Diary in the Sky: A Spatial Audio Display for a Mobile Calendar

Ashley Walker, Stephen Brewster, David McGookin & Adrian Ng

Glasgow Interactive Systems Group, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK

Tel: *+44 141 330 4966*

Email: *stephen@dcs.glasgow.ac.uk*

URL: *www.dcs.glasgow.ac.uk/~stephen*

We present a spatial audio display technique that overcomes the presentation rate bottleneck of traditional monaural audio displays. Our compact speech display works by encoding message semantics into the acoustic spatialisation. In user testing, this display facilitated better recall of events than a conventional small screen visual display. Moreover, results showed that this mapping aided in the recall of the absolute position of events — as opposed to merely their relative orders — in a temporally ordered data set.

Keywords: planar 3D audio, non-speech sounds, mobile devices, PDAs spatial mapping, earcons, sound, interface sonification.

1 Introduction

We live in a visual culture. The great worth of pictorial representations are reaffirmed over and over in the achievements of the 20th century — on billboards and the big screen, by graphic artists mousing over digital canvases, and school children clicking at multimedia PCs. As technology evolves in the 21st century, however, we may begin to see things differently. At least, we are likely to see things through a smaller display. As the mobile and miniature devices of the new millennium replace older forms of communication and computation, the fabric of our visual culture must stretch to accommodate other display and interaction techniques.

Concerns about the limits of the visual display are not new. We have known for decades that visual representations of information, including graphics and written

text, can be hard to read — causing 'eye-strain' and visual overload. This is particularly true in multi-tasking computer interfaces involving many windows of information. Moreover, people who interact with information 'on the go' — via the small screen of a personal digital assistant (PDA) or mobile phone — have further reduced visual (and attentional) resources. Mobile phone displays, in particular, have a small fraction of the pixel display space of desktop monitors and they are employed in use-contexts that are themselves visually intensive. One place in which we can seek display alternatives — alternatives not wed to the diminishing resource of screen space — is in the audio domain.

Sonic displays have been developed in a number of special purpose application areas (Gaver, 1989; Schmandt & Mullins, 1995; Kobayashi & Schmandt, 1997; Crease & Brewster, 1998; Mynatt et al., 1998; Sawhney & Schmandt, 1999; Walker & Brewster, 2000). Where these have succeeded, they have been based upon a firm understanding of hearing. Where audio displays have failed, they have naively attempted to translate a (visual) stream of information into an audio one — ignoring important differences between how the eye and ear process information.

The ear differs from the eye in that it is omni-directional — a true three-dimensional (3D) display space that does not suffer from occlusion. Its fabric is course grained — with angular resolutions approximately 10 times more coarse than the eye across the sensorially richest regions (Howard & Templeton, 1966). However, what the ear lacks in spatially sensitivity, it more than makes up for in temporal sensitivity (try watching an action movie with the sound turned off: the eye rarely perceives a punch land, but the ear satisfies your need to know that justice has been served).

The ear analyses information temporally; and this is both its strength and weakness. Audio displays involving speech are often dismissed as too slow because of the supposed delay involved with rending a stream of text (or, worse, a description of a graphic). A simple translation of information into a single audio stream, however, fails to exploit the third important strength of the ear: its ability to simultaneously monitor more than one stream of information (Cherry, 1953; Arons, 1992). Compare the layers of audio in a movie soundtrack — including music, dialogue, ambient sounds, auditory feedback from footsteps, dripping faucets, gun-shots, etc.— with the single layer of visual information on the screen (Chion, 1990).

Here we present a novel audio display technique that overcomes the presentation rate bottleneck of traditional monaural audio displays by spatialising audio streams. Moreover, we encode message semantics into the spatialisation to further increase presentation rate. This display technique is general and may be used in a variety of application interfaces. To test the utility, however, we built a prototype that encompassed the same functionality as a popular mobile device application: the *DateBook*, or calendar. This paper covers design and implementation (Section 2) of that prototype, as well as user-testing (Section 3 and 4). In Section 5 and 6, we draw together insights.

2 Materials

Here we describe an experiment to investigate the usability of an auditory interface to a DateBook application. We chose to work with the DateBook Application in

Figure 1: Palm DateBook visual display.

particular because the inherent temporal separation of the data (i.e. DateBook events) is naturally amenable to the mappings described below. We chose to compare our audio DateBook interface with a model of the visual DateBook interface that runs on Palm, Inc. PDAs because of the wide user-base of the latter (Palm, Inc.) — at the present time these are the most popular PDAs on the market.

The Palm is a small and light PDA with a 6cm×6cm rectangular screen (see Figure 1). Most applications — including the DateBook — present their data vertically in scrollable lists. In the case of the DateBook, events are typically displayed in 1 hourly denominations in a long vertical list. Due to the screen size limitations, approximately half a day's worth of events is typically visible at one time. Scrolling between events, however, requires some visual attention, due to the problem of mating the tip of a stylus with the small scroll bar area.

