Auditory Scene Analysis as the Basis for Designing Auditory Widgets

Evangelos N Mitsopoulos & Alistair D N Edwards
Department of Computer Science
University of York
York
England
YO1 5DD
email: enm@cs.york.ac.uk
WWW: http://www.cs.york.ac.uk/~enm

ABSTRACT
This paper presents a methodology for the design of fast-rate auditory presentations based on the Auditory Scene Analysis of Bregman. The auditory scene is hierarchically organized and based on the concept of auditory streams described in terms of two types of structures, one across streams (at an instant) and one within each stream (over time). Each stream constitutes a perceptual entity on which attention can be focused. Integrating information across streams requires effort and practice because the auditory system is inclined to derive properties such as temporal order or rhythm on a within-stream basis. The methodology distinguishes between the structure of the scene and the physical dimensions of sound used to produce this structure. Task performance may be critically affected by the structure of the scene. For this reason the structure is designed with respect to the tasks to be supported. When the structural aspects of the auditory scene have been defined, the physical dimensions of sounds are selected so as to provide the desired structure.

Keywords
Auditory displays, fast-rate auditory presentation, auditory scene structure, auditory widgets, blind users.

INTRODUCTION
In designing an auditory alternative to a visual interface (as one might do in providing an adaptation for a user who is blind) it is necessary to take into account the inherent differences between auditory and visual information. One of the most important differences is the dynamic character of sound. Visual interfaces are mostly static (although animation is becoming increasingly important), and the user can inspect any part of the display at his or her convenience. However, in auditory displays, information is inevitably presented over time. Thus, there are tasks for which it is crucial to present as much information as possible within a short time interval. In such cases, human memory limitations imply that a presentation of longer duration is useless; by the time the sound is finished the listener may have forgotten its beginning or its content.

This paper presents a methodology for the design of fast-rate presentation based on the Auditory Scene Analysis of Bregman [2]. This is part of a larger scale project, the aim of which is to develop an approach to the design of auditory adaptations, based on established psychological principles.

THEORETICAL BASIS OF THE APPROACH
The auditory scene is hierarchically organised and based on the concept of auditory streams[4]. It can be described in terms of two types of structures, one *across* streams (at an instant) and one *within* each stream (over time).

Classical music is a typical example of the across-stream structure. The top level of the structure (Figure 1a) consists of only one stream, the music itself. The two streams in the next level correspond to the orchestra and the soloist. Further down the hierarchy, streams may correspond to single instruments (e.g. piano) or groups of instruments such as strings and drums.

Each stream constitutes a perceptual entity on which attention can be focused. Hence, the above across-stream structure diagram illustrates where attention can be focused *at an instant*. The lower-level streams cannot be perceptually decomposed further (e.g. the group of violins). Consequently, although it is easy to attend to the group of violins, the untrained listener cannot pay attention to a single one of them, unless it forms a separate stream itself (for example, when it is not in tune with the other violins). The factors (physical dimensions of sound) that affect this structure have been reviewed in Bregman (*op. cit.*).

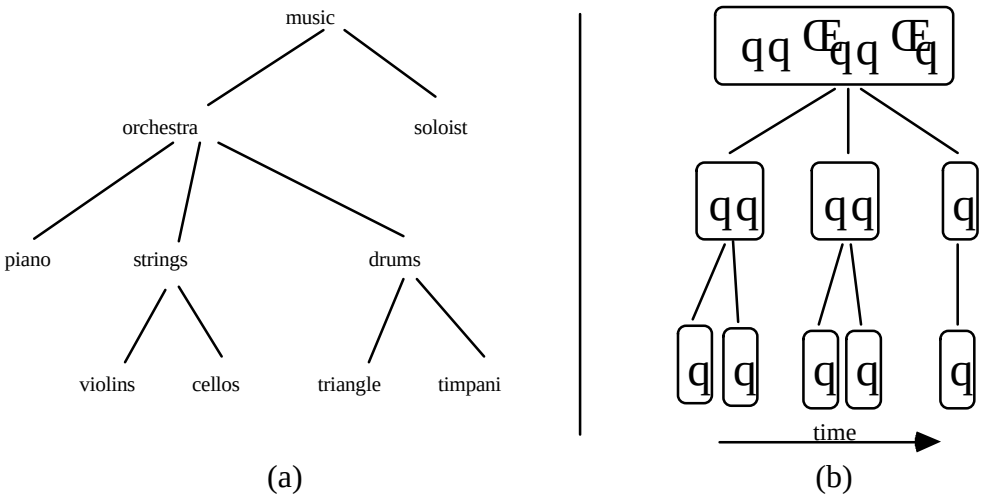

Figure 1: An example of (a) the across-stream structure in classical music and (b) a within-stream structure.

