THE DESIGN AND EXPLORATION OF INTERACTION TECHNIQUES FOR THE PRESENTATION OF FOREGROUND AND BACKGROUND ITEMS IN AUDITORY DISPLAYS

David Dewhurst

www.HFVE.org

david.dewhurst@HFVE.org

Tony Stockman

Queen Mary University of London
(School of Electronic Engineering and Computer Science)
Mile End Road, London, UK
t.stockman@qmul.ac.uk

ABSTRACT

This work forms a part of a wider project, in which one of the authors is developing a system to sonify images (and other material) via sets of audio (and tactile) effects. The main contribution of this paper is to describe the design, and examine the effectiveness, of "multi-talker focus effects" in directing the user's attention to particular items, while at the same time making them aware of other co-located (or separate) items.

Additionally, the paper describes approaches to presenting and navigating multi-level representations of visual scenes, and of non-visual and non-spatial information and entities. It describes how external client application-generated (or manually produced) material can be submitted to the system, and considers several interaction methods, including using multiple taps on parts of images to command the system.

Initial results are reported from informal assessment sessions with a totally blind person, and a sighted person.

1. INTRODUCTION

It is estimated that there are about 39 million blind people in the world [1]. Several attempts have previously been made to present aspects of vision to blind people via other senses, particularly hearing and touch. The approach is known as "sensory substitution" or "vision substitution".

1.1. Previous work

Work in the field dates back to Fournier d'Albe's 1914 Reading Optophone [2], which presented the shapes of characters by scanning across lines of type with a column of five spots of light, each spot controlling the volume of a different musical note, producing characteristic sets of notes for each letter.

Other systems have been invented which use similar conventions to present images and image features [3, 4], or to sonify the lines on a 2-dimensional line graph [5]. Typically height is mapped to pitch, brightness to volume (either dark- or light- sounding), with a left-to-right column scan normally used. Horizontal lines produce a constant pitch, vertical lines produce a short blast of many frequencies, and the pitch of the sounds representing a sloping line will change frequency at a rate that indicates the angle of slope.

Previous work in the field is summarised in [6, 7]. Previous approaches have allowed users to actively explore an image, using both audio and tactile methods [8, 9]. The BATS (Blind Audio Tactile Mapping System) [10] presents maps via speech synthesis, auditory icons, and tactile feedback. The GATE (Graphics Accessible To Everyone) project allows blind users to explore pictures via a grid approach, with verbal and non-verbal sound feedback provided for both high-level items (e.g. objects) and low-level visual information (e.g. colours) [11, 12]. An approach used by the US Navy for attending to two or more voices is to accelerate each voice, and then serialise them [13].

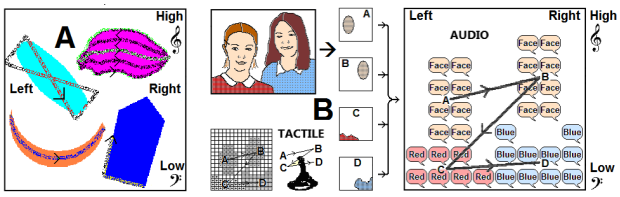

Figure 1. *Presenting items via "Tracers", and "Imprints".*

One of the authors has previously reported other features of the HFVE (Heard and Felt Vision Effects) system [14], notably using moving audio and tactile effects that trace out shapes, with corners emphasised ("Tracers" and "Polytracers") (A) Fig 1; using buzzing sounds and other effects to clarify the shapes of items; and using groups of voices, speaking in unison, that rapidly convey the properties, and the approximate size and location, of items ("Imprints") (B) Fig 1 [15, 16, 17].

1.2. Multi-Level Multi-Talker Focus effects

We describe the design, and examine the effectiveness of multi-level multi-talker focus effects Fig 2.

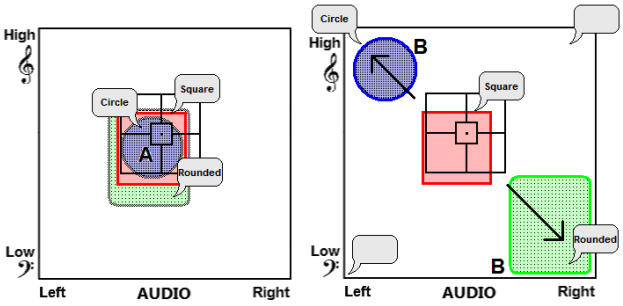

Figure 2. *Multi-talker focus effects, and effect relocation.*

Multi-level multi-talker focus effects Fig 2 are designed to work as follows:-

The system presents the items that are currently the primary focus of attention via crisp non-modified sounds, for example via speech sounds. At the same time the system presents the

speech sounds for items that are not at the focus of attention, but applies a distinct differentiating effect on them, for example by changing the character of the speaker, or by applying echo or reverberation effects.

The intention was that the effects might be perceived in a similar manner to the effect of shallow depth of field in a photograph, where the focused elements are accentuated, and out-of-focus elements are also present which the observer is aware of but not directed towards. The user can interactively control the focus of attention presented by the Focus effects.

The Focus effects Fig 2 will typically have higher user interaction than the previously-developed tracers and imprints Fig 1, as the user will generally want to actively control the items presented by the new effects. Tracers and imprints can be combined with and presented using the new effects. For tracers, imprints, and Focus effects, vertical position is mapped to frequency, and horizontal position to left-right stereophonic positioning in "soundspace" Fig 2. (Similar mappings are used in other sonification systems, for example "the vOICe" [3].)

