A Principled Methodology for the Specification and Design of Non-Visual Widgets

Evangelos N. Mitsopoulos

Department of Computer Science, University of York
York, UK.

Alistair D. N. Edwards

Department of Computer Science, University of York
York, UK.

Abstract

When the visual channel of communication is unavailable because the user is blind, non-visual user interfaces must be developed. The proposed methodology consists of three interrelated specification levels. Information and supported tasks are specified in abstract terms at the conceptual level, taking into account requirements imposed by manipulation of interaction devices and information provided by analysis of the visual representation. The perceptual structure of the auditory scene is specified next at the structural level and then the physical dimensions of sound are defined at the implementation level. The methodology is applied to the specification of a simple listbox widget.

1 Introduction

Sight is a very powerful means of communication. It is used in every-day communications through diagrams, pictures, gestures and many other forms; it is also the basis of most human-computer interaction. When that channel of communication is unavailable because the person is blind, then non-visual alternatives must be developed. A significant amount of research effort has been expended in recent years on the problem of how to substitute the visual communication of the graphical user interface (GUI) so that it can be used by blind people (see, for instance, [1, 2, 3, 4]). Most approaches are based on translation of the *surface* visual representation to a non-visual equivalent. The clearest example of translating at the surface level is the GUIB project [5]. A contrasting approach is that of the Mercator project [2]. In the Mercator interface, all reference to the visual appearance of the (X-Windows) interface is lost and the user interacts with the components arranged in a logical, hierarchical structure.

The objective of the research effort discussed herein is to develop a more principled basis for the design of non-visual interfaces. Our approach comes somewhere between those of the GUIB and the Mercator project. It uses the visual representation of the interface, but acknowledges the fact that there is more to the visual layout than a surface-level scan would reveal. Widgets are used as exemplars since they are essential components of GUIs and clearly capture interaction issues such as ear-hand co-ordination and auditory-haptic multi-modal integration.

The focus of this paper is on the specification of the auditory part of non-visual widgets. Properties of visual widgets that should be preserved in their non-visual counterparts are identified by perceptual analysis of the visual representation. The influence exerted by issues emerging from the manipulation of interaction devices is considered, too. However, the methodology need not be restricted to widgets because it is based on general psychological principles (those on auditory perception have been presented in [6]).

2 Aspects of Visual Perception

Understanding the perceptual organisation of visual widgets may provide important insight into the underlying mechanisms that support and enhance task performance. These mechanisms could be replicated in the auditory widget as far as possible, to preserve its efficiency. The Interacting Cognitive Subsystems (ICS) model [7] encompass the majority of visual perception issues and could serve as a framework accessible to the designer.

According to the ICS model, the visual scene is hierarchically organised into a structure of objects which group into larger ones and can be decomposed into smaller objects. For instance, the four checkboxes (in each of the figures 1a and 1b) group together, as well as their four labels. However, only in Figure 1a, each checkbox and its label are clearly related, forming together a group. The structure of the visual scene may significantly affect task performance because it restricts how people may search through it for a particular object. Thus, as a result of their structure, both widgets efficiently support tasks such as the identification of all available effects (Bold, Italic, Underline and Shadow) because browsing through objects in the same group requires simple

(a) **(b)** **(c)**

Figure 1: (a) a checkbox widget (b) its presentation by some screen readers, (c) a listbox widget

transitions of attention compared to those between objects in different groups. However, finding if "Underline" is checked is harder in Figure 1b, because the necessary objects (the label and the associated checkbox) are in different groups. Identification of groups which are necessary for efficient task performance in the visual representation can be useful in the specification of *streams* in the auditory scene.

