AN INVESTIGATION INTO THE IDENTIFICATION OF CONCURRENTLY PRESENTED EARCONS

David K McGookin

GIST
Department of Computing Science
University of Glasgow
Glasgow, Scotland G12 8QQ
mcgookdk@dcs.gla.ac.uk
www.dcs.gla.ac.uk/~mcgookdk

Stephen A Brewster

GIST
Department of Computing Science
University of Glasgow
Glasgow, Scotland G12 8QQ
stephen@dcs.gla.ac.uk
www.dcs.gla.ac.uk/~stephen

ABSTRACT

In this paper we describe an experiment investigating the ability of participants to identify multiple, concurrently playing structured sounds, called earcons. Several different sets of earcons were compared, one "state of the art" set based on the guidelines of Brewster [1], and other sets of earcons modified to take account of auditory scene analysis principles. The effect of the number of concurrently playing earcons on identification was also investigated, with instances of 1, 2, 3 and 4 concurrently playing earcons tested. Overall, performance was low, with less than two earcons being successfully identified in any condition. However it was found that both staggering the onset times of each earcon, as well as presenting each earcon with a unique timbre, had a significantly positive effect on identification.

1. INTRODUCTION

Mobile computing devices are becoming increasingly more popular, with greater functionality constantly being added by manufacturers. The usability of these devices is, however, open to debate, as mobile computing has several usability issues that need to be addressed. Notably, due to the form factor of mobile devices, the visual display space available is severely limited in comparison to other computing devices, such as the personal computer. For example, the Palm Tungsten personal digital assistant (PDA) has a display of only 6x6 cm. The low resolutions of such displays also contribute to limiting the amount of data that can be usably presented. Also, because mobile computer users are likely to be on the move whilst using their device, they cannot devote their entire attention to the computing task. They must constantly monitor the environment for danger and react accordingly. This places further strain on the visual sense.

One of the potential ways with which these issues can be overcome is in the use of audio feedback to the user. Brewster [2] showed that the addition of simple sounds to a PDA interface allowed for the reduction in size of visual buttons, whilst still leaving the interface usable. Other systems, such as Sawhney and Schmandt's Nomadic Radio [3] have been able to go further and totally remove the visual interface.

Because of the usefulness of audio with mobile devices and the limitations on mobile device display resources, developers may wish to push more information into audio. This means that it is possible that multiple items of audio information may be concurrently presented to the user. It is important therefore that developers understand the interactions that will occur when two concurrent items of audio feedback are presented together, and how they can design audio messages to avoid interfering with each other. The work presented here examines how to reduce such interactions when structured audio messages called earcons [4] are concurrently presented. As will be explained later, earcons are an interesting case, as they are inherently more susceptible to interference from each other than other auditory feedback, such as auditory icons [5] or speech.

2. AUDITORY SCENE ANALYSIS

Auditory Scene Analysis (ASA) [6], is the study of how the multiple, complicated waveforms that are detected by our auditory system are separated into meaningful representations, e.g. how a mobile telephone ring is separated from a symphony orchestra. Auditory Scene Analysis has been heavily studied, and is based on gestalts [7]. It shows that the greater the similarity between two auditory sources, along a number of dimensions, such as similarity (similar timbres, similar frequency etc.), familiarity, common fate etc., the more likely it is that they will be perceived as one composite stream.

3. EARCONS

Earcons are short structured abstract audio messages which can be effectively used to communicate information in a computer interface [4]. There are four main types of earcon, one-element earcons, compound earcons, hierarchical earcons and transformed earcons. One-element earcons are the simplest type and can be used to communicate only 1 bit of information, for example a sound used to indicate a "save" operation had occurred. Compound earcons are more extensible than the one-element type. Here one-element type earcons can be concatenated together to create more meaningful messages. For example, a one-element "save" earcon and a one-element "file" earcon can be played after each other to represent the "save file" operation. The hierarchical and transformational earcon types are the most flexible and are constructed around a "grammar". In the hierarchical type of earcon, each auditory parameter (generally, timbre, rhythm, pitch and register) of the earcon is manipulated to

provide more detailed information about what it represents. For example, a rhythm may represent an error, the pitch of that rhythm the type of error etc. Brewster [8] has performed extensive studies of the usefulness of both compound and hierarchical earcon types and has suggested guidelines [1] on how auditory parameters should be employed to produce usable earcons. These guidelines are summarised in Table 1.

Attribute	Guideline
Timbre	Musical Timbres that are subjectively easy to tell apart should be used for earcons.
Register	If absolute judgments are required then register should not be used. If relative judgments of register are to be made then there should be gross differences between the registers used.
Pitch	Introducing complex intra-earcon pitch structures can be effective when used with another attribute such as rhythm.
Rhythm	Putting different numbers of notes in each rhythm is an effective way of differentiating them. Brewster [1] also notes that changes to tempo are useful in differentiating earcons.