The user can control which items are being presented, for example via a mouse pointer, or via touch; or the system can automatically sequentially step around or list the most important items found within a user-defined area (including the whole image). Several related new interaction methods are also available and are described, for example coded tapping, and touchpad control, and their application to Focus effects, for example to drill down and up levels of view.

The degree of directed focus presented via Focus effects for any item can be based on its "focus property value" i.e. a value assigned to the item, representing the wideness (high-level) or detail (low level) of particular properties for that item. Several such item focus property values can be present simultaneously at a single point in a scene (A) Fig 2. For example for a computer spreadsheet (B) Fig 3, at any one point the "level of categorisation"/"level of view" emphasised can be the (low-level/detailed) cell at that point; or alternatively the wider, high-level cell block containing the cell can be emphasised (with the cell column, and cell row, containing the cell being of intermediate level). The system allows rapid navigation between such levels of view, for example by using a mouse wheel. The focus property value can be for spatial properties such as the item's distance (or lateral distance), or can be a visual property value, or level of view, or non-visual and non-spatial property (as explained below).

The amount of the de-emphasising effects can be related to the difference in the item's focus property value from the focus property value currently being emphasised; or alternatively there can be a sharp step-change in the effects, so that the emphasised items at the centre of attention are clearly different in perceived quality from non-emphasised items.

The system makes use of the cocktail party effect i.e. being able to focus one's auditory attention on a particular presented item while filtering out other sounds [18]. The system can artificially separate the presented items (B) Fig 2, so that the cocktail party effect is maximised. (**Note:** The term "cocktail party effect" is sometimes used to refer to the effect wherein certain words, typically your own name, suddenly catch your attention, though they are being spoken in a conversation which you are not part of. In this paper the term is used for its other meaning of following one speaker when several are speaking.)

This paper includes an initial evaluation of the approaches, which allow managing of complexity and awareness of items, as well as providing for different levels of view of items in complex auditory scenes.

The nature and aesthetics of the sonification effects can be experienced by visiting the website of one of the authors [14], which includes demonstration videos.

(Note that not all of the features described in this paper are fully implemented at the time of writing, notably some of the locking commands, and some combinations of effects.)

2. MULTI-TALKER FOCUS EFFECT FEATURES, THEIR PRODUCTION, AND USE

By using Focus effects, the system allows several properties and items of the same region to be presented and investigated at the same time. This feature may produce a qualitatively different impression on the user from the previous approaches.

2.1. Overview

The approach is illustrated by the following examples, which feature two different scenarios:-

Fig 3 shows two scenes, one relating to the countryside (a bird perched on a branch of a tree (A)), and the other relating to office administration (a computer spreadsheet (B)). In both cases a pointer is positioned over part of the scene. In the first example (A) the pointer is over one of the bird's feathers. If a sighted person's centre of gaze was similarly positioned, without moving their gaze the sighted person's attention could be concentrated on either:- one of the bird's feathers; or the bird's wing; or the bird; or the branch on which the bird is perched; or the part of the tree in their field of view.

In a similar manner for the spreadsheet (B) the pointer is over a particular cell, but is also over a column of cells, a row of cells, a block of cells, and the spreadsheet. Likewise the user's focus of attention can be drawn towards any one of these spreadsheet items (cell, column, row, block etc.) while at the same time the user can be made aware of the other co-located items, which are at different levels of view.

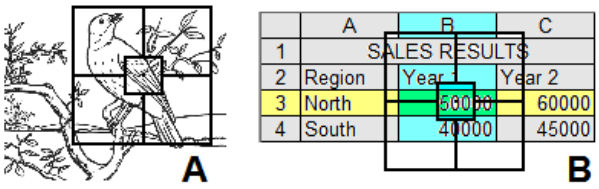

Figure 3. *Items at different levels of view in two scenarios.*

A blind user can rapidly navigate between such levels, for example by using a mouse wheel or "dial" (E & F) Fig 9, while hearing the Focus effects speaking the Focus level (e.g. cell, column, row, or block) that is currently emphasised, and at the same time being made aware of the levels above and below the current level of view, which have distinguishing effects applied (voice character etc., and optionally echo and/or reverberation).

The cocktail party effect [18] helps users to focus their auditory attention on the item emphasised by the system, or switch their attention to another item that is also presented but not emphasised. They can then cause the system to highlight that other item instead.

Initial tests (and previous work [19]) show that the cocktail party effect works best as a stereophonic or binaural effect i.e. with speech stereophonically separated (with voice character, pitch, etc. also contributing). However as the several levels being presented will typically be co-located or in close proximity (A) Fig 2, the system can artificially separate the

items in soundspace i.e. both in pitch and left-right stereophonic positioning (B) Fig 2, so that the cocktail party effect is maximized. Deliberately spreading out (i.e. relocating) the voices in soundspace is not as confusing as might be expected, as the currently-emphasised subject of attention is mapped to its unadjusted corresponding location via pitch and left-right stereophonic positioning, and the relocated de-emphasised effects are identified as such via their audio properties (particularly voice character), and by their apparent locations (e.g. in the corners of the audio display (B) Fig 2).

