured that indeed, some of the subjects from group 2 whose performance were quite poor rated their performance high. On the other hand, subjects from Group 1 seemed to be more consitently modest. Finally, the annoyance: these data have only been collected for Group 1 as they refered specifically to the sounds. The average value of annoyance reached 34.7%. Compared to the physical demand (20%), which can stand for a reference as a low value, given that the experiment was not physically demanding, and also compared to the time pressure, which we assumed to be low, the annoyance has been rated fairly low. 6 out of 12 subjects rated the annoyance of the sounds below 20% whereas 1 only rated it above 80%. All the differences mentioned above were not significant (see Table 4).

CONCLUSION AND FUTURE WORK

This case study has allowed us to take a step forward in understanding how sounds can help support navigation in a complex menu structure by tackling a real-world problem. In particular, we have outlined a framework for divising sound design principles from looking at actual navigation issues. The evaluation of the mobile phone simulations investigated have shown that user performance can significantly benefit from the use non-speech sound. The main results have shown that users who used the sonified simulation took less keypresses to complete the tasks and completed more tasks successfully. Moreover their performance improved more over the experiment than that of the subjects using the unsonified simulation of the mobile phone.

	Group 1	Group 2	T-Test
Mental Demand	64.5	51.3	0.185
Physical Demand	19.9	20.5	0.940
Time Pressure	38.2	53.8	0.155
Effort Expanded	64.1	51.3	0.227
Performance Level Achieved	54.9	51.3	0.710
Frustration Experienced	46.2	49.5	0.804
Annoyance	34.7		

Table 4: Results of the workload test. The last three columns compile respectively: the means (in percent) for each component of the workload for Group 1 and for Group 2, and the probability ($p(T_{12})$) associated with a t-test that indicates whether the differences of means are significant or not.

However, some work still needs to be done in both HCI and sound design sides of the project. For instance, an important issue which needs to be addressed is that of *levels* of sonification: design sets of sounds that meet different users' needs and preferences. The next stage of this project will involve producing guidelines to support the creation of sounds for similar purposes, involving both usability and design issues.

ACKNOWLEDGMENTS

This research has been founded by EPSRC grant number GR/L66373. We would also like to thank Nokia for providing us with the mobile phone used in this case study.

REFERENCES

1 M. Blattner, D. Sumikawa, and R. Greenberg. Earcons and icons: Their structure and common design principles. *Human Computer Interaction*, 4(1):11–44, 1989.

2 S.A. Brewster, V.-P. Raty, and A. Kortekangas. Earcons as a method of providing navigational cues in a menu hierarchy. In A. Sasse, R. Cunnigham, and R. Winder, editors, *Proceedings of BCS HCI'96*, pages 169–183, London, UK, 1996. Springer.

3 NASA Human Performance Research Group. Task load index (nsa-tlx) v1.0 computerised version. Technical report, NASA Ames Research Centre, 1987.

4 David C. Howell. *Statistical Methods for Psychology*. PWS-KENT Pusblishing Company, third edition, 1992.

5 Andrew Howes. A model of the acquisition of menu knowledge by exploration. In *Proceedings of CHI'94*, pages 445–451, Boston, Massachusetts USA, 1994. ACM Press.

6 Grégory Leplâtre and Stephen A. Brewster. An investigation of using music to provide navigation cues. In *Proceedings of ICAD'98*, Glasgow, UK, 1998.

7 Martin Maguire. A human-factors study of telephone developments and convergence. *Contemporary Ergonomics*, pages 446–451, 1996.

8 K. L. Norman. *The psychology of menu selection: Designing cognitive control at the human/computer interface*. Ablex Publishing Corporation, 1991.

9 Teresa L. Roberts and George Engelbeck. The effects of device technology on the usability of advanced telephone functions. In *Proceedings of ACM CHI'89 Conference on human factors in computing systems*, pages 331–337, Austin, Texas, USA, 1989. ACM Press.

10 R.M. Schumacher, M.L. Hardzinski, and A.L. Schwartz. Increasing the usability of interactive voice response systems. *Human Factors*, 37(2):251–264, 1995.

11 Nicole Yankelovich, G. Levow, and M. Marx. Designing speechacts: Issue in speech and user interfaces. In *Proceedings of CHI'95 Conference on Human Factors in Computing Systems*, Denver, CO, 1995. ACM Press.