Table 1. *Guidelines for the construction of effective earcons* [1].

Almost all of Brewster's work however, has looked at cases where only one earcon is presented at a time. Because the most powerful earcon types, hierarchical and transformational, are constructed from a grammar, the members of the sets of earcons produced are very similar. For example, several will have the same rhythm or be played in the same register. Because ASA states that the greater the differences between two sounds are, the more likely it is that they will be perceived as being two sounds (instead of one), we can conclude that simultaneously playing earcons from the same set are likely to interfere with each other. In order to avoid this, it is not possible to arbitraly make the members of a set of earcons different, as this will destroy the "grammar" that makes earcons powerful communicating sounds. It is important therefore, if multiple earcons from the same set are likely to be concurrently played, that designers know both how many earcons a user can concurrently attend to, as well as how the interactions between individual earcons can be minimised.

4. RELATED WORK

The issues of identification of multiple concurrent audio sources have long been known about. Papp [9], noted that *"At worst, the entire sound presentation will be an auditory "smearing" of each individual source, and of no informational value to the user"*. He proposed an audio server which would apply ASA rules to select the most appropriate form of auditory feedback to the user. However he did not perform an evaluation of his system. McGookin and Brewster [10], also identified these problems with their map navigation system. Here multiple spatialised earcons interfered with each other, such that individual earcons could not be identified.

There has been little work investigating the topic of concurrently playing structured audio. Gerth [11], performed several experiments on identification of a limited number of synthetic timbres. Brungart, Ericson and Simpson [12] looked at improving the identification of concurrently presented speech in aircraft cockpits. They found that having different talkers for each spoken text significantly improved identification.

5. EXPERIMENTAL OVERVIEW

In order to answer the questions posed at the end of Section 3, an experiment was designed to identify how both varying the number of concurrently playing earcons, as well as redesigning those earcons to take into account ASA principles, affected recognition. The experiment was of a between groups design and involved 16 participants per condition identifying simultaneously playing earcons in a common spatial location. Although spatial location is an important factor in ASA, we decided not to include it in this study because many mobile devices do not have good quality spatial positioning ability. It may also be inconvenient for the user to wear headphones to use spatial feedback, e.g. it is unlikely that a user would wear headphones to specifically interact with a mobile telephone menu. Also, even when using spatial positioning, it is difficult to know how far apart sound sources would need to be in order for them to be identifiable and distinct. In real world scenarios it may not always be possible, even when using spatial positioning, to keep important audio objects apart (for example, cartographic data).

Ride Parameter	Description
Type	This parameter defines the type of the ride. There are three possible ride types. We have taken care to ensure that we choose obviously different instruments. We use a trumpet to represent a rollercoaster, a banjo to represent a water ride and a piano to represent a static ride.
Intensity	This parameter defines how intense the ride is and is mapped to a rhythm with a complex pitch structure. Three distinct combinations were used to represent low, medium and high intensity rides. In accordance with the guidelines of Brewster *et al.* [1], we used a varying number of notes to help differentiate the rhythms, with 2, 4 and 6 notes used respectively for low, medium and high intensities
Cost	This defines how much it would cost to go on the ride. This attribute was mapped to register, with a higher register representing a greater cost. As absolute pitch perception is difficult for most people, we ensured that there was a gross difference (at least an octave) between the registers used.

Table 2. *The "grammar" used to construct earcons.*

In this experiment we looked at the worst-case scenario where earcons were given the same spatial location. We do not argue that spatial position is unimportant, rather that there are advantages in studying it in isolation to other factors. The earcons used in the experiment were the same as those used in the Dolphin

system [10]. These were based around a variation of the transformational earcon type, and represented rides that may be found in a theme/amusement park. The "grammar" used to construct the earcons is given in Table 2. The earcons produced from the grammar provided a "state of the art" set with which to compare ASA modifications.

5.1. Conditions

Overall there were nine conditions in the experiment. Before each, participants were trained such that they could identify 3 individually presented earcons without help. In each condition participants heard 4 concurrently presented earcons, repeat 7 times. They attempted to identify these earcons and record their choices in a clickable list in a computer interface. There were twenty sets of stimuli for each condition. The conditions are described below.

5.1.1. Original Earcon Set Condition

In this condition participants performed the previously outlined experiment with the earcons formed from the grammar in Table 2. The results of this condition were treated as a baseline with which to measure the other conditions.

5.1.2. Three Earcon Condition

Three earcons were simultaneously presented instead of four.

5.1.3. Two Earcon Condition

Two earcons were simultaneously presented instead of four.

5.1.4. One Earcon Condition

Here only one earcon was presented at a time. This condition was used to identify the "quality" of the earcon set used.

