MULTIMODAL SCANNING INPUT WITH NON-SPEECH SOUNDS

Stephen A. Brewster[1]*, Veli-Pekka Raty*[2] *& Atte Kortekangas*[3]

[1]SINTEF DELAB,
N-7034 Trondheim,
Norway.
Tel: +47 73 59 42 52
stephen.brewster@delab.sintef.no

[2]VTT Information Technology,
Kanslerinkatu 12 B,
Tampere, FIN-33101,
Finland.
Tel: +358 31 316 3328
veli-pekka.raty@vtt.fi

[3]VTT Information Technology,
Tekniikantie 4B,
Espoo, FIN-02044 VTT,
Finland.
Tel: +358 0 456 4311
atte.kortekangas@vtt.fi

ABSTRACT

This paper proposes the addition of non-speech sounds to aid people who use scanning as their method of input. Scanning input is a temporal task; users have to press a switch when a cursor is over the required target. However, it is usually presented as a spatial task with the items to be scanned laid-out in a grid. Research has shown that for temporal tasks the auditory modality is often better than the visual. This paper investigates this by adding non-speech sound to a visual scanning system. It also shows how our natural abilities to perceive rhythms can be supported so that they can be used to aid the scanning process. Structured audio messages called *Earcons* were used for the sound output. The results from a preliminary investigation were favourable, indicating that the idea is feasible and further research should be undertaken.

KEYWORDS

Non-speech sound, earcons, scanning input, multimodal interaction.

INTRODUCTION

Scanning input is employed by users who cannot operate a standard mouse, often because of a physical disability, and are only able to use a single switch for selecting items. In this technique, a set of items is displayed on-screen and a cursor moves slowly between them. The user presses a switch when the required item is reached. This method of input is slow. In this paper we propose the use of non-speech sound to improve the scanning process. We aim to show that adding sound is possible and that is has benefits for the user.

There is a growing body of research which indicates that the addition of non-speech sounds to human-computer interfaces can improve performance and increase usability [2, 4, 10]. Non-speech sound is an important means of communication in the everyday world around us and the benefits it offers should be taken advantage of at the interface. Such multimodal interfaces allow a greater and more natural communication between the computer and the user. They also allow the user to employ the appropriate sensory modalities to solve a problem, rather than just using one modality (usually vision) to solve all problems.

This work is part of the TIDE ACCESS Project 1001. The aim of this project is to create a mobile communication device for speech-motor and/or language-cognitive impaired users. People will use the device to create messages they want to communicate and then play those messages via synthetic speech. Such users often utilise pictographic languages (for example, Bliss [1]) to communicate. The pictures represent words or actions and can be combined to create complex messages. Users must be able to interact with the system as fast as possible so that they can communicate effectively. In this paper we investigate the use of non-speech sound to facilitate communication.

Scanning

Figure 1 and Figure 2 demonstrate scanning input. Figure 1 shows a layout of sixteen pictographic symbols, arranged as a four by four grid. In this example each row is highlighted in turn with the system moving on to the next row after a fixed amount of time (if no selection is made). The figure shows the second row highlighted. If no selection is made then the highlight moves on to the next row and so on through all of the rows and then back to the first row again.

