Designing Spatial Audio Interfaces to Support Multiple Audio Streams

Yolanda Vazquez-Alvarez and Stephen A. Brewster
Glasgow Interactive Systems Group, Department of Computing Science
University of Glasgow, Glasgow, G12 8QQ, UK
+44 (0) 141 330 8430

{yolanda, stephen}@dcs.gla.ac.uk www.gaime-project.org

ABSTRACT
Auditory interfaces offer a solution to the problem of effective eyes-free mobile interactions. However, a problem with audio, as opposed to visual displays, is dealing with multiple simultaneous outputs. Any audio interface needs to consider: 1) simultaneous versus sequential presentation of multiple audio streams, 2) 3D audio techniques to place sounds in different spatial locations versus a single point of presentation, 3) dynamic movement versus fixed locations of audio sources. We present an experiment using a divided-attention task where a continuous podcast and an audio menu compete for attention. A sequential presentation baseline assessed the impact of cognitive load, and as expected, dividing attention had a significant effect on overall performance. However, spatial audio still increased the users' ability to attend to two streams, while dynamic movement of streams led to higher perceived workload. These results will provide guidelines for designers when building eyes-free auditory interfaces for mobile applications.

Categories and Subject Descriptors
H.5.2. User Interfaces: Interaction styles.

General Terms
Design, Experimentation, Human Factors.

Keywords
Auditory interfaces, mobile systems, spatial audio, multiple audio streams, divided-attention task.

1. INTRODUCTION AND BACKGROUND
Auditory interfaces are used for interaction in mobile environments when access to the visual display can be too distracting or might not even be a possibility. Often, mobile users may wish to perform multiple tasks simultaneously, for instance, checking the time of the next bus while listening to a favourite team playing the most important soccer game of the season, or checking the name of a music track while on the move without having to stop the music. In an eyes-free environment the effective presentation of multiple audio streams can facilitate such multitasking [1]. The experiment presented in this paper focuses on the most challenging multi-stream environment, where the user is required to attend to both streams (divided-attention task). Our goals were to:

1. Determine to what extent users can successfully operate an audio menu while attending to a second audio stream, and to show whether spatial and audio minimization can mitigate performance loss compared to simple sequential presentation.
2. Given recent interest in spatial audio techniques and eyes-free interactions, provide some useful guidelines to designers who are required to implement interactions with simultaneous audio streams in a mobile environment.

We can present multiple and simultaneous information streams in an auditory form either sequentially or simultaneously. Presenting them sequentially will prevent information sources from competing with each other but this could result in a more lengthy interaction when switching between sources, poorer recall of earlier information, and irritation caused by continuous interruption. The Cocktail Party effect [2] provides evidence that humans can, in fact, monitor several audio streams simultaneously, selectively focusing attention on any one and placing the rest in the background.

Although audio is often seen as a single stream coming from a fixed point, if users are wearing headphones, 3D audio techniques can create the perception that a sound is coming from a specific spatial location [3]. Previous work has investigated such spatial audio techniques to present multiple streams [4,5]. Just as the visual field can be used to present information to the user, a 3D sound field can be used to differentiate between information sources. By modelling the filter based on the transfer function between the sound source located at certain positions and the eardrums of a listener (the Head Related Transfer Function, HRTF), it is possible to position audio effectively all around a user. A spatial representation of the auditory display provides orientational information that aids segregation and attention switching between the audio streams to maintain intelligibility when auditory information is being used [6,7]. However, it is less clear how 3D audio techniques might be implemented in an interactive environment, where we need to consider how to manage multiple audio streams without overloading the user.

The use of 3D audio also raises the issue of how streams are presented in the 3D sound field over time. Not only can audio appear to come from a specific position, this position can be dynamically moved. The movement of items in visual interfaces is commonly used to enhance the interface. For example, animating a window

MobileHCI 2010 September 7 - 10, 2010, Lisboa, Portugal.
ACM 978-1-60558-835-3.

as it is minimized. Such techniques can also be used in an audio interface. For example, moving an audio stream to the side (we will term this spatial minimization), while a second stream is played from the front.

