AUDITORY ICON SUPPORT FOR NAVIGATION IN SPEECH-ONLY INTERFACES FOR ROOM-BASED DESIGN METAPHORS

Daniel Skantze

CTT, Department of Speech, Music, and Hearing
Royal Institute of Technology (KTH)
SE-100 44 Stockholm, Sweden
daniels@speech.kth.se

Nils Dahlbäck

Department of Computer and Information Science
Linköping University
SE-581 83 Linköping, Sweden
nilda@ida.liu.se

ABSTRACT

In this paper, a navigation support approach for speech-only interaction based on auditory icons for room-based designs is presented, i.e. naturally occurring sounds that have a natural mapping to the system's underlying design metaphor. In contrast to many recent investigations that have focused on multi-modal or augmented reality systems, this paper concerns a unimodal speech and sound-only system. Auditory icons and earcons were evaluated in a prototype system for building maintenance support based on this design approach. The subjects' subjective attitudes toward auditory icons were significantly more positive than to earcons.

1. INTRODUCTION

Designers of telephone-based and other speech and sound-only interfaces face several challenges; one of the most important being the problem of designing complex systems that are easy to navigate. One way of supporting user navigation has been to use sounds to indicate certain locations and events within the system, e.g. as seen in Vodafone's voice enabled entertainment portal [9]. Another approach to improve navigation in these kinds of interfaces has been to investigate alternatives to the dominating menu-metaphor. There is some support that a room based metaphor, or more specifically, the warehouse metaphor, can improve user performance and their subjective attitudes towards the system, compared to a traditional menu-metaphor based system [3].

In this study, the use of auditory icons was evaluated in a voice-based system for buildings maintenance support; earcons were developed as a reference. The system is based on a room metaphor although it is a prototype and does not model an existing building.

1.1. Earcons vs. Auditory Icons

The most common type of sounds discussed today is so-called earcons. Brewster describes earcons as: "*Earcons are abstract, musical tones that can be used in structured combinations to create sound messages to represent parts of an interface*"[1]. Although the term "earcon" sometimes have been used in a wider sense, it is in this meaning it will be used in this article.

It has been shown that users effectively can remember large hierarchies of earcons, which therefore would make them suitable for navigation cues in telephone-based interfaces [1]. Since there is no natural mapping between each earcon and the phenomenon it represents, this mapping has to be learnt. Earcons are primarily tested in menu metaphor-based interfaces and it is unclear if their benefits are present in other design metaphors such as the room metaphor.

In distinction to earcons, auditory icons are *natural* sounds whose natural associations are used, e.g. the sound of rewinding tape from an answering machine in a voice mail system. One potential advantage with this approach is that it in many contexts does not require a lengthy period of learning. Another benefit is that it is possible to express several different information categories using natural sounds [6]. In the voice mail example, the length of the speedy speech sound could indicate how many messages there are. Lemmens [5] showed that auditory icons have some advantages to earcons in multi modal environments. The use of auditory icons in augmented reality systems has also shown promising results [7]. However, we are not aware of any empirical comparisons between earcons and auditory icons in a unimodal speech and sound-only interface.

2. SYSTEM DESIGN

Using this platform, a voice-enabled application for building maintenance personnel was developed. This is an expanded version of a demonstrator developed by us for ABB. In this application it is possible to check status of fuses, water temperature and ventilation, and also control certain things such as open and close valves, adjust the speed of ventilation fans, enable and disable fuses etc. The application is intended as a quick diagnosing tool when no maintenance man is in place. A "basement" (i.e. room based) metaphor, inspired by natural maintenance environments was developed. These environments tend to be complex, and may in many situations not be intuitively structured (e.g. the room structure may unnecessary complex), which is essential for speech interfaces. Therefore the basement metaphor did not mirror a real environment. Instead it was used to create an optimal structuring of the domain, e.g. the amount of rooms was kept to a minimum, and each room was devoted to a particular function (e.g. fuse boxes and water valves did not exist in the same room).