Focus effects can also be used to present property values of non-visual and non-spatial properties, for example levels of categorisation and analysis, as found in many academic fields. Some perceptual and cognitive models, and some social science models use 3- or 4-level models [20, 21], and these could be presented using Focus Effects. For example the Dewey Decimal classification system [22] could be presented and navigated round using Focus Effects, as described in section 2.4.3 below.

2.2. Producing Multi-talker effects

The system is implemented mainly in Microsoft Visual Basic, and runs on a standard Windows PC. The open source library OpenCV [23] is used to perform computer vision tasks such as face recognition, optical flow motion detection, and Camshift tracking; and the open source engine Tesseract [24] is used to perform optical character recognition (OCR). The force-feedback Logitech mouse (A) Fig 9 and Microsoft Sidewinder joystick (B) are controlled via Microsoft's DirectInput methods.

The audio is primarily speech-like. For earlier versions of the system, a limited number of words would be presented, for example colours and certain recognised items such as faces or motion, and recorded speech samples were used. For the current version, any words may need to be spoken, so Windows SAPI Text-to-Speech synthesis (TTS) [25] output is saved to a standard sound (.WAV) file, which can then be pitched and panned on replay as and when required (using Microsoft's DirectSound [26] SetFrequency and SetPan methods).

It was advantageous to use an even-level voice for the main talker voice (most modern TTS voices speak with considerable intonation/prosody present). The eSpeak [27] open source SAPI speech synthesizer software is used for the main talker voice, as it can be set to produce a flat voice output, and is therefore more suitable for conveying the pitch-to-height mapping. Other TTS voices can be used for the secondary focus effect voices, as they are generally stationary and not attempting to convey precise location through pitch.

When multiple voices are speaking, the voices can be differentiated via:- voice character of the speaker (sex, accent, etc.); pitch; left-right pan positioning; volume; special effects such as echo, reverberation, "flange", "gargle", etc.; and speaker start time offset.

Typically the main talker voice will move to convey location and shape, while the extra voices, presenting the additional information, will be located in fixed positions, for example near the corners of the audio display (B) Fig 2.

One useful feature is to "flip" the location of the extra voices if the main voice gets too near to them in pitch or pan separation. For example if an extra voice is located in the top left corner of the audio display, as the main talker voice moves left, when it gets to within ¼ of a screen-width of the left edge, the extra/secondary voice panning is flipped to the centre of the audio display, and later flips back to the left edge as the main talker voice moves back towards the centre. A similar effect is performed with the pitch of the extra/secondary voices as the main voice moves in the in the vertical direction.

2.3. Visual-domain processing, and client-domain views

In the visual domain, the system can produce higher-level consolidations of image content. The filter GUI Fig 4 allows users to select the Level 3 (D) categories of basic visual items that they want to have presented e.g. Reds (A), Faces (B), OCR Text (C) etc.; and to select higher-level (Level 2 up to Level 0) group item consolidations (D, E, F, and G), as described below.

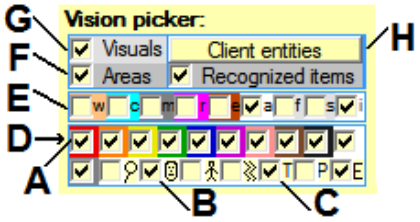

Figure 4. *The filter GUI for selecting visual item categories.*

The system performs standard computer vision processing, reducing the image (A) Fig 5 to a set of "blobs" (B) both of areas of particular properties e.g. colours (C), and recognised items such as faces (D), or text (E). These are referred to as "basic items". The system can then consolidate the blobs into higher-level items, referred to as "group items". For example from e.g. Level 4 individual coloured blobs and recognised items (e.g. Red 2, Face 3, Text 1 etc.) the system can consolidate to Level 3 groupings (e.g. Reds, Faces, etc.) (D) Fig 4, to Level 2 (e.g. monochrome areas, "rainbow"/spectral-coloured areas, found items etc.) (E), to Level 1 (Areas of colour, and Recognized group items) (F), and to a single "Level 0" group item for the all items in the visual image (G). The Level 0 item identifies the type of entity and domain view (e.g. general visuals domain view), and can be switched to and from other entities (H) that may be available, and that may use a client-domain view, as described below.

Figure 5. *Computer vision processing of an image, including extracting face and text data.*

Furthermore, bespoke combinations of properties can be specified for particular tasks. For example for highlighting red tomatoes, specifying the bespoke combination of colours "Red or Pink or Orange or Purple" will generally produce clearer tomato-shaped blobs, as they cover the range of found shades.

Additionally, cascaded items can be produced from basic items, and are at lower levels. For example if a face is detected, then standard facial features can also be deduced from a standard library face that includes e.g. a Level 5 feature Eyes, Level 6 Left eye, Level 7 Iris etc. Such levels and items can be interacted with in the same way as for higher-level items.

While HFVE knows how to consolidate general images, it does not know about other domains such as, for example, Excel spreadsheets. Instead such entities can be submitted to HFVE as client entities, for HFVE to present.

For example consider the spreadsheet (A) Fig 7. Although it could be presented as a visual-domain view i.e. as a series of patches of colour and perhaps some text recognition, it is more meaningful to be able to inspect it via a spreadsheet-domain view (B), consolidating cells (Level 4) to columns and rows (Level 3), then to individual blocks (and objects such as charts and pictures) (Level 2), then to all blocks (and all objects) (Level 1), then to top level Spreadsheet (Level 0).

Such higher-level view groupings facilitate obtaining meaningful summaries/overviews of content, and help with navigating around the items of the image/entity.