Normally, tasks involving checkboxes do not require their decomposition. As far as these are concerned, the *lowest* level objects are the checkboxes themselves which may be either checked or unchecked. All information about their constituent objects (in this example, a square and possibly a cross) can be discarded since no such task would enquire about it. Consequently, an *auditory* checkbox need not contain any information about these constituent visual objects. It could be merely represented by a note (an auditory object) as long as this note would contain sufficient information to represent the two distinct states of a checkbox. Consideration of the perceptual structure of a widget and of the set of associated tasks may facilitate the distinction between information that must be conveyed in sound and information which is essential to the visual widget only.

A phenomenon usually exploited by widgets is the 'pop-out' effect. In figures 1a and 1b, checkboxes form two sub-groups. The unchecked checkbox is spatially proximal but dissimilar to the checked ones, and thus, it pops-out. There is no need to search for it; when the group of checkboxes is attended, the unchecked checkbox effortlessly becomes the focus of attention. A typical example of the pop-out effect exploited by visual widgets is the listbox widget in figure 1c, where attention can be immediately focused on the highlighted entry ('Normal').

In conclusion, visual analysis offers a substantial volume of information that can be exploited in the specification of the auditory widget. Of course, the final specification of the auditory widget may significantly depart from what visual analysis suggests. It should be acknowledged that apart from the similarities between the two sensory modalities, there are important differences, too. However, visual analysis is worth undertaking because the insight gained may considerably facilitate the initial steps in the specification of non-visual widgets.

3 Specification of Auditory Widgets

The design methodology consists of three levels of specification: conceptual, structural and implementation. These will be illustrated using the checkbox widget (figure 1a) but they can be applied abstractly to non-visual widgets in general: The foundations of the methodology rely on general visual and auditory perceptual theories.

3.1 Conceptual Level

The set of tasks associated with the widget under design as well as the (abstract) information necessary to perform them are specified at this level. For the checkboxes example, these tasks would include identification of the status of the checkbox column (whether all checkboxes are checked or unchecked, useful to deduce if current style is normal), browsing of available effects, identification and/or toggling of the value of an effect, and identification of the purpose and the type of the widget (a number of checkboxes controlling the style Effect).

The necessary information to perform the related tasks can be defined in terms of abstract dimensions which can be of nominal, ordinal, interval, or ratio type. For example, because the two states of a checkbox are distinct but not ordered they can be represented by two values along a nominal dimension. When physical dimensions of sound are selected to convey the information specified at the conceptual level, they should be of the same scale as their abstract counterparts. Timbre is a nominal dimension and it would be appropriate to represent the status of a checkbox. Pitch (on its own) would be a less appropriate dimension since it is ordinal and the user might erroneously deduce that the two states are ordered (for example, low - high).

Abstract dimensions are the constituents of semantic entities. A checkbox widget is an entity that has two dimensions, a nominal one representing its status and another one for its label. Depending on the set of

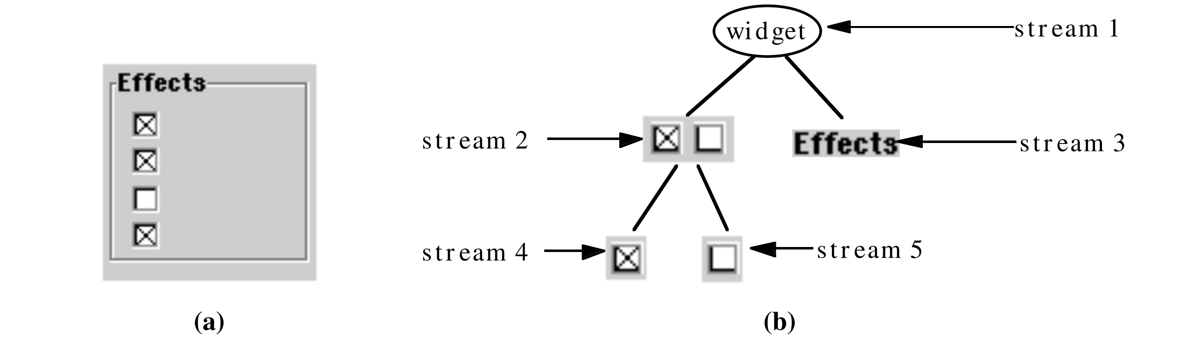

Figure 2: A checkbox widget at a glance and its across-stream structure

supported tasks, entities may be semantically related forming higher-level entities. Tasks such as identification of the status of the checkbox column or browsing of labels imply that checkboxes as well as labels should be interrelated. Higher-level entities can be related, too. For example, there may be a number of widgets in a dialog box. In other words, there may be a conceptual hierarchy of entities.