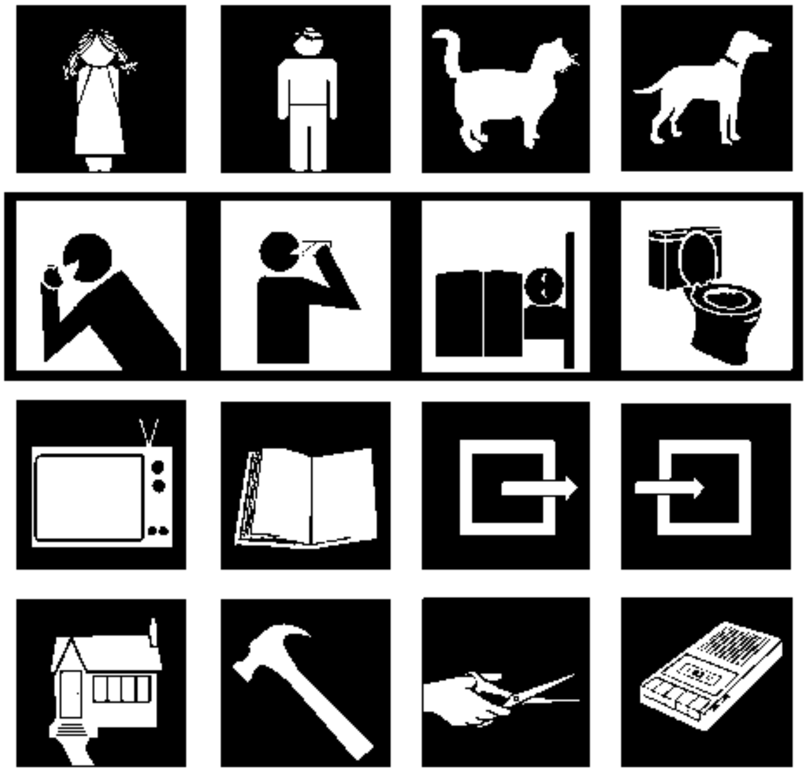

Figure 1*: Scanning a four by four grid of pictographic symbols. The second row is highlighted.*

If the user presses the switch whilst the highlight is on row two then scanning item by item begins. The highlight now moves from one item to the next on the same row. Again, the highlight moves on if, after a predefined period of time, the switch has not been pressed. Figure 2 shows the second item highlighted. If the user now presses the switch the second item will be chosen. If the switch is not pressed the highlight will move on to the next item in the row and continue until the end of the row is reached. It then returns to the first item of the row once again.

This method allows any item to be selected but can take a long time. The highlight delay, the length of time the highlight stays on one item, may be as much as five seconds. The time is necessary so that users have time to decide that it is the item required and then to press the switch. With certain physical impairments, moving the hand to the switch and then pressing it can take a long time.

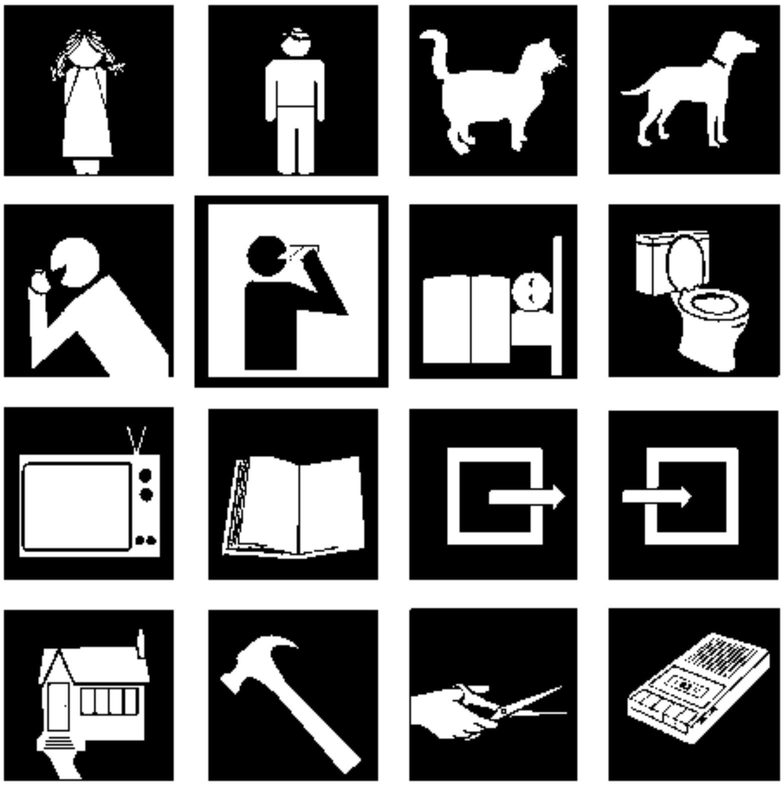

Figure 2*: The second row of the grid was selected and now item by item scanning is taking place. The second item on row two is highlighted.*

The items to be scanned are laid out spatially but the selection task is not a spatial but a temporal one: The user must press the switch at the correct *time* to choose the required item. Walker & Scott [14] suggest that the auditory modality is a more appropriate means of processing information in the temporal domain whereas the visual modality is better in the spatial domain. Wenzel agrees saying ([15], p 137): "Another advantage of audition is that it is primarily a temporal sense and we are extremely sensitive to changes in an acoustic signal over time". Gaver [9] suggests that sound continues over time and is therefore good for the display of changing events, such as those that occur with the changing highlight when scanning. Therefore a graphic highlight alone may not be the best method of guiding scanning selection. It may be more effective to combine sound and graphics to cue the user when to press the switch.