The system can use an in-box folder, into which client applications can deposit entity files for potential presentation. The system can then process and present all, or some, of them (or none), and can optionally delete them after presentation.

As well as presenting the content of the current item, the system can also present items etc. via the extra talkers, for example:- a) higher- and/or lower- level items; or b) adjacent items or nearby items; or other properties; and these arrangements can be per entity type, with a default arrangement being used for entities whose type is not recognised.

2.4. Interfacing to external entities

In order to present externally-processed images and other entity types via HFVE, a straightforward interfacing method has been devised. This comprises submitting a standard 24-bit colour bitmap (.BMP) file Fig 7 that includes all of the required basic item blobs (referred to as the "ItemMap" file); and a standard text (.TXT) file Fig 6 that describes how those blobs are marked via particular bit settings on the bitmap, and specifies how those blobs are consolidated to higher-level items (referred to as the "ItemKey" file). This pair of files, that fully describes the blobs of the image/entity, and how they are consolidated, can be created manually using a simple image painting application and a text editor, or can be created via an external application.

```
4,b2,Face 1|Pink,$200 $30200
4,b3,Face 2|Pink,$10000 $30200
...
4,b10,Blue 1,$4 $107
4,b14,Blue 2,$100 $107
...
3,g19,Blues,b10 b14
...
3,g24,Faces,b2 b3 b4 b5 b6
```

A B

Figure 6. *Part of an "ItemKey" file describing the blob bits, and how they are consolidated into higher-level group items.*

For more complex entities some blobs may overlap (for example faces and colour blobs Fig 5), and the system can reserve a certain number of bits in the 24-bit bitmap for particular sets of non-overlapping blobs. Such content is resolved by the ItemKey text file Fig 6, which specifies which bits are significant (B), and their values (A) for particular items.

2.4.1 Interfacing to a spreadsheet

For the Spreadsheet entity example described above, it would be an arduous task for someone to mark-up all of the cells and objects of an Excel spreadsheet, and then create a text file describing them. Instead an Excel Add-In has been developed, which can be triggered for typical Excel spreadsheets. It paints corresponding rectangles etc. equal in size to each filled cell or object (graph, chart, picture etc.) (B) Fig 7, each such item having a unique colour shade. The add-in also produces a corresponding ItemKey text file that describes the content of each blob, with one line for each item, and details of consolidations for columns, rows, blocks etc.

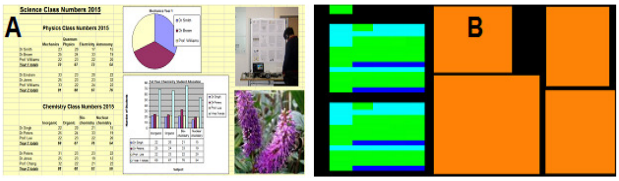

Figure 7. *A spreadsheet and the corresponding "ItemMap".*

A snapshot of the spreadsheet (A) Fig 7 is taken, and merged with the ItemMap bitmap (B) (the marker bits use the least significant bits of the 24-bit colour bitmap, and their presence is typically invisible to sighted users).

HFVE does not know about Excel, but processes the resultant pair of files like any other, getting item identifier bits from the ItemMap bitmap pixels, then getting the corresponding item details (e.g. words to speak) from the ItemKey text file.

2.4.2 Interfacing to other client entities

The interface has proved to be versatile, and many different client application-created entities, or manually-created entities, can be submitted using it. Many client applications such as movie players (with or without specially marked-up items), graph and charting applications, and drawing applications, can pass item information to the interface, for presentation via the system's audio (and tactile) effects.

It is not always necessary to submit both an ItemMap and ItemKey. The ItemKey text file content can be directly added to the end of the bitmap file (which will still be presentable as a standard image file), and can later be separated by the system.

Alternatively, one of either of the files can be used to create the other, as illustrated in the following two examples:-

2.4.3 Pseudo-visual representations

Non-visual multi-level/structured entities may be presented as pseudo-visual/spatial representations. For example for the Dewey Decimal classification system [22] the levels might be Level 1 Class (e.g. 500 / Science & Maths) – Level 2 Division (e.g. 510 / Maths) – Level 3 Section (e.g. 516 / Geometry) – Level 4 Sub-section (e.g. 516.3 / Analytic Geometry) (with Level 0 giving the entity/domain view name). The lowest level items i.e. Sub-sections can be automatically marked on a bitmap as block patterns of rectangles, each of a unique colour shade, which can then be consolidated up through the levels to the higher-level group items in the same manner as is done for standard visual entities. Then when presented as audio (and tactile) effects, the user can obtain an impression of the size and distribution of the items at each level of the entity.

This can be achieved by initially counting the lower level items that comprise each group item, then splitting the "pseudo-image" into rectangular areas each sized according to the basic item count for the group items at Level 1 (i.e. Class), then within each such rectangular area splitting further according to the next level content, until a pattern of similar-sized small

rectangles representing the basic items is produced, grouped according to their higher-level classifications.

In use, the user can freely move the pointer to find a higher-level group item, lock on it, and then explore the lower level items within that item. In this way a spatial/dimensional impression of a non-visual entity can be achieved.