Currently, there are four different room categories in the application: corridor, water room, electricity room and ventilation room. All categories except the corridor have three subtypes, e.g. fuse room (electricity), section water valves (water) and air-cooling unit (ventilation). Both category rooms, e.g. water room, and subtypes e.g. main water valve room can occur in the application.

When entering a room, the user hears the name of the room, which objects the room contains, and which rooms he or she can go to from the room. If earcons are enabled, the user will hear the earcon of the room before the text-to-speech synthesis says which room he or she entered. If auditory icons are enabled, the user hears the room's characteristic background sound, to support orientation in the same way that a background color can support orientation in a graphic interface. Apart from interacting with rooms and objects by giving commands like: “inactivate fuse a3”, the user can always choose to "examine" at objects to get information about their state. The system was designed to be as consistent as possible, both in terms of room and object descriptions as well as in the choice of available voice commands.

The system was built on Aspect Integrator Platform (AIP), ABB’s software platform for ABB’s Industrial^{IT}, using speechWeb^{TM}, a VoiceXML browser from PipeBeach for dialog handling. AIP is a suitable platform for this purpose, since it has a strong connection to the real world environment – in AIP, objects such as pumps, valves, rooms etc. can be modeled, thus making modeling of a real world maintenance environment possible. On each object can be put a collection of aspects, such as a faceplate, name, alarm and event list etc. The objects are essentially meta-objects; their aspects determine their behavior. All interaction is done through aspects. For this project a voice aspect have been developed which can be added to an object to make it generate VoiceXML code. Since speechWeb cannot handle broadband audio and the auditory icons that were played continuously in the background needed to be in high quality, a special audio application server was developed.

2.1. Auditory icon design

There is an auditory icon for each room. For example, the section valves and the water heater are in different “water” rooms (see figure 1). Therefore both share the same basic background sound, but with different valve or heater sounds mixed together with it.

Category – subtype	**Sound**
Electricity room	A buzzing sound
Main electricity central	Ground noise
Fuse room	Electricity crackle
Electricity meter room	Processed “rotating” ground noise
Water room	Dripping water
Main valve room	Flowing water
Section valve room	Multiple flowing water sounds
Water heater room	Bubbling sound like boiling water
Ventilation room	Flowing air
Air cooling central	Refrigerator noise
Main fan room	Rumbling fan noise
Section ventilation room	Brighter fan noise
Corridor	Reverberated distant noise

Figure 1. *Auditory icons for room categories and types*

The possibility of overlaying more than one sound in a natural way is one of the advantages of auditory icons over earcons. Auditory icons are also used to indicate that the user has performed an action. Furthermore, the auditory icons have been used not only for navigation support and action feedback, but also as indicators for certain states of the system. For example, there is a flooding sound if the user is located in the water main valve room, and the valve is open, if it is closed, a muffled flooding sound is heard.

A problem when designing the auditory icons was that it was hard to find representative sounds for some of the phenomena to be sonificated. However an important issue is that the sounds do not have to be realistic; they need to be representative, the purpose of the sounds is to give the user unambiguous associations to the phenomena the sounds illustrate. This can be experienced in films; film sound designers (Foley artists) often create sounds that exaggerate or caricature the natural phenomena [8]. This strategy was adopted in the design of the sounds used in this study; it was particularly evident in the electricity central, where ground hum was mixed with intense electrical bursts and crackles along with a ticking sound. Examples of the sounds used in the system can be found at http://www.speech.kth.se/~daniels/sound_examples.html.

Although the sound quality of today’s mobile phones is somewhat limited (monophonic, with a sample rate of 8 kHz), the auditory icons that were played in the background had high audio quality (44.1 kHz, stereophonic), which was necessary since many of these sounds relied on high frequency content. This was not unreasonable, considering that the forthcoming 3G standard for mobile computing and telephony will allow multimedia transfers at data rates up to and possibly higher than 2 megabits per second.