There is rhythm inherent in the scanning process. The highlight moving from one row to the next is like a beat and the end of the bar comes when the last row is reached. The rhythm begins again when the highlight goes back to the first row. As Fraisse [8] says the perception of rhythm is a basic but powerful human ability and is present even in very young children. It is an extremely important aspect of music and humans are very accurate at timing sound pulses [8]. An aim of the research described here is to allow users of scanning input systems to user their basic human abilities to improve and speed-up the scanning process. If users can follow the rhythm of scanning then it may help them select items more quickly because they will be able to predict when the highlight will move to the next item.

There can be a problem when scanning because some users need very long scanning delays, as mentioned above. Humans can normally perceive a rhythm when there is less than an 1800 msec. gap between the repeating units [8]. When the inter-unit gap is longer than this the stimuli are no longer perceptually linked. Listeners do not perceive the units as being part of a rhythmic whole but as separate and disconnected. With a long scanning delay, of perhaps five seconds, users cannot use their natural rhythmic abilities to cue them as to when to press the switch to choose an item. Part of the work described here will use non-speech sound to overcome this problem.

Earcons

The non-speech sounds used for this investigation were based around structured audio messages called *Earcons* [3, 4, 13]. Earcons are abstract, synthetic tones that can be used in structured combinations to create sound messages to represent parts of a human-computer interface. Detailed investigations of earcons by Brewster, Wright & Edwards [5, 6] showed that they are an effective means of communicating information in sound.

Earcons are constructed from *motives*. These are short, rhythmic sequences that can be combined in different ways. The simplest method of combination is concatenation to produce *compound earcons*. By using more complex manipulations of the parameters of sound (timbre, register, intensity, pitch and rhythm) *hierarchical earcons* can be created [3]. Using these techniques structured combinations of sounds can be created and varied in consistent ways. The earcons described here were created using the guidelines proposed by Brewster *et al.* [7].

SONICALLY-ENHANCED SCANNING

We will now propose a method of using non-speech sounds to improve the scanning process. Earcons were added to the standard graphical scanning process (described above) to improve usability. In our investigation the scanning process could be varied in two ways. The scanning delay could be changed (from one to five seconds) and the size of the grid could be changed (from 2x2 to 4x4). Scanning could be row by row or column by column. The sounds will be described here in terms of rows, for columns the sound for the top row was on the left and the sound for the bottom row on the right.

With large scanning delays and/or small grids problems of rhythm perception can arise because there can be delays of greater than 1800 msec. between the items. To avoid this we played repeating notes which were reduced in intensity to fill up the time gaps.

The system was based on a HyperCard stack that ran on an Apple Macintosh. The sounds were played on a Yamaha TG100 multi-timbral sound module controlled by MIDI. The sounds were presented via loudspeakers.

Row scanning

When scanning a four by four grid of symbols (as in Figure 1 above) one row at a time, each row was given a base octave. The earcons were at a high pitch at the top of the screen and low at the bottom. The top row was in the octave of C_1 (1056Hz), second row C_2 (523Hz), third row C_3 (261Hz) and the bottom row was at C_4 (130Hz). For each row the notes C, E, G, B were played (in the octave of that row), each for 100 msec. as an *arpeggio* (see Figure 3). There were four notes corresponding to the 4 items in the row. In order to make the sounds for each row into rhythmic units, the first item was accented (played slightly louder than the rest) and the last item was slightly lengthened [11]. The notes of the first row were played at a lower intensity to equalise the loudness across the rows (the high-pitched notes of the first row were perceived as louder than the lower-pitched notes of the other rows [4], Chapter 2).