The main difference between the conceptual and the other levels of specification is that only the former is medium-independent. Because it mediates between the visual and the auditory representation, had it not been medium-independent, it would be possible to erroneously include some information necessary to the visual widget only in the specification of the auditory representation. Another reason for its medium-independence is that a number of multi-modal issues are considered and resolved at this level. Use of an interaction device (such as the mouse or touchpad) introduces a number of navigation tasks which require further information to be accomplished. For example, while browsing, current position relative to either end of the list of checkboxes is necessary to co-ordinate motor movements. However, available haptic information might not be adequate. The 'missing' information is defined in conceptual terms so that it can be accommodated in the auditory representation. Moreover, ear-hand co-ordination may pose additional requirements (such as advance auditory feedback) which affect conceptual specification. These design implications are illustrated with the design of a simple listbox widget in Section 3.4.

The influence of conceptual specification to the other two levels is significant. As mentioned above, the choice of physical dimensions of sound suitable to represent an abstract dimension is restricted by its scale. Hence, a number of constraints are introduced at the implementation level. Moreover, the semantic relations between entities should be perceptually encoded in the auditory structure of the widget. For example, the list of checkboxes should be *perceived* as such. If all checkboxes form a single auditory stream, their semantic relation will be readily perceived; it need not be deduced. This is to say that the conceptual level may also impose a number of constraints on the structure of the auditory scene which is defined at the structural level.

3.2 Structural Level

This level aims at the specification of the auditory scene structure. The psychological theory and specification tools used in this level have been presented in [6] and are briefly mentioned here. According to Bregman's Auditory Scene Analysis [8], the auditory scene is hierarchically organised and based on the concept of auditory streams. It can be 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. It is easy to attend a stream over time, but it may be almost impossible to integrate information between two streams which have emerged from *auditory stream segregation* which takes place at fast rates of presentation. Since fast presentation is of paramount importance to efficiency, it is necessary to analyse supported tasks with respect to their information requirements. If a task is based on perceiving temporal relations (such as order or rhythm) between objects in different streams, it will be obstructed at fast rates by auditory stream segregation. Either all information must be in a single stream or a slow rate of presentation should be used to eliminate segregation. Moreover, temporal relations might not be inferred even at relatively slow rates of presentation which do not favour segregation but are too fast for cognitive processes to keep up with.

An across-stream diagram describes the across-stream structure of an auditory scene. It has the interesting property of being independent of the rate of presentation and can be used in the specification of the desired structure of the auditory scene. For instance, consider the information extracted by the user at a glance. This is shown in figure 2a. If each state of a checkbox is represented by a distinct timbre, then, at fast rates, a sequence of them will segregate into streams 4 and 5, as shown in figure 2b (adapted from [6]). But even at slower rates, where segregation does not occur, it is still possible to attend to either group of checkboxes. That is, streams 4 and 5 are still present as far as the interface designer is concerned, although they are the product of a psychological mechanism different to that employed at fast rates. In both cases, the title of the sequence

('Effects'), presented using synthesized speech, forms a distinct stream, too.

At fast rates, tasks involving order detection will become almost impossible: either the users will not perceive a coherent sequence of checkboxes because of the stream segregation taking place, or the sequence will be just too fast to be meaningful. Nevertheless, they will be able to perceive that the sequence consists of both checked and unchecked checkboxes. If temporal relations are not important to a task, segregation need not be considered in the design of the appropriate across-stream structure. On the other hand, the designer should ensure that dividing attention among segregated streams is necessary for no task.