To evaluate potential techniques for an interface using multiple audio streams three factors should be considered: sequential versus simultaneous sound presentation; fixed point versus spatial audio; and static versus moving audio presentation. While these factors have been suggested in previous research, to our knowledge, they have not been evaluated formally against each other. We set up a divided-attention task experiment in which users were presented with a continuous audio stream as well as an audio menu task. By using a divided-attention task we hoped users would perceive higher workload levels, which in turn would help us assess usability of simultaneous sound presentation when under cognitive load. Our research questions were: 1. Can users maintain coherent attention on dual audio streams in a mobile interface? 2. What 3D audio techniques can be used to alter focus on the streams and move them from foreground to background and *vice versa*? 3. How efficient and usable is such an interface?

2. EXPERIMENTAL DESIGN

2.1 Stimuli

There were two audio streams in this experiment. One continuous and the other user activated. The continuous stream was a podcast selected from the BBC radio programme 'From our own correspondent'. It was mono, 16-bit and sampled at 16 kHz, approximately 3 minutes long and with a male speaker. A total of 5 different podcasts were selected (1 training session + 4 different conditions). These podcasts were chosen because they all shared similar topics and are narrated in a similar journalistic format.

The user-activated audio stream was a hierarchical audio menu (see Figure 1a). The audio menu items were synthesized using Cereproc's[1] British English female RP voice. All prompts were mono, 16-bit and sampled at 16 kHz. The audio items in the menu were different for all conditions. We used the Amplify filter in Audacity[2] to normalize the volume of both the podcast and all the audio menu items to 70% of the audio dynamic range, which equals to a normal conversation typically 60-70dB.

Figure 1. (a) Audio-menu structure (b) Experimental setup.

2.2 Tasks

For each condition, participants performed 3 different tasks using the hierarchical audio menu: 1) Finding the next track title, 2) Checking an appointment for Tuesday, and 3) Finding the current time. The audio menu was presented at 0° azimuth (in front of the user's nose) and always at a distance of 1m in the frontal horizontal plane.

2.3 Conditions

There were four conditions in the experiment varying sound location and continuous vs. interrupted presentation:

1. *Baseline*: The podcast was paused or *interrupted* while the participant carried out the audio menu tasks and then resumed after the tasks were completed. Podcast and audio menu were both located at the origin (0° azimuth) and at a distance of 1m in the frontal horizontal plane.

2. *Concurrent*: The podcast was playing while the participant carried out the audio menu tasks. Podcast and audio menu were located at the origin (0° azimuth) and at a distance of 1m in the frontal horizontal plane.

3. *User-activated spatial minimization*: The podcast was located at the origin (0° azimuth) 1m away from the listener in the frontal horizontal plane (see Figure 2a), and moved to the right hand-side (90° azimuth) only when the participant was engaged in the audio menu tasks (see Figure 2b). We based our decision for moving the podcast from the origin to the right hand-side on our previous evaluation results [8]. This specific location showed less variation in the localization perception by listeners. The volume level of the podcast was attenuated by approximately -10 dB by moving the source to the right hand-side (-3 dB intensity drop) and doubling the perceived distance by placing it 2m away from the listener. Listeners have been shown to perform best when monitoring an audio stream at -10 dB, compared to lower levels [7].

4. *Fixed spatial minimization*: The podcast was fixed to 90° azimuth and 2m away from the listener for the entire duration of this condition. The audio menu was at 0° azimuth. The audio streams were presented continuously.

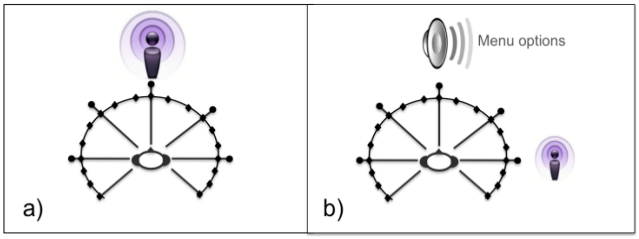

Figure 2. (a) Single continuous stream. Black filled circles show different azimuth locations for the audio streams 1m away from the listener. (b) Stream moved to the background – from front to right – and perceived distance of the source increased, as user interacts with foreground spatialized audio menu.