These limitations are inherent in a small screen device and can only be overcome via alternative display techniques. Figure 2 shows a mapping of DateBook events onto an alternative audio display space. In this space, an imaginary clock-face is projected onto a slice of the auditory sphere surrounding a user's head — with 9am/pm as the extreme left, 12am/pm as the direct front, 3am/pm as the extreme right and 6am/pm as the direct back. The mapping is displayed within the horizontal plane containing a listener's ears because this is the most sensorally rich listening region (Begault, 1994).

We hypothesised that this horizontal ('clock-face') display orientation was more natural — exploiting existing knowledge of time-space mappings — than a vertical list of time-ordered data or a stream of non-spatialised audio items. Given the limitations of current spatialisation technology it is much harder to make a solution for the general population that works well in azimuth (due to pinnae differences between listeners); the transverse plane is much easier to work with for a general solution. In particular, we hypothesised that such a clock-face display would facilitate better recall of events and incur lower workloads. A description of hypothesis testing is given in the next section.

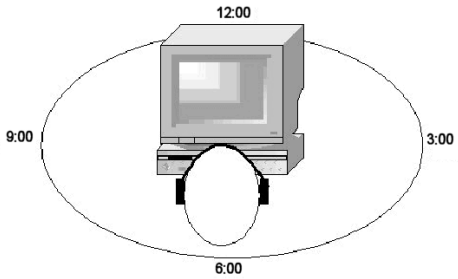

Figure 2: Time-space mapping of the auditory DateBook display.

3 Methods

3.1 Experimental Design

Sixteen students from University of Glasgow served as participants. This group comprised six women and ten men between the ages of 18 and 24. Participants were paid. The experiment was a counter-balanced within-groups design with modality of cue as the single independent variable. Each of the participants performed the task described below using a visual display and the spatial audio display. Dependent variables included recall performance (of four diary items) and several subjective workload measures. Participants also gave informal feedback following the experiment.

3.2 Experimental Scenario

Users were presented with DateBook contents and told that they would be asked to perform a series of recall questions about the day's events. Events consisted of simple keyword phrases of 4 words or less, preceded by a verbal or written time stamp. Care was taken to ensure that the semantics did not overlap with those in a participant's real-life by telling participants that they were seeing/hearing the diaries of a variety of professionals (e.g. Surgeon, Reporter, Circus Clown, etc.). After being exposed to each condition as described below, the display was hidden/muted and the participant was asked four recall questions — one per calendar item. The responses were verbally cued and tested relative recall of item order ("Did A occur before or after B?") as well as absolute recall of the temporal ordering of items ("What time did X occur?" or "What occurred at Y time?"). The questions were equated in that they sampled the entire temporal window in an effort to control for memory order effects that could favour recall of primacy/recency items over items from the middle of the list. Following this recall test, workload ratings were collected. Each participant performed the experiment three times and the order of presentation of each modality was varied.

In the visual condition users were presented with the interface to the standard Palm DateBook. As the Palm is the biggest selling PDA its diary application is one

of the most commonly used. Diary applications on many other PDAs follow a very similar list-based design so we took this as the control condition. Participants were allowed to scroll between events over a period of 8 seconds (this period was chosen to correspond with the two seconds per event playback scheme used in the audio condition). Events were vertically separated by space proportional to their temporal separations. Users had to scroll between events, as they did not all fit on one screen. Because the Palm device does not yet support audio of the type required by this study, the experiment ran in a 6×6cm rectangular window on the screen of a desktop computer. Participants scrolled between events using a standard desktop mouse. To make the audio and visual conditions consistent both used this desktop simulation.

In the audio condition events were speech synthesised sequentially using Lucent's Text-to-Speech technology (Lucent Technologies, 1999) in intervals of two seconds (none of the audio cues lasted more than 1.5seconds). In this condition events were spatialised via convolution with head-related transfer functions (HRTFs) included with Microsoft's Direct X multimedia API and a Creative Labs SoundBlaster Live! Platinum soundcard. Events were not preceded by a verbal time stamp, as that information was available in the semantics of the spatial audio mapping. The sounds were presented through a pair of Sennheiser HD25 headphones. There was no visual display in this condition.

3.3 Measures

Recall performance was calculated using the percentage correct in each condition, as well as an intra-condition performance comparison of absolute vs. relative event knowledge. Subjective workload assessments — on a modified set of NASA TLX scales (Hart & Staveland, 1988) — were collected after each condition. The workload ratings included mental and physical demand, time pressure, effort expended, frustration and performance. We added a seventh factor: Annoyance. This is one of the main concerns that users of auditory interfaces have with the use of sound. In the experiment described here annoyance due to auditory feedback was measured to find out if it was indeed a problem. We also asked our participants to indicate overall preference, i.e. which of the two interfaces they felt made the task easiest. Participants had to fill in workload charts after both conditions of the experiment.