The auditory icons used as background sounds consisted of several, sometimes up to 15, sounds mixed together. Some of the sounds were taken from a sound effects library; some were created on a subtractive analog synthesizer. The majority of the sounds however, were authentic sounds recorded using a binaural stereophonic microphone. When possible, authentic sounds were used as basic material for the auditory icons, e.g. the fan sound that was based on a recording of a real fan. The mixing and choice of sounds was inspired by Gaver [4]. The sounds were mixed and thoroughly processed in a multi-track audio application. Some of the sounds were based on the same audio material but were processed using different effect chains. For example, the main fan room and the section ventilation room use the same audio material, but the fan noise in the main fan room was processed using a downward pitch-shifter and flanger effect, giving a lower, darker sound, characteristic of a large fan.

2.2. Earcon design

The earcon design was inspired by Brewster [2]. By manipulating various dimensions of sounds Brewster created a musical language that could be used to represent nodes in tree-structures. Each level of nodes had its own characteristic; on the first level, rhythm was manipulated, on the second level pitch (i.e. a melody based on the rhythm played with a sine wave sound) and on the third timbre (the melody played on e.g. an electric piano). All levels would inherit the characteristics of their ancestors. Thus an earcon representing a node on level 3 would have the same rhythm as its ancestor on level 1, the same melody as its parent (level 2) and a special timbre (level 3 characteristic).

In this study, the room types were represented by earcons. Four rhythmical themes were developed, one for each room category. In [2], the next level of earcons would consist of that rhythm followed by a melody based on that rhythm. However, the pilot studies suggested that a stronger family coherence within each category was needed. In order to improve that, melodies were added to the category themes; e.g. the electricity room earcon consisted of a rhythm followed by a sine-wave melody based on that rhythm. On the next level, the earcon consisted of a rhythm, a sine-wave melody and then the melody played again with a special timbre (e.g. saxophone in the earcon for fuse rooms).

3. USER EVALUATION

The aim of this study was to compare different sound types' effect on user performance and their effects on the users' subjective attitudes toward the application.

3.1. Experimental design

A three conditions (auditory icons, earcons and no sounds) counterbalanced within-groups design was used. Nine participants took part; four females and five males in the ages 23 - 31, all were students of the Royal Institute of Technology in Stockholm. Although the majority was recruited from the Department of Music and Hearing, it was believed that they did not have too much knowledge about speech technology to be representative for real-world users. To confirm this, each subject was interviewed before performing the experiment. During the experiments, the participants wore stereophonic headphones and held a microphone. Since the test domain was not modeling a real building, the subjects did not have any pre-existing knowledge about the rooms.

Before the test, the subjects were interviewed as described previously. The test session consisted of the following phases: Instructions, training and test condition repeated 3 times, and finally another interview.

During the instructions, the subjects were given general instructions about the testing environment, including listening to an example dialog between a subject and the application.

During each training session, the subjects trained on either the sounds and the name of the room types available in the application, or only the name of the room types, depending on the current testing condition. The training environment consisted of a web page showing the room types, see figure 2. Depending on the test condition, the subjects would hear a certain sound associated with the room if he clicked it. The subjects were instructed to memorize the sounds and room names for as long as they needed to understand and learn associations between sounds and rooms. If the current condition was without sounds, the subjects were instructed to familiarize themselves with the room names. In average, the subjects trained approximately for 3 minutes on the associations between sounds and rooms (both in the earcon and auditory icon condition), and for 1.5 minutes in the condition without navigation support sounds.

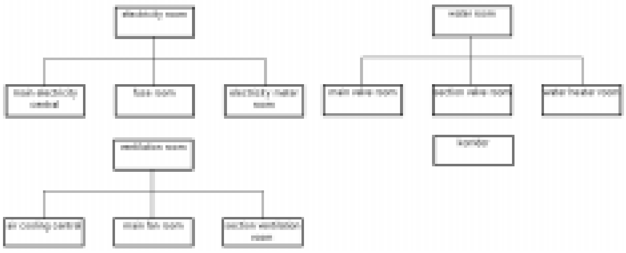

Figure 2. *The training page. The subjects could click each room box to hear the corresponding sound.*