2.4 Methodology

There were 24 participants (12 male, 12 female, aged 19 to 53 years). We balanced the number of males and females to control for gender effects as it is sometimes said that females are better than men at multitasking. All subjects were native speakers of British English, reported normal hearing and were right-handed. We used a within-subjects design and conditions were presented in a randomized order to control for ordering effects. Podcasts were always presented in the same order. All conditions were tested in a static lab environment where users were seated on a

[1] www.cereproc.com

[2] http://audacity.sourceforge.net/

chair holding the mobile phone in an upright position wearing a pair of headphones (see Figure 1b). The experiment was run on a Nokia N95 8GB [9] using the built-in Head Related Transfer Functions (HRTFs) and the JAVA JSR-234 Advanced Multimedia Supplements API [10] to position the audio sources. The audio was played over a pair of DT770 PRO – 250 OHM Beyerdynamic headphones.

The experiment consisted of two training sessions followed by the four different conditions. First, a training session was exclusively devoted to familiarizing the participants with the audio menu structure. There was no time constraint and they were monitored and guided by the experimenter via a separate set of earphones connected to the mobile device using a headphone splitter. The second training session used the concurrent condition in order to familiarize participants with a continuously streamed podcast while interacting with the audio menu while performing the tasks.

In all four experimental conditions, participants started listening to a podcast and after approximately 1 minute, the user was prompted with a 25-ms sine wave beep at 1500 Hz to start interacting with the menu and complete the three tasks described above in any order. To initiate this interaction the participant pressed the central navigation key on the phone. The arrow keys on the phone were used to browse the menu items. Once the tasks were completed and the audio menu was exited by pressing the central navigation key, the user continued listening to the podcast until it was over. The participant was instructed to monitor the continuous podcast. After the end of the podcast for each condition, the participant was asked to answer a set of six questions as in Stifelman's study [6]. Five of the questions requested information that was located at evenly spaced points in time over the length of the podcast to confirm the participant had paid attention to it. The last question requested information about one of the menu tasks. Following the recall questions, participants were asked to complete a NASA-TLX subjective workload assessment [11]. NASA-TLX is a well validated multi-dimensional rating scale designed to obtain workload estimates from one or more operators while they are performing a task or immediately afterwards. After all four conditions were completed, participants were instructed to rank all of the conditions in order of preference: '1' preferred the most, and '4' preferred the least. This experiment took approximately 45 minutes in total and participants were allowed to rest between conditions.

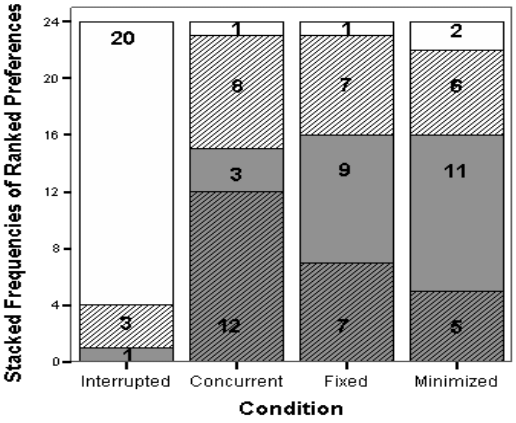

Figure 3. Frequencies stacked bar chart showing ranked preferences per condition. Preferences show: 1- most preferred (solid white), 2 - (striped/white), 3 - (solid grey) & 4 - least preferred (striped/grey).