4 Results

4.1 Recall

Recall rates were significantly affected by modality, with users performing better in the audio than the visual condition ($T_{15} = 4.49$, $p = 0.0002$). The mean percentage of correct recalls was 88.3% and 70.2% in the audio and visual conditions respectively.

Analysis of a subset of the data showed that, in both conditions, recall of the absolute time of an event was worse than recall of an event's order relative to other events. However, in the case of absolute event time recall, the mean percentage of correct recalls dropped more markedly in the visual than the audio conditions (84.4% and 64.6% in the audio and visual conditions, respectively).

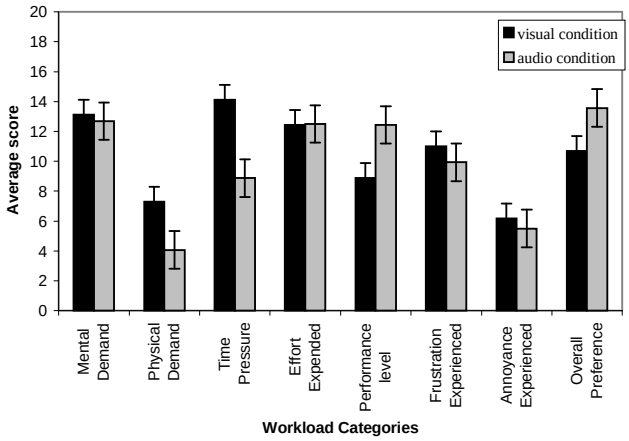

Figure 3: Average workload scores including annoyance and overall preference. Standard error bars are shown.

4.2 Subject Workload Ratings

The audio condition resulted in significantly lower subjective workload ratings in three of the six workload categories measured. Participants reported that the audio condition caused significantly less physical demand ($T_{15} = 2.31$, $p = 0.018$) and time pressure ($T_{15} = 3.97$, $p = 0.0006$). This ultimately resulted in a significantly higher sense of performance in the audio condition ($T_{15} = 2.98$; $p = 0.005$). As shown in Figure 3, the other workload ratings are fairly equal across the conditions. The audio condition was not rated as more annoying than the visual.

5 Discussion

5.1 Spatial Axes

Many participants said that the audio condition required less ‘interpretation’. Some participants explained that event time came ‘for free’, thereby reducing the task to memorising a keyword associated with an easily recalled spatial landmark. By contrast, the visual condition required the memorisation of ‘two things’: time and event. This insight could be valuable in visual layout design as well. Although one could argue that a list of items is spatially extended, the vertical extendedness of such a space does not inherently encode the relevant semantics of this task as effectively as a horizontal, clock-face space.

5.2 Audio

On top of advantages associated with an audio DateBook display, a number of participants also said that audio, *per se*, registered more automatically (“like someone telling you what to do”). Many participants admitted to feeling surprised that an apparently visual task could be performed without much effort using the auditory cues.

Although most participants found that heard (as opposed to read) events were easier to memorise/recall in a verbally cued recall test, it is not necessarily the case that a well-designed written recall test would elicit the same response.

5.3 Serial vs. Parallel Presentation

The audio and visual conditions differed in the presentation bandwidth. In the visual condition, users could see several events at once; while, in the audio, events were presented serially. We expected that this would frustrate audio users, but only a minority complained that the audio was presented too fast or that they only had 'one chance' to memorise the audio events.

In pilot studies, we broadcast spatial audio events in parallel and tracked users' head movements — adjusting the volume (audibility) of each event depending upon listening behaviour. However, this volume control mechanisms appeared to be too crude and listeners felt overloaded. Certainly more training in this technique — or more sophisticated volume control (Schmandt & Mullins, 1995) — could yield better results.

5.4 Active and Passive Displays

The audio and visual conditions differed in the degree of interactivity. While we expected users to prefer some control over the display, a few complained about losing time scrolling through the visual list. Recall that users experienced significantly greater time pressure and physical demand in the visual condition. Moreover, one participant highlighted a potential difficulty: although the event list was sorted according to time of occurrence, the order of events appeared reversed when he scrolled back up through it. Again, this may be more a reflection of the difficulty of scrolling as opposed to a commentary on control in general. Good audio interaction techniques — i.e. input devices symmetrical to the 3D display space — are certainly worth pursuing in future work.

5.5 Laboratory vs. Mobile Interface Testing

Applications with great potential to enhance mobile device interfaces should be tested on those devices to confirm that they work as well in the field as they do in the lab. We regret that we did not have the technology to conduct a field study in this case. From other fieldwork on mobile audio, e.g. Brewster (2001), we expect that the following factors would bear on the same study conducted there.