2.4.4 OCR-read key/legend

A simple bitmap comprising a few coloured areas could just be presented as coloured areas. Alternatively, a simple key/legend can be included on the bitmap, in which the meaning of each colour shade is written next to a patch of that particular shade (A) Fig 8. OCR can recognise the legend text, then the system can link the text to the shade, to give it meaning, allowing the bitmap alone to be presented meaningfully to the user : the system can build a small ItemKey file based on the text and adjacent shades. Higher-level group items can be included by writing the higher-level terms next to patches containing the several shades that represent the basic items that comprise the higher-level items (B). (The topmost non-key/legend wording (C) is assumed to be the title / Level 0 entity name.)

The user can then access the map as if it was set up as a standard pair of text and bitmap files, hearing meaningful terms.

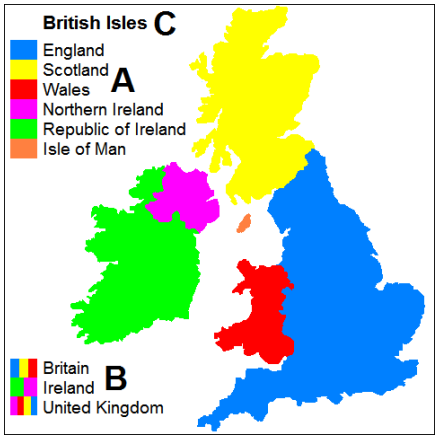

Figure 8. *OCR-read key/legend describing the blob shades.*

2.5. Using multi-level, multi-talker Focus effects

In use, there are three main ways that the user typically accesses the entity being presented, and they can be used concurrently.

Pointer : The user can freely move a pointer (e.g. via mouse or touch) over the items in the rectangular area of the entity image (which can occupy the entire computer monitor area). The system presents the item (according to the current level of view) that the pointer is over at any time (represented by a basic item blob or a consolidated group item blob or blobs). Optionally the system can present an audio (and/or tactile) cue when the pointer crosses the border between two items. At any moment in time the user can lock on the item being presented. (There is also a mode which presents the underlying pixel colour, with no blob consolidation performed.)

In addition to the spoken information, a pitched and panned buzzing sound conveys the location of the pointer within the image area, which, as previously reported, greatly improves the perception of shape and location [16]. An additional tracer, of differing timbre, can convey distance information (if available) via pitch. Alternatively, the pitch of either the standard speech or standard buzzing sound can convey distance information, with the other conveying height. (A similar approach can be used for presenting distances for shape tracers and polytracers.)

Automatic : The user can command the system to automatically step around the items found within a user-sizable and user-moveable frame Fig 3, which can follow the pointer (the frame can encompass the entire image area). The system attempts to pick the most important items given the current level of view and other settings, and this can depend on activity. The user can at any time lock on the item being presented.

Search : The user can type the name of an item (basic item or group item) into a search box and the system then locks on it.

Filters can be used to control which categories of items are presented, for example via the vision filter GUI Fig 4.

2.5.1 Locked-on items

Once an item is locked on, the subsequent interaction depends to some extent on the equipment being used to access the entity.

Force-feedback : If a force-feedback mouse (A) Fig 9 or joystick (B) is being used, the system can restrict the free movement to the area(s) of the current item – when pushed by the user away from the item, a spring force will attempt to push the mouse or joystick handle back to the centre or nearest part of the selected item (or to the point at which they left the blob). When within the area of the item, the mouse or joystick handle will be loose/"floppy" and can be moved freely. The user can feel around the edge of the item, and get audio feedback as well. (Alternatively the user can command the force-feedback device to perform an audiotactile tracer of the item's outline, with corners emphasised, as was previously available.)

If the item is multi-blob, e.g. a group item or fragmented basic item, then the user can command a jump to the next blob, then explore that shape and content. Alternatively, with a force-feedback device the user can simply push the handle around the image and it will tend to "snap" to the nearest applicable blob.

Mouse : If a standard mouse is being used, an audio cue can signify and warn that the user has attempted to leave the area of the item. However the cursor pointer can be locked at the edge of the item (via a Windows SetCursorPos action), so that the user does not need to find the item again and can simply move their mouse back in the opposite direction. In this way the user can gain an impression of the extent of the item (as well as from the other audio effects that are presenting it).

Touch : If a touch-screen, or an absolute mode touch-pad, is being used, then the system cannot easily restrict the physical movement of the user's finger, so needs to directly tell the user or give non-speech cues to indicate how to move back to the locked item area. However users will typically be better able to recall the approximate location of the item within the physical fixed area of the touch-screen or touch-pad, than when using a standard relative mode mouse.

Obtaining shapes for mouse and touch access : The user can get an immediate impression of the locations and shapes of the locked-on items or group items via sound – section 3.2 below describes using a mouse or touch device to perform a drag following a coded tap or click sequence, and this can command the system to move an audio (and tactile) shape tracer around the blob perimeter via one of the following approaches:-

a) The audio tracer's position in its path around the perimeter of the item or items at any time can correspond to the distance of the drag from its start point. Hence by dragging back and forth the user can move the tracer correspondingly back and

forth along the perimeter, and so get an impression of the shape, size and extent, and location, of the items. The system measures the distance from the initial vertical or horizontal location, so that the drag does not need to return to the exact start spot.

b) The user can keep moving the tracer forwards around the perimeter by constantly moving the drag in any direction. They can reverse the drag direction to cause the tracer to reverse.

Both tracers and imprints can be presented, and either can move forwards or backwards, and present the current item, or all items in an item group. The type and combination of effects can be signified via combinations of:- the initial direction of drag (up, down, left, right, etc.); the screen quadrant or screen half that the drag starts in; and the direction of circular motion (clockwise or anticlockwise) of a rotational drag.