Sequences of two or more different sounds which are repeatedly played (recycled) are frequently used in the study of factors which affect the formation of streams. Van Noorden [5] demonstrated that for a recycled sequence of two pure tones 'A' and 'B' (that is 'ABAB...') streaming is affected by rate (duration of tones) and frequency separation between them. When either the 'A' or the 'B' tones are *selectively* attended, there is little or no sensitivity to the factors mentioned above. In contrast, the ability to divide attention between the 'A' and 'B' tones in the sequence (that is to follow the 'ABAB...' sequence over time) is significantly affected. Fast rates and large frequency separations render this task impossible.

Bregman, interpreting Van Noorden's results, has suggested that there are two types of auditory organisation, the *primitive* and the *schema-based*. The primitive organisation is responsible for the formation of *primitive* streams. When conditions are in favour of the streaming phenomenon (for example fast rates, large frequency separation, different timbres, spatial separation etc.), the sequence 'ABAB...' will segregate into two primitive streams; one which consists of the 'A' tones and another which contains the 'B' ones. It is easy to follow either of these two primitive streams over time. However, integrating information across these primitive streams requires effort and practice, because the auditory system is inclined to derive emergent properties (such as temporal order or rhythm) on a within-stream basis.

If conditions do not favour primitive segregation, the sequence 'ABAB...' will remain coherent in a single stream and can be followed over time. Moreover, it may still be possible to impose a *schema-based* organisation, that is to selectively attend to either the 'A' or the 'B' sub-sequence. Thus, following the 'A' or the 'B' sub-sequence tends to be insensitive to factors that affect streaming since the corresponding streams can be products of the schema-based organisation as well as the primitive organisation. In contrast, the coherency of the sequence depends on the absence of streaming.

If coherence is necessary to the performance of a task, then one option would be the use of similar sounds and/or slow rates. Alternatively, the nature of the streaming phenomenon can be exploited. The first few tones in the recycled sequence will be coherent because stream segregation does not occur right away. Moreover, the introduction of a 4-sec silence at any point seems to reset the streaming process. Thus, the designer may choose to use short, non-repeating sequences (*one-shot* sequences) or to recycle them with intermediate silence intervals.

Each stream consists of units which are hierarchically organised over time. The within-stream structure diagram (Figure 1b) identifies these units and their relations. Whether the listener is able to directly perceive the lower level of the hierarchy (as a sequence of distinct units) or only the higher levels of the structure depends on factors such as rate of presentation.

Another factor important to the perception of the auditory scene is the top-down influence of the listener's knowledge and expectations. The task to be performed can affect perception since it influences which stream will be attended. The nature of tasks may also vary with the perceptually available level of within-stream units. Hence, the design of the auditory display should aim at the analysis of both types of structure (within-stream and across-stream) in order to support a full range of tasks.

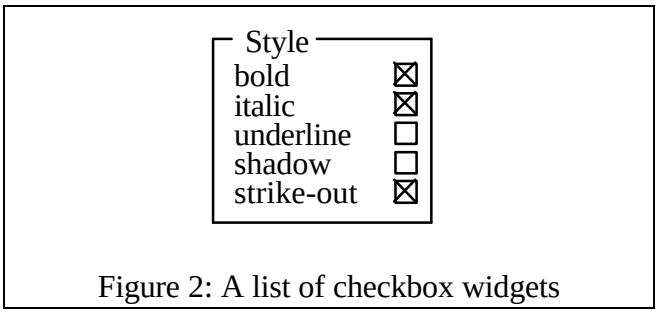

Figure 2: A list of checkbox widgets

THE PROPOSED METHODOLOGY

A distinction is adopted between the structure of the auditory scene and the physical dimensions of sound that give rise to this structure. Different configurations of the physical dimensions may result in exactly the same structure. That is to say that there may be more than one way to implement a given structure. Moreover, the structure of auditory scene may significantly affect task performance. Hence, the abstraction of the structure from the physical dimensions and the study of the interrelations between the given set of tasks and the structure is envisaged as a flexible and powerful approach; the outcome of the analysis would be applicable to different configurations. At the same time, any such configuration could be analysed in structural terms with respect to its suitability for supporting the related tasks.

The list of checkbox widgets in Figure 2 will be used as a simple exemplar of the methodology. The specification of the set of related tasks is the first concern. Let us assume that the following tasks should be supported.