Structural specification is kept separate from the conceptual one because it is medium-dependent. Moreover, a distinction is adopted between the perceptual structure of the auditory scene and the physical dimensions of sound that give rise to this structure, since different configurations of physical dimensions may result in exactly the same structure. Also, as discussed above, task performance when integration of information is involved is affected rather by the structure of the auditory scene than the physical dimensions of sound. For these reasons, structural specification is distinguished from implementation level specification.

Visual analysis may provide some initial insight about groups which might be realised as auditory streams. Nevertheless, there might be considerable differentiation between what visual analysis suggests and the end-specification of this level. In any case, considerable constraints are imposed on the implementation level, since the physical dimensions must be such that give rise to the specified perceptual structure.

3.3 Implementation Level

In the implementation level the physical dimensions of sound are defined. Restrictions that have been imposed by the conceptual and the structural specification have to be satisfied. Bregman's [8] extensive review of factors that affect segregation may be used to suggest which dimensions of sound are appropriate for a particular across-stream structure. Apparently, their suitability might be further restricted by their type (nominal ordinal, interval or ratio). If the imposed constraints cannot be satisfied then alterations in the structural and/or conceptual level are required. Whenever feasible, the designer should opt for redundancy by using more than one physical dimension for each abstract one as well as for achieving the specified auditory structure. In this way, the segregation phenomenon is intensified and thus, auditory streams become more distinct. In other words it is easier to focus on a particular stream and attend it undistracted over time. Moreover, because hearing abilities may significantly vary, the robustness of the design is enhanced by the provision of multiple cues.

3.4 An Example: the Specification of a Non-Visual Listbox Widget

To highlight the application of the methodology, some important aspects in the specification of a simple listbox widget (shown in figure 1c) are briefly considered. Associated tasks include identification and inspection of the selected entry ('Normal'), selection of a new entry and comparison of an entry with the selected one. Performance in most of these tasks mainly relies on fast identification of the selected entry and quick browsing through the list.

The interaction device used in this example is a flat resistive touchscreen with a tactile grid on its top. The main purpose of the tactile grid is to provide information that allows the user to move horizontally or vertically while browsing the touch screen. Apart from this added information, the interaction device resembles a single-button mouse in its functionality.

Although information appears to be mostly verbal, visual analysis reveals that graphical information is of key importance, too. Because the selected entry is highlighted, it pops-out. As soon as the listbox is attended, it can be instantly recognised without knowledge of its content ('Normal'). The listbox window is a mechanism imposed by potential spatial restrictions. Yet, its added functionality of scrolling up or down a window at a time is important: Verbal presentation implies that the user may inspect one entry at a time (corresponding to the sighted user who can look at a single item, referred to as the 'inspected entry'). If the list is sorted, the added functionality of this mechanism becomes very useful for fast browsing. In addition, spatial constraints are re-introduced by devices such as touchpads, rendering the window mechanism necessary.

For the inspected entry, conceptual level information consists of its verbal content and contextual information in the window. What information is considered as context depends on the supported tasks. Fast identification of current selection implies that the user is able to quickly infer if the selected entry is in the window, as well as, its position relative to the inspected entry. To avoid errors such as moving off the window while browsing, it is important to know the position of the inspected item relative to either end of the window. Ear-hand co-ordination necessitates the provision of advance feedback on either end of the window and the selected entry. Hence, at the conceptual level, the inspected entry is an entity <content, relative position of selected entry, position of inspected entry in window>. If the selected entry is in the window, the second dimension is ordinal: {before (far) selected entry, before (close), on, after (close), after (far)}. The third dimension is also ordinal {"top" of window, close to "top", middle, close to "bottom", "bottom"}.