5.1.5. Melody Altered Earcon Set Condition

Here the earcons used were based on those described in Table 2. However the pitch/rhythm combinations were altered, such that each melody "glided" in one direction. I.e. one melody continually rose in pitch, one melody continuously fell in pitch and another kept the same pitch. The objective was to attempt to take advantage of the common fate principle of auditory scene analysis. It has been noted by some that tone sequences composed in this way may promote better streaming [6].

5.1.6. Multi-Timbre Earcon Set Condition

Although there is no universal definition of timbre [5], several researchers have shown that known elements of timbre can influence how sequences of sounds are perceived. Also as shown by Brungart, Ericson and Simpson [12], described in section 4, having different voices speaking similar texts (effectively modifying the timbre of the speaker) had a significant improvement on recognition.

In this condition, whenever two rides of the same type were presented simultaneously, each was presented with a different instrument from the same instrument group. Hence if two rollercoasters were presented simultaneously, instead of both being presented with the same piano timbre, one would use an acoustic grand piano timbre, the other would use an electric grand piano timbre. It was hoped that this would allow earcons representing the same ride type to sound different enough so that they could stream separately. The instrument groupings used were based on those of Rigas [13].

5.1.7. Extended Training Condition

Although all of the participants were trained to identify individually presented earcons before performing the experiment, they were not given specific training on how to listen to concurrently presented earcons. In this condition participants were given a tool where they could listen to a specific combination of four earcons, which were not used in the experiment, and switch on and off individual earcons in order to understand the impact of adding or removing individual earcons on the composite sound.

5.1.8. Staggered Onset Condition

Here instead of all four earcons being simultaneously presented, there was a 300ms onset to onset delay between the starts of each individual earcon. ASA research indicates that sounds which start at the same time tend to be related causing them to stream together [6].

5.1.9. Final Condition

In this condition all of the previous modifications that preliminary analysis had shown to be effective, were combined to measure the overall improvement in recognition. This condition combined the staggered onset features and the multi-timbre features.

6. RESULTS

For each of the conditions, the number of correctly identified earcons and the number of correctly identified earcon parameters (number of correctly identified ride types, ride intensities and ride costs) were collected.

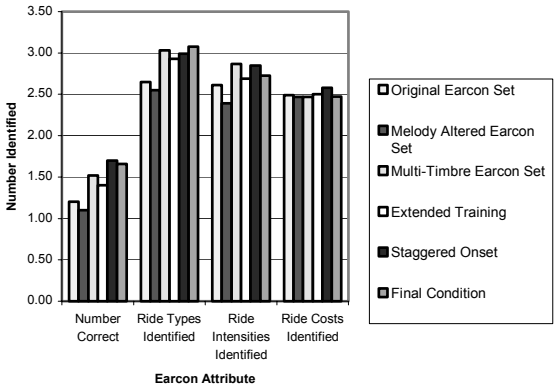

Figure 1. *Summary of correctly identified parameters.*

This data, excluding the results for the one, two and three eareon conditions, is summarised in Figure 1. In order to determine if any of the results were statistically significant, four one-way analysis of variance (ANOVA) tests were performed, one for each parameter (number of eareons identified, types identified, intensities identified and costs identified). The ANOVA for the correctly identified number of eareons was significant ($F(5,90)=7.12$, $p<0.001$). *Post hoc* Tukey tests showed that the staggered onset condition had significantly better identification than the original eareon set condition, as did the final condition. For the number of ride types identified the ANOVA also showed significance ($F(5,90)=7.84$, $p<0.001$). *Post hoc* Tukey tests showed that the multi-timbre eareon condition, the staggered onset condition and the final condition were significantly better identified than the original eareon set condition.

For the intensities of rides identified, the ANOVA again showed significance ($F(5,90)=3.16$, $p=0.011$). Post hoc Tukey tests showed that the multi-timbre eareon set was significantly better identified than the melody altered eareon set as was the final condition. The final condition did not perform significantly better than either the multi-timbre eareon set condition or the staggered onset condition in any of the ANOVAs. The ANOVA for ride cost was not significant ($F(5,90)=0.31$, $p=0.907$).