1) Identification of the list name.

2) Identification of the list status (e.g. all checkboxes are checked or unchecked).

Obviously, there are many more tasks that can be performed with the visual counterpart, but here the focus is on the fast presentation of the list, equivalent to just having a *glance* at it. Tasks based on the label of each checkbox will not be discussed here, as this information is not available in a glance. Furthermore, for reasons of simplicity, the analysis is constrained to the above two tasks, although many more tasks can be supported with fast presentation.

One appropriate across-stream structure is illustrated in Figure 3. The first task suggests that the spoken list name forms a stream (3) on its own. The representation of the status of each checkbox (either checked or unchecked) requires two different types of sound. These two types of sound will segregate at fast rates of presentation, giving rise to stream 4 (all the checked boxes) and stream 5 (all the unchecked ones). Since streams 4 and 5 are probably much more similar to each other than to stream 3, they are grouped under stream 2. In this way the second task is also supported because there is no interference in stream 2 from stream 3. Nevertheless, these two streams should be perceptually integrated as they are conceptually interrelated. Hence, they should be grouped under stream 1.

Each stream should be examined over time, too. Streams 2, 4 and 5, which provide information about the status of the checkboxes, are of particular interest because they may affect performance in the second task. Each of these streams can be analysed in units over time as shown in Figure 4.

There are several different ways to perform the second task. At fast rates of presentation, stream 2 will segregate into streams 4 and 5. The rhythm (or the absence) of either stream 4 or 5 could provide the required information in most cases. For example, the gap in the sequence of stream 4 (‘XX– –X’ as shown in Figure 4) would indicate that there are unselected

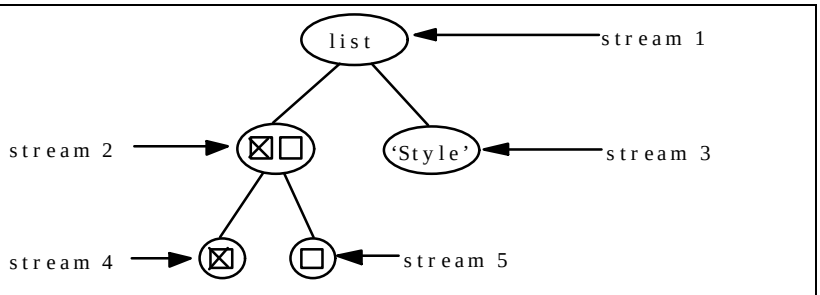

Figure 3: The across-stream structure diagram of the widget list.

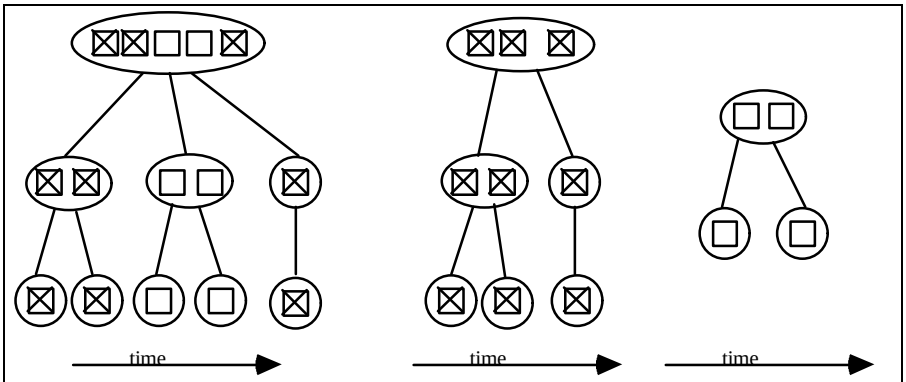

Figure 4: Within-stream structures for streams 2, 4 and 5

as well as selected checkboxes. Although the user will be able to attend to the sequence of either stream 4 or stream 5 exclusively, it is not possible to attend both of them at the same time because this would require integration of information across primitive streams.

At slow rates of presentation, where stream 2 remains a coherent sequence, it is possible not only to accomplish the second task but also to perceive the *order* of the individual checkboxes in the sequence. Thus, order-report tasks become feasible.

Which method will be adopted in the performance of the second task depends on factors such as:

- the status of each checkbox in the list;
- the user's strategy, preferences and expectations;
- the rate of presentation which affects the lower level of each within-stream structure where units remain perceptually distinguishable, and
- the ability of the user to mentally reconstruct the lower levels of the structure when these are perceptually unavailable.