During each test condition, each subject was instructed to navigate through a basement that consisted of a collection of rooms each of which being of one out of the fourteen room types as seen in figure 3. The subject's task was to follow a certain route (given on a separate sheet of paper) and make sure everything was normal in each room. The route description consisted of an ordered collection of codes e.g. A1, B2, C2 etc. Each code corresponded to a certain room in the system. To get to a room the subject needed to say both the name of the room and the code, e.g. "water room A1", just saying "A1" or "water room" was not sufficient. Thus, the subject needed to "look around" in each room, to learn which room to go to next. The subject was told to check the status of a special control panel in each room he or she entered, to make sure everything was normal. If the control panel reported a problem, then a certain action to remedy the problem was recommended by the system. Regardless if the subject succeeded in solving the problem, the test was interrupted and the subject was asked to specify the rooms he or she had passed (the subject did not have to specify the room codes). Then the route would continue, starting in the room where the subject found the problem. In total, the subject would find anomalies and report the rooms he or she passed three times. Then the route was finished and the subject would fill out a questionnaire about his or hers subjective attitudes toward the application. The questionnaire had special questions about the sounds that were to be filled in if the current test condition was one with sounds (earcons or auditory icons). Before each of the remaining two test conditions, the subject would have a new training session matching the current test condition. The remaining test conditions were executed as the first one, the only difference being the sounds used.

After all test conditions, the subject was interviewed about his or her attitudes towards the application and the test.

3.2. Results

Three types of data were collected; data from the subject's room recalling performance, data from the questionnaire and data from the final interview. The results from the recalling performance and the questionnaire were analyzed statistically (although 9 subjects are close to being too few to motivate statistical tests).

The questionnaire consisted of a collection of statements that the subject would rate from –2, "I disagree completely with the statement", to 2, "I fully agree with the statement". In the questionnaire, significant differences were found among the specific sound related questions. These results are summarized in table 1.

Question	Avg. earcon score	Avg. auditory icon score	t-test, two tailed)
The sounds were annoying	0.33	-0.89	p = 0.038
the sounds made the system easier to use	-0.56	0.78	p = 0.002
the sounds made it easier for me to remember where I was	-1	1.11	p = 0.006

Table 1: *The results from the questionnaire were significant differences between the earcons and auditory icons were found*

The subjects' room recalling performances were corrected along the following rules: A subject would not get punished for omissions or insertions. For each correct room the subject got 2 points, for the correct family (water, electricity or ventilation) but not the correct subtype, the subject got 1 point. The condition without sounds had an average score of 23.78, the earcon condition 21.44 and the auditory icon condition 25.11. Given the few subjects in the study, and hence a low statistical power, it is interesting to note that the ANOVA almost reaches significance (F(2,8) = 2,89, p = 0.085).

The final interview showed that none of the subjects preferred the earcon condition, 6 subjects explicitly said they preferred the auditory icon condition, 1 subject preferred the no sound condition, 2 were uncertain. Without being asked, 3 subjects said that earcons was the worst condition. Furthermore, the interview showed that 8 subjects preferred room-based metaphors before menu-based metaphors in systems similar to this one, 1 preferred menus. The attitudes toward the degree of trust of a high security system, e.g. a banking system, using a room metaphor and auditory icons were mixed. 3 subjects believed they would take such a system seriously, 3 would not take such a system seriously, 2 were uncertain and 1 subject could not answer the question. On the question that followed; if such a design would cause them not to use it, 3 subjects reported it would not, 2 that it would, 2 were uncertain and 2 subjects could not answer that question.

4. DISCUSSION

Since the room metaphor is not widely used, it was particularly important to ensure that the subjects were comfortable with the system design. If not, the results from the evaluation may have been irrelevant, since the auditory icons were strongly related to the room metaphor. The final interview confirmed that 8 out of 9 subjects would prefer a room-based metaphor in a similar system (i.e. a system that can be effectively modeled using a room-based metaphor) to a menu-metaphor based system.

The lack of significance in the recall task evaluation is not surprising, considering the low number of subjects and hence a low statistic power. An interesting note in the recall task is that one of the subjects reported that he used a special memorizing strategy of pretending he was checking rooms in his grandparent’s basement (the so called *method of loci*). If the user could customize the interface and the auditory icons to some extent, these kinds of mnemonic devices could help navigation even more.