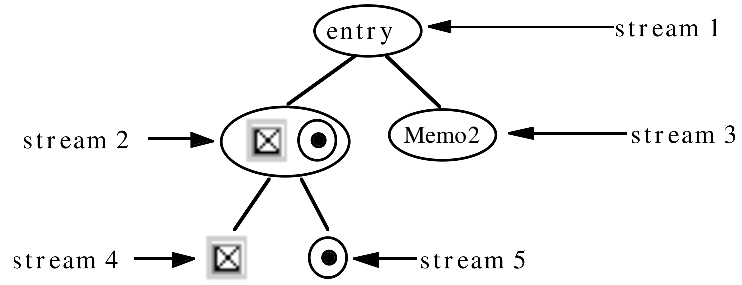

Figure 3: An across-stream diagram for the listbox widget

At the structural level, the inspected entry consists of three simple objects: the spoken content (provided that sufficient presentation time is allowed by the user), the selected entry (if present in the window) and the position of the inspected entry in the window. That is, each one of the objects corresponds to an abstract dimension. When the user is browsing the entries, three corresponding streams emerge giving rise to the across-stream hierarchical organization depicted in figure 3. Stream 3 is the spoken content, stream 4 is the position of the selected entry and stream 5 conveys information on the position of inspected entry ('Memo2' in Figure 3) within the window. It is important to consider fast presentation issues and, in particular, to verify that across-stream integration of information is not required for any of the supported tasks. All information necessary for locating the selected entry is contained in stream 4. Information for moving to either end of the window is contained in stream 5. The only case where the user would have to attend to more than one stream at the same time occurs while browsing the entries in the window. Then one would have to attend to both stream 3 and 5. However, browsing would be performed at slow rates because of the time required to present even a part of the spoken content; stream segregation would not take place. Hence, the analysis of the emerging across-stream structure suggests that there are no problems with the structure and the associated tasks.

The following is a possible implementation level specification of physical dimensions. In general, speech is used for the verbal content of entries, but presentation is interrupted when the user moves to another entry to resume with the contents of that entry. Its delayed onset makes stream 3 even more distinct from the others and thus easier to attend to.

Sighted users (and perhaps late-blind users) who probably have some experience with the visual listbox widget, might use their visual imagery to structure their understanding of the non-visual list-box. In other words, they could take advantage of a mental model which has specific spatial properties. Therefore, it would be better to arrange the entries on the touch pad manipulation device in a vertical layout, similar to that of the visual widget. A horizontal layout would be in conflict with their mental model. However, if a blind user has no previous experience of a listbox widget, then it might be the case that other mental models could be exploited. For example, a horizontal layout of the entries could be desirable since it resembles a sentence in braille and blind users can be more familiar with this type of browsing. Consequently, the entries in the window of the listbox widget have been arranged from left to right.

The user can inspect one entry at a time which consists of the spoken content of the entry, a pure tone denoting the position of the inspected entry relative to the window, and a noise-like auditory object which represents the selected entry and its position relative to the inspected entry. Distinct pitch and timbres have been used to keep apart the emerging streams (4 and 5 in figure 3) and to make sure that the desirable across-stream structure is achieved.

An example of slowly browsing the entries in the listbox window depicted in figure 3 is given here (sound1[1]). As the user is browsing the entries from left to right, they can listen to the voice speaking each entry and to the pure tone denoting the position of the inspected entry in the window. Sounds are presented in binaural stereo and the tone is displaced to the left when 'Manuscript 1' or 'Manuscript 3' is reached. Similarly, it is displaced to the right for 'Presentation 1' and 'Press Release 1'. For the rest of the entries it is in the middle. Hence, the user may know when either end of window has been reached. For the entries just before either end of the window ('Manuscript 3' and 'Presentation 1') advance feedback is provided using dissonant pitch (two tones with dissonant pitch are simultaneously played). Also, the loudness of these two entries is higher than the others, so that their identification becomes easier. In other words, these two entries are different from the others not only in terms of pitch but also in terms of loudness, which increases the robustness of the widget, as discussed in Section 3.3.