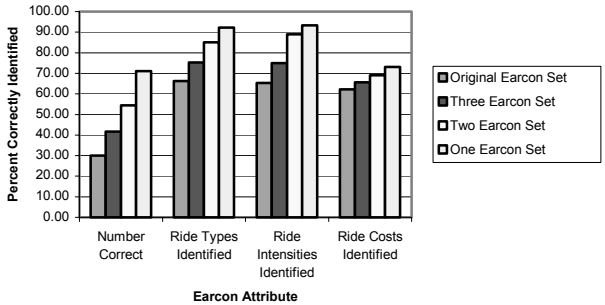

Figure 2. *Percentage of attributes correctly identified for the 1,2,3 and original eareon sets.*

Because of the design of the experiment, it is not meaningful to directly compare the numeric results for the one, two, three and original set conditions. Figure 2 therefore, shows the percentage of correctly identified attributes for these conditions. ANOVAs were performed on the percentage correct for each condition. As with the previous ANOVAs, the results for the number correct ($F(3,60)=23.28$, $p<0.001$), ride types identified ($F(3,60)=23.28$, $p<0.001$), and ride intensities identified ($F(3,60)=31.16$, $p<0.001$) were significant. The ANOVA for ride costs identified was not significant ($F(3,60)=2.24$, $p=0.093$). Post hoc Tukey tests agree with Figure 2 and generally show that the one eareon condition is better than the two eareon condition and so forth.

7. DISCUSSION

From the results given in Figure 1, it is clear that the identification of multiple concurrently playing eareons is difficult, with no more than two eareons on average being correctly identified in any condition. We can conclude that the eareon set used is of high quality since for the one eareon condition, the level of recognition was around 70%. This is similar to the work of Brewster [8], which although using a different procedure found similar levels of recognition for single eareons. For the multi-timbre eareon set, not only was timbre identification improved, but also melody identification. We believe this was caused in part by those combinations of eareons which shared the same timbre and register, differing only in melody. Presenting these stimuli with slightly different timbres allowed the melodies to be separated.

The result in Figure 2, although showing that identification between the one eareon and original set conditions varied by 40%, fails to show that the actual numerical differences were less than 0.5 eareons. Therefore, it cannot be assumed that simply reducing the number of eareons concurrently presented will have a useful impact on identification. Rather, it seems that the problems of concurrently presented eareons stem from interactions between eareons rather than the amount of audio presented.

In conclusion, it seems clear from the results that in general it is difficult to identify eareons in cases where more than one is presented concurrently, and designers should be aware before using such a technique, of the amount of information expected to be retrieved from each eareon. We have looked at the worst case where all information needs to be retrieved from each eareon. This may not always be necessary. In cases where such a technique is employed, the multi-timbre and staggered onset techniques should be exploited to improve robustness and increase recognition.

8. ACKNOWLEDGMENTS

This work was supported by EPSRC studentship 00305222.

9. REFERENCES

[1] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, "Experimentally Derived Guidelines for the Creation of Eareons," in *Proceedings of HCI 95*, Huddersfield, UK, 1995, pp. 155-159.

[2] S. A. Brewster, "Overcoming the Lack of Screen Space on Mobile Computers," *Personal and Ubiquitous Computing*, vol. 6, pp. 188-205, 2002.

[3] N. Sawhney and C. Schmandt, "Nomadic Radio: Speech & Audio Interaction for Contextual Messaging in Nomadic Environments," *ACM Transactions on CHI*, vol. 7, pp. 353-383, 2000.

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

[5] W. W. Gaver, "Auditory Interfaces," in *Handbook of Human-Computer Interaction*, vol. 1, M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds., 2nd ed. Amsterdam: Elsevier, 1997, pp. 1003-1041.

[6] A. S. Bregman, *Auditory Scene Analysis*, vol. 1, 1st ed. London, England: MIT, 1994.

[7] S. M. Williams, "Perceptual Principles in Sound Grouping," in *Auditory Display: Sonification, Audification, and Auditory Interfaces*, vol. 1, G. Kramer, Ed. Santa Fe, New Mexico: Addison Wesley, 1994, pp. 95-125.

[8] S. A. Brewster, "Providing a structured method for integrating non-speech audio into human-computer interfaces," Ph.D. dissertation, *Department of Computer Science*. University of York, York, 1994.

[9] A. L. Papp, "Presentation of Dynamically Overlapping Auditory Messages in User Interfaces," Ph.D. dissertation, *Department of Computing Science*. University of California, Davis, 1997.

[10] D. K. McGookin and S. A. Brewster, "DOLPHIN: The Design and Initial Evaluation of Multimodal Focus and Context," in *Proceedings of ICAD 2002*, Kyoto, Japan, 2002, pp. 181-186.

[11] J. M. Gerth, "Performance Based Refinement of a Synthetic Auditory Ambience: Identifying and Discriminating Auditory Sources," Ph.D. dissertation, *Department of Psychology*. Georgia Institute of Technology, Atlanta, 1992.

[12] D. S. Brungart, M. A. Ericson, and B. D. Simpson, "Design Considerations For Improving The Effectiveness Of Multitalker Speech Displays," in *Proceedings of ICAD 2002*, Kyoto, Japan, 2002, pp. 424-430.

[13] D. I. Rigas, "Guidelines for Auditory Interface Design: An Empirical Investigation," Ph.D. dissertation, *Department of Computer Studies*. Loughborough University, Loughborough, 1996.