Additionally a mouse wheel or dial (E & F) Fig 9 can control the movement of the tracer, in a similar manner.

2.5.2 Navigating with locked-on items

When an item is locked on, and the user moves the pointer within the area of the item, typically the items at *lower*-levels than the locked item are presented – the user will normally know which item is locked on (via an earlier announcement), and so is instead told about the lower-level items that they are currently moving over, and that comprise the locked-on item.

The items above and/or below the item being presented can also be presented at the same time via multi-talker focus effects, so that the user can be aware of items in adjacent layers (or items nearby on the same layer), and can switch to being locked on one of them. For example, if locked on a spreadsheet column (B) Fig 3, the main voice can present the cell being moved over, at the same time as which two of the extra focus effect voices can present the column and row respectively in which the cell is located (and optionally a third voice could present the block containing the cell, column and row). As these extra voices are typically re-located at the corners of the audio display area (B) Fig 2, it is straightforward for the user to indicate which of these items to switch the lock to if required.

The user can also command the system to switch to any level of view above or below the current item; and if appropriate automatically step round the items below (or above, or adjacent to) the current item in the levels of view. They can then switch the locked item to be any of the listed items, so that directly pointing at particular items in the image is not required.

Figure 9. *Logitech's Wingman Force Feedback Mouse (A), Microsoft's Sidewinder Force Feedback 2 joystick (B), an "MMO" mouse (C), Gyration's Air Mouse (D), 3Dconnexion's Space Navigator (E) and Contour Design's Shuttle (F) "dials".*

2.5.3 Multiple properties and item types

In the visual domain, an image can be presented via several types of property, for example colour, distance, texture, the nature of recognised items, etc., and the user could select which of these to present. However they might also wish to be aware of several property types and consolidations at the same time.

(B) Fig 8 shows an example of basic blobs (countries) which could be consolidated in two ways (as geographical islands, and via political grouping). Similarly the cells of a spreadsheet (B) Fig 3 can be consolidated into columns, and/or rows, both of which are on the same level of view.

Some users would want to follow only one or two extra talker voices Fig 2. One simple approach to presenting several different items, even if in separate entity views (e.g. visual and spreadsheet), via a limited number of extra talkers, is to get each talker to present several items, or properties, in sequence.

To resolve and simplify the presentation and navigation of multiple properties and classification/grouping methods, the following approach can be used:-

i) In order that a client application can request presentation of more than one property type or item at the same time, the client can specify which extra voice should present each property or item when not being presented via the main voice, and so keep separate, if required, particular types of item. For the Excel example, the column details, and row details, can each be directed to separate voices (via a field in the ItemKey file).

ii) The system can then inspect the various items to be presented, and direct selected items to particular extra voices, speaking them in sequence. Optionally the system can apply varying focus effects (e.g. reverberation effects) if required; and can temporarily alter the apparent position of the extra talkers, for example to reflect the relative location of the items.

iii) The user can navigate between items, properties, and entities, by selecting them when their corresponding words are spoken the talkers. Alternatively the user can indicate the ordinal of the required item within a spoken list of items. With either method, that item then becomes the locked-on item.

In this way, the system can stream information to separate speaker channels, allowing the user to be simultaneously aware of several entities, and related items and properties.

3. INTERACTION

Methods of interacting with the system have previously been described [17], and Fig 9 illustrates several less common interaction devices, namely a force-feedback mouse (A), a force-feedback joystick (B), a "MMO" mouse with 12 extra programmable buttons (C), an air mouse (D), a "dial" controller (E), and a dial controller with 15 programmable buttons (F). Pen input, voice input, touch-screen and touch-pad, as well as standard mouse and keyboard control, can also be used.

3.1. Ordered control

One effective approach is to have up to 48 ordered control actions available via, for example, the numeric keys located along the top of a standard "QWERTY" keyboard, plus the two following keys (typically "-"/minus and "="/equals) totalling 12 keys. These 12 keys can be combined with two modifier keys, e.g. Control and Shift, giving a total of 48 possible command actions. Such an arrangement can be operated via a numeric keypad, or via a touch- or mouse-operated on-screen grid

("OSG") Fig 10, where the elements can be arranged 4x4 (A), or arranged around the image area (B), with combinations of the lockable Ctrl- and Shift- keys modifying the function of the 12 command keys. An "MMO" mouse with 12 extra programmable buttons (C) Fig 9 could also be used for this purpose.

3.2. Tapping and other control methods

One effective method of commanding the system is to tap Morse-like commands onto a touch-screen or touch-pad i.e. combinations of short and long taps. The three possible modifier key combinations (Ctrl-, Shift-, and Ctrl+Shift-) can also be signified on a mouse or touch-screen or touch-pad, by the user doing a single long click or tap; a short then long click or tap; or two long clicks or taps; followed by up to 12 short taps for the appropriate 1 to 12 command.

This was found to be straightforward to perform, though if necessary an extra modifier key can be used to reduce the maximum number of short taps to six. Similarly a combination of short and long taps can precede a drag across the touch-screen or touch-pad, for example to specify an area for tracking, a section of the image to zoom into, to pan a zoomed-in image, and to perform the shape inspection described in 2.5.1 above.