When the user is not on the selected entry the noise-like sound is appropriately displaced. When this sound is on the right, the user should move to the right to find 'Normal'. Similarly, a displacement to the left

[1] The listener should use headphones to make the binaural displacement of auditory objects clearer.

denotes that the user should move to the left. To provide advance feedback, loudness and spectral brightness are increased just before and after the current selection (i.e. for the entries 'Newsletter Wizard' and 'Presentation 1'). A rhythmic pattern is used to distinguish these entries from the selected entry. These cues can be used to quickly locate the currently selected entry (sound2), which is one of the essential tasks that the widget should support. Reaching either end of the listbox window can also be performed at fast rates, by attending to the pure tone (sound3).

Evidently, the above is just a part of the specification of even a simple listbox widget. Issues related to scrolling the window have been omitted for simplicity. Also, it should be noticed that there is a difference between passively listening to the last two auditory examples and actually interacting with the auditory scene. In the latter case, the user may have already formed his or her expectations of the target sound and is comparing them to the auditory scene at a perceptual level. This is the reason that the above tasks can be performed quickly. On the other hand, trying to understand each of the constituent sounds is a cognitive process which requires considerably more time. Consequently, it is much easier to attend to the first example rather than the last two as a passive listener.

4 Conclusions

The proposed methodology has been applied to the design of a number of visual widgets such as checkboxes, radio buttons, scroll bars and listboxes. Its experimental evaluation may take a number of forms. Comparison with screen-readers would be superficial since the methodology explicitly tackles the problem of considering only the surface representation. A more meaningful comparison would be to contrast its artefacts with other non-visual widgets. A comparison between non-visual and visual widgets would be even more desirable, since the latter are regarded as highly efficient. This comparison is currently pending, as the tactile grid mentioned in Section 3.4 is currently under design. An appropriate tactile grid would ensure that the necessary haptic and, in particular, spatial information for interacting with the non-visual widgets is available to the user. Otherwise, differences in performance cannot be directly attributed to the auditory part of the non-visual widget.

References

1. Edwards A.D.N: Soundtrack: An auditory interface for blind users. Human Computer Interaction 1989; 4(1):45-66

2. Mynatt, E.D. & Weber, G: Nonvisual presentation of graphical user interfaces: contrasting two approaches. In C. Plaisant (ed.), Celebrating Interdependence: Proceedings of Chi’94, Boston, New York, ACM Press, 1994. pp 166-172

3. Harness, S., Pugh, K., Sherkat, N. and Whitrow, R: Fast icon and character recognition for universal access to wimp interfaces for the blind and partially sighted. In E. Ballabio, I. Placencia-Porrero and R. P. d. l. Bellcasa (ed.), Rehabilitation Technology: Strategies for the European Union (Proceedings of the First Tide Congress), Brussels, IOS Press, 1993, pp. 19-23.

4. Schwerdtfeger, R: Making the GUI Talk. In Byte Magazine, December 1991, pp. 118-128.

5. Weber, G: Access by blind and partially sighted people to interaction objects in MS-Windows. In Proceedings of Ecart2, Stockholm, 1993, pp. 2.2.

6. Mitsopoulos, E. & Edwards, A.D.N: Auditory Scene Analysis as the basis for designing auditory widgets: Proceedings of International Conference on Audio Display (ICAD’97) Xerox Corp, 1997, pp 13-18

7. May, J., Scott, S. and Barnard, P: Structuring Displays: a psychological guide. In Eurographics Tutorial Notes Series. EACG: Geneva, 1995

8. Bregman, A.S: Auditory Scene Analysis. Cambridge, Massachusetts, MIT Press, 1990

9. Zhang, J: A Representational Analysis of Relational Information Displays. In International Journal of Human-Computer Interaction, vol. 45, 1996, pp. 59 - 74.