Hence, the heuristics a user might employ are diverse. It is important to support as many of them as possible if tasks are to be performed with ease and without requiring extensive training.

Once the analysis of tasks and structures (across and within stream) has been accomplished, the physical dimensions of the sound can be specified. Bregman (*op. cit.*) has extensively reviewed the factors that affect sequential and simultaneous integration. In our implementation, the segregation of the spoken list label from the checkboxes is reinforced by spatial separation (binaural stereo). Distinct timbres are used to segregate stream 2 into streams 4 and 5. The segregation is emphasized by a pitch separation of an octave. Streams 2 and 3 are presented simultaneously to integrate them under stream 1 (temporal proximity). Nevertheless, their onset asynchrony and some frequency separation keep them distinct. The length of each note representing a checkbox is kept fixed (in a range of 35 msec to 75 msec). The silent gap between successive notes can be varied (from 10 msec to 200 msec) to control the perceived lower level of the within-stream structures. That is to say that the rate of presentation can be manipulated to support the perception of different levels of the within-stream structures.

CONCLUSIONS AND FURTHER WORK

The design of most auditory interfaces is somewhat *ad hoc*. This work is unusual in that it uses a theoretical basis for its designs. The methodology has been developed using widgets such as checkboxes, radio-buttons and sliders. Exemplars of checkbox widgets have been implemented and are being tested in experiments involving several tasks and at different rates of presentation. The results are encouraging and promising. For tasks which involve order perception, rates as fast as 75 msec per widget have been achieved. Furthermore, for the tasks discussed in the previous section, rates as fast as 40 msec have been achieved with minimal training. The formal experimental evaluation of the method is in progress and current results comply with the predictions of the methodology.

There is a close linkage to the proposed methodology and the general notion of auditory earcons ([1], [3]). Earcons have been used not only for widgets but also for conveying information such as menu hierarchies. Similarly, the proposed methodology is not confined to the design of auditory widgets. It is intended to be applicable to the design of all aspects of auditory user interfaces, as well as several categories of information such as auditory bar-charts and graphs. Moreover, many designs derived using the proposed methodology can be thought of as earcons since the concept of the latter is particularly general. Looked at another way, any earcon can be analysed in terms of its suitability to support a particular task, using the methodology.

Nevertheless, apart from the relationships between the methodology and earcons, there are some important distinctions, too. The methodology, at its current state, is mostly focused on the *fast* presentation of *multiple* 'earcons' and the exploitation of their emergent gestalts. On the other hand, research on earcons has rather been more vigorous in aspects such as the detailed presentation of the interface (which takes place when the user interacts with the auditory user interface). Moreover, manipulation of the dimensions of sound in the methodology is theory-driven, in contrast to the design of earcons which is mostly empirical. This is to say that the methodology provides the means to systematically drive the designer's intuition, to assess the suitability of the resulting design, and to reject inappropriate alternatives prior to any experimental evaluation.

There are two further major research objectives. The first is to apply the methodology to the complete hierarchy of widgets — not just its lowest level. Second, links must be established between 'at a glance' presentation and detailed presentation. These two modes of presentation are clearly interrelated, since fast presentation could provide contextual information necessary to the formation of the user's expectations of the detail presentation. For example, in the list of checkboxes, if 'strike-out' was the only selected option and the user would like to switch back to normal style (all checkboxes unselected), then he or she could infer that the desired box is close to or at the end of the list and move faster towards the target. Hence, it is envisaged that the specification of the auditory scene should take into account this issue, in order to maximize the information communicated to the user.

REFERENCES

1. Blattner, M. Sumikawa, D. & Greenberg, R. (1989). *Earcons and icons: Their structure and common design principles*. In Human Computer Interaction, 4(1), pp 11 - 44
2. Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts: MIT Press.
3. Brewster, S. A., Wright, P. C. & Edwards, A. D. N. (1992). *A detailed investigation into the effectiveness of earcons*. In Auditory Display, sonification, audification and auditory interfaces. G. Kramer (ed.). The Proceedings of the First International Conference on Auditory Display, SFI. Addison-Wesley, pp 471 - 498.
4. Williams, S. M. (1994). *Perceptual Principles in Sound Grouping*. In Auditory Display, Ed. Gregory Kramer, SFI Studies in the Sciences of Complexity, Proc. Vol. XVIII, Addison-Wesley, 1994. pp. 95 - 125.
5. Van Noorden, L.P.A.S. (1975). *Temporal Coherence in the Perception of Tone Sequences*. Unpublished doctoral dissertation, Eindhoven University of Technology.