One somewhat surprising fact that came out is the subjects’ negative attitudes toward earcons. One reason for this is most likely lack of training. Note however that letting the users decide for themselves how much time they would like to spend appears as an ecologically valid assumption. Furthermore, the training was successful to some degree: The majority of the subjects (6) claimed that they would remember the sounds of each category (electricity, water, ventilation, and corridor) fairly well, albeit not the exact room type, after their earcon training session. However, the difficulties in remembering the exact room type from the sounds were present in the auditory icon condition as well (8 subjects reported this problem), but all subjects claimed they could easily remember the categories from a particular sound. Thus it seems the subjects had comparable knowledge about the sound-room associations in the earcon and auditory icon condition. Still subjects rate earcons significantly more negative. We believe this effect is likely to occur since the earcons are artificial and foreign to the domain and room-based metaphor.

The question whether sound should be used at all remains. But since only 1 subject preferred the no sound condition (an interesting note is that this subject still said that the auditory icons were helpful to some extent), we believe auditory icons are to be preferred to a design without sounds.

5. SUMMARY AND FUTURE WORK

The described system has, in our opinion, the following characteristics and advantages compared with earcon-based systems: No (or little) need of user training; differentiates naturally between place and state-change or action sounds; better possibility for presenting many simultaneous or overlaying sounds without making it difficult to interpret for the user.

The advantage of auditory icons is supported further by our results from the evaluation; the subjects preferred auditory icons before earcons. Furthermore the final interviews suggest that using auditory icons is better than not using any sounds at all. Therefore we believe that the design philosophy used behind the auditory icons used in this study is a good choice when designing room-metaphor based systems.

Strategies using auditory icons to facilitate and encourage the user’s use of mnemonic devices should be further investigated.

The results from this study further suggest that the issue of trust and auditory icon design in high security systems should be addressed in future research.

We are aware that auditory icons cannot be used in all domains and for all applications where earcons have been or could be used. But we believe that our present work suggests that where they can be used, they have an advantage over earcons. Furthermore, it is possible that by taking inspiration from Foley art and taking some artistic liberty of inventing new sounds may extend their possible range of application. But this is something that requires further research to clarify.

6. ACKNOWLEDGMENTS

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

The support from ABB and PipeBeach is gratefully acknowledged.

7. REFERENCES

[1] Brewster, S.A. (1997). Navigating telephone-based interfaces with earcons. In *Proceedings of HCI'97*.

[2] Brewster, S.A. (1994) Providing a structured method for integrating non-speech audio into human-computer interfaces. PhD Thesis, University of York, UK

[3] Dutton, R.T., Foster, J.C. & Jack, M.A. (1999) Please mind the doors – do interface metaphors improve the usability of voice response services? *BT Technology Journal*, 17 (1), 172-177.

[4] Gaver, W.W. (1993). What in the World do we Hear?: An Ecological Approach to Auditory Event Perception.. *Ecological Psychology*, 5(l), 1-29.

[5] Lemmens, P.M.C, Bussemakers, M.P. & de Haan, A. (2001) Effects of Auditory Icons and Earcons on Visual Categoriztion: The Bigger Picture. *Proceedings ICAD2001*.

[6] Macaulay, C., Benyon, D., and Crerar, A. (1998) Voices in the Forest: Sounds, Soundscapes and Interface Design, In Dahlbäck, N. (ed.) *Exploring Navigation*. SICS Technical Report T98:01, SICS, Stockholm.

[7] Mynatt, E.D., Back, M., Want, R. Baer, M., & Ellis J.B. (1998) "Designing Audio Aura". *Proceedings of CHI '98.*

[8] Somers, E. (2000) Abstract Sound Objects to Expand the Vocabulary of Sound Design for Visual and Theatrical Media. *Proceedings of ICAD2000*.

[9] Vodafone official www site, visited 2003-02-02. http://www.vodafone.se/150_1.jsp?service_id=2197