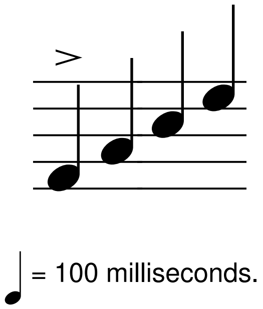

Figure 3*: The arpeggio used to indicate a row when scanning a 4x4 grid.*

The system's scanning delay could be varied to suit the needs of the user of the system. Users with good motor control could use a scanning delay of one second. This meant that the user had one second in which to press the switch to choose a row. For this second the row was visually highlighted and the earcon played. The scanning delay could be increased to a maximum of five seconds per row. The sounds just described were for a scanning rate of one second. As the scanning delay increased the earcons were repeated. For example, when the scanning delay was set to three seconds, the earcon for one second was repeated three times. The volume was reduced for each repeat so the earcon sounded like it was fading away. This helped with the problem of rhythm perception (mentioned above) where humans cannot recognise groups of units as a rhythm if there is too large a gap between the units. By repeating the groups we made sure that there was no large gap between the sounds and listeners could use their natural rhythmic abilities to cue them when to press the switch to choose the item required.

The earcons changed if the grid-size changed. For example, when scanning a two by two grid an *arpeggio* of two notes instead of four was played. The top row was at pitch C_1, the bottom row at C_2. These notes were again repeated if the scanning delay was increased. With a scanning delay of five seconds and a grid size of two by two, there was more than 1800 msec. gap between the two sounds. In this case, the earcons were repeated three times to avoid the rhythm perception problem.

Item scanning

When a row was chosen on a four by four grid, item scanning started and the individual notes of the *arpeggio* from above were played (C, E, G, B) in the octave of that row. The first note of the row (C) was accented. For a scanning delay of one or two seconds, each note was played twice. The first time for 300 msec. and then for 100 msec. (with a 100 msec. gap between them). The second note was played more quietly to give the impression that the sound was fading away (as for row scanning). The next note was played when the next item was highlighted. The last note of the four (B) was played only once but for 500 msec. This again helped with grouping into rhythmic units (as the last note of the group was lengthened [11]).

For a scanning delay of three or four seconds a sequence of four notes was played for items one to three in the row and for the last item two long notes were played. For a scanning delay of five seconds five notes were played. This allowed us to avoid the long gaps that would have affected rhythm perception.

When item scanning on a two by two grid with a scanning delay of more than two seconds longer notes were needed so that there were no rhythm perception problems. With a delay of three or four seconds four notes were played, with 200 msec. between each note (see Figure 4) and with a delay of five seconds five notes were played. The notes decreased in volume to give the impression the sound was fading away.

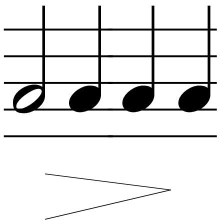

***Figure 4**: Item scanning on a 2x2 grid with a scanning delay of three or four seconds required four notes to avoid rhythm perception problems.*

We have shown that sound can be added to the scanning process to utilise the basic human ability of rhythm perception. This would allow the user of such a sonically-enhanced scanning system to predict when the highlight would move on to the next item to be scanned. The user would be able to ‘feel’ the rhythm of the scanning and perhaps allow shorter scanning delays to be used. It would also reduce problems due to long scanning delays when it can be difficult to predict when the highlight will change because the gap between changes is longer than the human ability to group the units into a rhythmic whole.

EVALUATION

To investigate the addition of sound a simple evaluation of the scanning process was conducted. We had access to one ten year old child at Folkhälsan (a children’s hospital) in Helsinki who used scanning as her normal form of input. She needed a scanning delay of five seconds to be able to select the items she wanted.

She was trained on the standard scanning program (without sound) by performing various selection tasks. When she could use this program and move around and select items as necessary, she was given the sonically-enhanced scanning program. Unfortunately, due to time constraints, we only had the chance to test the enhanced system for one afternoon. The nature of the evaluation was informal; it was not possible to persuade our user to do exactly the tasks we wanted, she was more interested in playing with the system and having fun! We therefore observed her using the system in general and doing certain selection tasks requested by her speech therapist. Subjectively, she seemed to enjoy using the sonically-enhanced scanning system and there was no obvious decrease in performance due to the addition of sound. She appeared to feel that, with sound, the system was more like a game and so enjoyed playing with it more. She was also more engaged with the system than without sound [12]. Her speech therapist indicated that she seemed to be performing better, although we were not able to measure any quantitative increase as we could not do any formal analysis due to a lack of time. However, these favourable preliminary results indicate that there may be advantages to be gained from sonically-enhanced scanning and therefore further investigations should be undertaken.