First, it is well established that device interface tasks with a high visual load and manual input requirement are difficult to perform when walking or driving. In this regard, we expect results of a field study to further accentuate the suitability of audio (indeed, such a belief motivated this study). Brewster (2001) showed that a real sonically-enhanced Palm PDA significantly improved usability and mobility in a real mobile situation. It is not possible to do more sophisticated audio interfaces on the Palm platform as it has only basic audio capabilities. The next stage of our work in this area is to use a wearable PC with soundcard that will allow us to run the type of experiment discussed here on a real mobile device in a real mobile situation.

Nevertheless, there remains an open question as to how the introduction of additional audio stimuli, on top of environmental sounds, will impact on the

perception of those sounds and vice versa. We performed the present study in a laboratory where there was no environmental sound targeted at the participant. The number and content of environmental plus artificial/display streams competing for a user's attention appears, to us, to be a more relevant issue than 'lab' vs. 'field' noise. This issue certainly requires more sophisticated and systematic study than any field test report available in the literature.

5.6 Conclusions

We consume audio and visual information in different ways, and different techniques are required for displaying to the ear and eye. GUI design, in collaboration with visual display hardware, has evolved to exploit the narrow-field, high resolution, space-scanning proclivity of the eye. Here we presented a simple and effective audio display device that exploits the ear's omni-directionality, its sensitivity to coarse-scale spatialisation, and its temporal sensitivity. In doing so, we constructed a display that overcomes some of the bandwidth limitations of traditional (monaural) speech displays and provided a more effective interface to an existing PDA application.

Acknowledgements

This work was funded by EPSRC GR/L79212 and a New Discoveries grant from Microsoft Corporation.

References

Arons, B. (1992), "A Review of the Cocktail Party Effect", *Journal of the American Voice I/O Society* **12**(7), 35–50.

Begault, D. (1994), *3-D Sound for Virtual Reality and Multimedia*, Academic Press.

Brewster, S. A. (2001), Overcoming the Lack of Screen Space on Mobile Computers, Technical Report TR-2001-87, Department of Computer Science, Glasgow University, Glasgow, UK.

Cherry, E. C. (1953), "Some Experiments on the Recognition of Speech", *Journal of the Acoustical Society of America* **25**, 975–9.

Chion, M. (1990), *Audio-Vision: Sound on Screen*, Columbia University Press.

Crease, M. & Brewster, S. (1998), Making Progress With Sounds — The Design And Evaluation Of An Audio Progress Bar, *in* A. Edwards & S. Brewster (eds.), *Proceedings of the International Conference on Auditory Display (ICAD'98)*, BCS. http://www.ewic.org.uk/.

Gaver, W. W. (1989), "The SonicFinder: An Interface that Uses Auditory Icon", *Human–Computer Interaction* **4**(1), 67–94.

Hart, S. & Staveland, L. (1988), Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, *in* P. Hancock & N. Meshkati (eds.), *Human Mental Workload*, North-Holland, pp.139–83.

Howard, I. P. & Templeton, W. B. (1966), *Human Spatial Orientation*, John Wiley & Sons.

Kobayashi, M. & Schmandt, C. (1997), Dynamic Soundscape: Mapping Time to Space for Audio Browsing, *in* S. Pemberton (ed.), *Proceedings of CHI'97: Human Factors in Computing Systems*, ACM Press, pp.194–201.

Lucent Technologies (1999), Lucent Speech Solutions, http:// www.lucent.com/speech.

Mynatt, E. D., Back, M., Want, R., Baer, M. & Ellis, J. (1998), Designing Audio Aura, *in* C.-M. Karat, A. Lund, J. Coutaz & J. Karat (eds.), *Proceedings of CHI'98: Human Factors in Computing Systems*, ACM Press, pp.566–73.

Sawhney, N. & Schmandt, C. (1999), Nomadic Radio: Scalable and Contextual Notification for Wearable Messaging, *in* M. G. Williams, M. W. Altom, K. Ehrlich & W. Newman (eds.), *Proceedings of the CHI99 Conference on Human Factors in Computing Systems: The CHI is the Limit*, ACM Press, pp.96–103.

Schmandt, C. & Mullins, A. (1995), AudioStreamer: Exploiting Simultaneity for Listening, *in* I. Katz, R. Mack & L. Marks (eds.), *Companion Proceedings of CHI'95: Human Factors in Computing Systems (CHI'95 Conference Companion)*, ACM Press, pp.218–9.

Walker, V. A. & Brewster, S. (2000), "Spatial Audio in Small Screen Device Displays,", *Personal Technologies* **4**(2), 144–54.

Author Index

Keyword Index