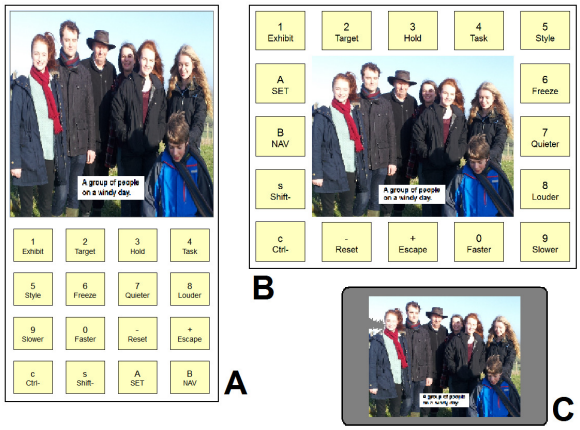

Figure 10. *Main image and on-screen grid arrangements.*

The same 48 ordered control actions can alternatively be triggered by gestures, multiple mouse clicks, etc. If gestures are used, simple swipes in the direction of the hour markers of a standard clock face can represent the numbers 1 to 12. The Air Mouse (D) Fig 9 could be used for this purpose.

3.3. Touch control

If a touch-screen tablet is being used (for example a Windows tablet), then the whole screen area can show the image being presented. The user can tap commands and drag over the computer monitor area, and touch the tablet screen to indicate parts of the image (C) Fig 10. Alternatively the screen can be split so that some of it is occupied by the image monitor and some of it by the commanding on-screen grid (A & B).

Blind users can slide their finger over the on-screen grid (a process known as "scrubbing"), with speech feedback informing them of the key that they are over at any moment, so that they can navigate to the required command, whereupon they can raise their finger in order to select that command.

One possible arrangement is to default to presenting only the image area (C) Fig 10, which is tapped to give common commands, and swiped to indicate areas of the image, and on a particular command the image is replaced with an array of command buttons for less-common instructions.

All of the above touch-based interaction methods were found to be effective to a degree, and a user can decide which approach is most appropriate for them, or they can use a combination of the methods.

A blind person is unable to benefit from seeing the OSG or the image being presented on a tablet computer's touch screen. Instead, a touch pad, as often found on laptop computers, may be used to control the system via taps and drags/swipes in the same manner. If a Synaptics® TouchPad® is available and set to absolute mode, it can be used to indicate locations within the Monitor and OSG, and to trigger touch screen-style tap and gesture commands.

4. INFORMAL ASSESSMENT SESSIONS

It was important to obtain assessments of the approaches described in this paper. "AB" (not his real initials), who has been totally blind since birth, and "CD" (not her real initials), who is sighted, participated in informal feedback and evaluation sessions, especially of the new multi-talker focus effects. AB has considerable prior knowledge and experience of computer access for blind people, and also assessed the system prior to ISon 2013 [17]. In free-format discussion sessions, the approaches were demonstrated, and the pros and cons considered.

Note that not all of the described features of the system are fully operational, nor were they when the assessments took place, particularly the lock on feature was not complete.

It is intended that further evaluation will be occur after the ISon 2016 workshop (see below).

AB found the "multi talker" feature promising, though preferred the different voice characters, and the separation in pan and pitch, to the echo and reverberation effects. The latter made the speech unclear, and AB thought that such effects might best be reserved for conveying particular information.

He found the "location flip" feature effective, whereby additional talkers are temporarily repositioned from their stationary location (in their left-right pan location, and pitch) in order to maintain separation from the primary/key talker if it moves close to them. AB also thought that the simple panning method for spatial voice separation was adequate (panning requires low processor load and allows multi-point effects such as Imprints and Polytracers to run smoothly on a regular PC).

An Excel spreadsheet demo allowed AB to experience freely moving over the area of the spreadsheet with a pointer, with varying client-domain levels of view.

AB suggested mapping might be a suitable application for the system. (This was previously done to demonstrate shape tracers, and the developer is currently trialling using multi-level and multi-talker focus effects to explore a political map of the type described in section 2.4.4 above.)

We discussed the interaction methods. AB was interested in using a laptop touchpad as a controller, and this feature is currently implemented for Synaptics TouchPads, and makes available the standard touch-screen actions, including multi-tap combinations, long and short taps, and "tap and drag" (e.g. to zoom, scroll, or select an area).

CD particularly preferred using tapping codes to command the system when in Tablet Style, and found these intuitive to use. Like AB, she had reservations about the echo and

reverberation effects, and similarly preferred using the other properties (voice character, pitch, pan position and volume) to differentiate the speakers. CD thought that three simultaneous voices was the most that she could comfortably listen to. She also found the "location flip" feature effective in keeping the voices separated in soundspace.

The lock on feature was demonstrated but was not fully functional, however CD liked the method of locking the mouse cursor for use with standard mice, and the technique of navigating by picking one of the extra Focus effect voices i.e. the one presenting the required next item.

The key feedback from these initial assessment sessions was that adding special effects such as echo and reverb was generally distracting and should be reserved for conveying special information (e.g. perhaps mild amounts can be reserved to identify and distinguish distance-presenting extra voices from level-of-view-presenting extra voices). Instead the other properties described should generally be used to identify particular voices and keep the voices separated.

5. CONCLUSIONS AND FUTURE WORK

Multi-talker focus effects are a way for blind people to gain information about the visual content of a scene, and, when combined with multi-level representations of visual scenes (and other entities), and the previously reported methods, allow a blind person to access several aspects of visual images. The initial results and feedback are encouraging, and indicate that the approach is worth progressing.