FUTURE WORK

This work has shown that it is possible to add sound to the scanning process. The preliminary results indicated that there were some benefits to be gained from adding sound. Therefore, these sounds will be used as described here in the communicator device for TIDE ACCESS Project 1001. The next stage of our work will be to conduct a more formal analysis. This will involve asking participants to perform certain tasks with and without sound and measuring the differences in time taken to select items. From this we will then be able to suggest some guidelines for designers of scanning systems wishing to incorporate sound.

CONCLUSIONS

Scanning input is used by many disabled users to select items on a display. The items to be scanned are laid out on a grid and presented as a spatial task. However, it is really a temporal task: The user must press the switch at the right time to select the required item. We supported this by adding non-speech sound which can be more effective with temporal tasks. This would allow users to ‘feel’ the rhythm and use their powerful natural abilities of rhythm perception to help predict when to press the switch to choose the item required. We also wanted to make sure that, when very long scanning delays were needed, users would still be able to utilise their rhythmic abilities to help them select more quickly. An initial evaluation showed that users may prefer sound and that it can help them perform better.

ACKNOWLEDGEMENTS

Thanks go to our participant and to Gitta Lönnqvist, her speech therapist at Folkhälsan in Helsinki for help and assistance with the evaluation. This research was supported by ERCIM Fellowship 94-04. It was undertaken whilst the first author was at VTT Information Technology in Finland.

BIBLIOGRAPHY

1. Baumgart, D., Johnson, J. and Helmstetter, E. *Augmentative and Alternative Communication Systems for Persons with Moderate and Severe Disabilities*. Paul Brookes Publishing Co., Baltimore, Maryland, 1990.

2. Blattner, M., Papp, A. and Glinert, E. Sonic enhancements of two-dimensional graphic displays. In *Auditory Display, sonification, audification and auditory interfaces. The Proceedings of the First International Conference on Auditory Display* (Santa Fe Institute, Santa Fe) Addison-Wesley, 1992, pp. 447-470.

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

4. Brewster, S.A. *Providing a structured method for integrating non-speech audio into human-computer interfaces*. PhD Thesis, University of York, UK, 1994.

5. Brewster, S.A., Wright, P.C. and Edwards, A.D.N. A detailed investigation into the effectiveness of earcons. In *Auditory display, sonification, audification and auditory interfaces. The Proceedings of the First International Conference on Auditory Display* (Santa Fe Institute, Santa Fe) Addison-Wesley, 1992, pp. 471-498.

6. Brewster, S.A., Wright, P.C. and Edwards, A.D.N. An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of INTERCHI'93* (Amsterdam) ACM Press, Addison-Wesley, 1993, pp. 222-227.

7. Brewster, S.A., Wright, P.C. and Edwards, A.D.N. Experimentally derived guidelines for the creation of earcons. In *Adjunct proceedings of HCI'95* (Huddersfield, UK), 1995.

8. Fraisse, P. Rhythm and tempo. In *The psychology of music.*, Deutsch, D. (Ed.), Academic Press, San Diego, CA., 1982, 149-180.

9. Gaver, W. The SonicFinder: An interface that uses auditory icons. *Human Computer Interaction* 4, 1 (1989), 67-94.

10. Gaver, W., Smith, R. and O'Shea, T. Effective sounds in complex systems: The ARKola simulation. In *Proceedings of CHI'91* (New Orleans) ACM Press, Addison-Wesley, 1991, pp. 85-90.

11. Handel, S. *Listening: An introduction to the perception of auditory events*. MIT Press, Cambridge, Massachusetts, 1989.

12. Kramer, G. An introduction to auditory display. In *Auditory Display, sonification, audification and auditory interfaces. The Proceedings of the First International Conference on Auditory Display.* (Santa Fe Institute, Santa Fe) Addison-Wesley, 1992, pp. 1-77.

13. Sumikawa, D., Blattner, M., Joy, K. and Greenberg, R. *Guidelines for the syntactic design of audio cues in computer interfaces*. Lawrence Livermore National Laboratory, 1986.

14. Walker, J.T. and Scott, K.J. Auditory-visual conflicts in the perceived duration of lights, tones and gaps. *Journal of Experimental Psychology: Human Perception and Performance* 7, 6 (1981), 1327-1339.

15. Wenzel, E.M. Spatial sound and sonification. In *Auditory Display, sonification, audification and auditory interfaces. The Proceedings of the First International Conference on Auditory Display* (Santa Fe Institute, Santa Fe) Addison-Wesley, 1992, pp. 127-150.