One possible future line of development is in presenting data from the Internet, for example interfacing with mapping data that is available online. Additionally, online artificial intelligence (AI) systems may be used in order to perform more accurate object recognition etc. For example basic face and blob detection can be provided standalone as previously described, yet when an Internet connection is available then more sophisticated processing can be provided – for example emotion detection is being developed, and this could also be presented.

Furthermore, online facilities exist to provide words summarising the content of images, so providing a top-level summary term for visual images [28].

Future work should include a detailed evaluation, with an examination of specific tasks and interaction approaches, detailed statistical analysis of results, and a qualitative analysis of post task interview data.

The authors intend to perform such an evaluation for inclusion in a subsequent paper that may be submitted to an ISon-related special issue.

The system's current state of development will be demonstrated at ISon 2016.

6. REFERENCES

[1] World Health Organization, "Visual impairment and blindness" in Fact Sheet No. 282, Updated October 2013, http://www.who.int/mediacentre/factsheets/fs282/en/.

[2] E. E. Fournier d'Albe, "On a Type-Reading Optophone" in *Proc. Royal Society of London. Series A*, Vol. 90, No. 619 (Jul. 1, 1914), pp. 373-375.

[3] P.B.L. Meijer, "An Experimental System for Auditory Image Representations" in *IEEE Trans on Biomedical Engineering*, Vol. 39, No. 2, pp. 112-121, 1992.

[4] U.S. Patent No. US 6,963,656 B1.

[5] D.L. Mansur, M.M. Blattner and K.I. Joy, ``Sound Graphs, A Numerical Data Analysis Method for the Blind," in *Journal of Medical Systems*, Vol. 9, pp. 163-174, 1985.

[6] A. Edwards, "Auditory Display in Assistive Technology" in *The Sonification Handbook*, T. Hermann, A. Hunt, J.G. Neuhoff (Eds.) 2011, pp. 431-453.

[7] T. Pun et al., "Image and Video Processing for Visually Handicapped People" in *EURASIP Journal on Image and Video Processing*, Vol. 2007, Article ID 25214, 2007.

[8] Roth P, Richoz D, Petrucci L, Pun T., "An audio-haptic tool for non-visual image representation" in *Proceedings of the Sixth International Symposium on Signal Processing and its Applications* 2001 (Cat.No.01EX467) : 64-7.

[9] Patrick Roth, Thierry Pun: "Design and Evaluation of Multimodal System for the Non-visual Exploration of Digital Pictures". In *Proceedings of INTERACT 2003.*

[10] Parente, P. and G. Bishop. BATS: The Blind Audio Tactile Mapping System. ACMSE. Savannah, GA. March 2003.

[11] Kopeček, I and Ošlejšek, R. "GATE to Accessibility of Computer Graphics" in *Computers Helping People with Special Needs: 11th International Conference*, ICCHP 2008. Berlin: Springer-Verlag, pp. 295-302, 2008.

[12] Kopeček, I and Ošlejšek, R. "Hybrid Approach to Sonification of Color Images" in Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technologies. Los Alamitos: IEEE Computer Society, pp. 722-727, 2008.

[13] Derek Brock, Christina Wasylyshyn, and Brian McClimens, "Word spotting in a multichannel virtual auditory display at normal and accelerated rates of speech" in *Proc. of 22nd International Conference on Auditory Display (ICAD-2016),* Canberra, Australia, 2016.

[14] *The HFVE system,* http://www.hfve.com.

[15] D. Dewhurst, "Accessing Audiotactile Images with HFVE Silooet" in *Proc. Fourth Int. Workshop on Haptic and Audio Interaction Design,* Springer-Verlag, 2009.

[16] D. Dewhurst, "Creating and Accessing Audiotactile Images With "HFVE" Vision Substitution Software" in *Proc. of ISon 2010, 3rd Interactive Sonification Workshop*, KTH, Stockholm, Sweden, 2010.

[17] D. Dewhurst, "Using "Imprints" to Summarise Accessible Images" in *Proc. of ISon 2013, 4th Interactive Sonification Workshop*, Fraunhofer IIS, Erlangen, Germany, 2013.

[18] *Cocktail party effect,* https://en.wikipedia.org/wiki/Cocktail_party_effect.

[19] Hawley ML, Litovsky RY, Culling JF. "The benefit of binaural hearing in a cocktail party: effect of location and type of interferer" in *J. Acoust. Soc. Am.,* Vol. 115, No. 2, February 2004.

[20] *Intentional stance,* https://en.wikipedia.org/wiki/Intentional_stance.

[21] *Level of analysis,* https://en.wikipedia.org/wiki/Level_of_analysis.

[22] *Dewey Decimal Classification,* https://en.wikipedia.org/wiki/Dewey_Decimal_Classification.

[23] *OpenCV(Open Source Computer Vision)*, http://opencv.org

[24] *Tesseract,* https://github.com/tesseract-ocr/tesseract/wiki.

[25] *Microsoft Speech API*, https://en.wikipedia.org/wiki/Microsoft_Speech_API.

[26] *DirectSound*, https://en.wikipedia.org/wiki/DirectSound.

[27] *eSpeak text to speech*, http://espeak.sourceforge.net.

[28] *IBM Watson Visual Recognition service,* https://www.ibm.com /watson/developercloud/doc/visual-recognition.