3. RESULTS

3.1 Ranked Preferences

Figure 3 shows a stacked count for the order of preference for each of the conditions. A non-parametric Friedman test showed a significant main effect for condition type ($\chi 2$ =32.650, df=3, $p< 0.001$, N=24). Pair-wise Wilcoxon signed ranks tests showed that the interrupted condition, as expected, was significantly the most preferred over the other conditions ($p < 0.001$) due to the reduced cognitive load. When audio streams were presented simultaneously, the spatialization of the audio sources was preferred over the concurrent presentation.

3.2 Overall Workload

Raw overall workload means were calculated from the NASA-TLX questionnaires completed after each condition (see Figure 4). A by-subjects repeated measures ANOVA on overall workload means grouped by gender and order of conditions showed a significant main effect for condition type ($F(3,48) = 15.651$, $p < 0.001$). There was no main effect for gender or ordering type and no interactions. *Post hoc* Tukey HSD tests with Bonferroni correction for condition type showed that the divided-attention task increased perceived workload levels. These levels were not significant amongst the simultaneous conditions.

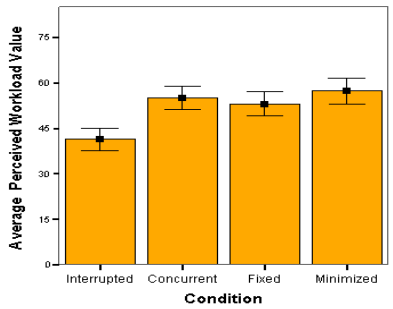

Figure 4. Average scores for perceived workload from NASA-TLX questionnaires. All error bars show Standard Error of Mean ± 1.0.

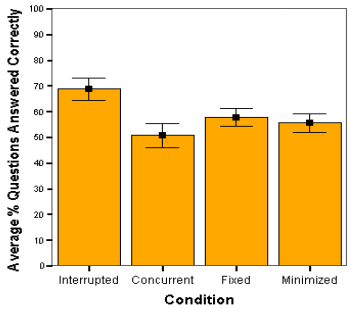

Figure 5. Average percentage of correct answers per condition.

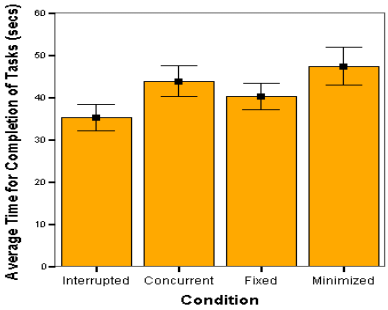

Figure 6. Mean task completion times for audio menu tasks per condition.

3.3 Performance

3.3.1 Recall performance

Recall performance was calculated using the percentage correct of answers in each condition (see Figure 5). A by-subjects repeated measures ANOVA on recall performance means grouped by gender and order of conditions showed a significant main effect for condition type ($F(3,48)=5.109$, $p<0.010$). There was no main effect for gender or ordering type and no interactions. *Post hoc* Tukey HSD tests with Bonferroni correction for condition type showed that the spatially fixed ($p < 0.050$) and concurrent conditions ($p < 0.050$) suffered a significant performance drop caused by the cognitive load.

3.3.2 Task completion times

Total time taken to complete the audio menu tasks was also computed (see Figure 6). A by-subjects repeated measures ANOVA on task completion mean times grouped by gender and order of conditions showed a significant main effect for condition type ($F(3,48)=4.997$, $p<0.010$). There was no main effect for gender or ordering type and no interactions.

Post hoc Tukey HSD tests with Bonferroni correction for condition type showed that the pattern seen in recall performance was repeated for task completion time with the concurrent condition ($p < 0.050$) having performed significantly worse than the baseline.

4. DISCUSSION AND CONCLUSIONS

Results showed the divided-attention task experiment presented in this paper increased users' cognitive load in simultaneous conditions. In these situations one could expect a drop of performance (in our study from 70% to 50% recall and an increase in task time from 35.32 to 47.43 secs).

Amongst the simultaneous conditions, there is a tendency for spatial audio to improve recall against single point presentation, for the user-activated spatial minimization to increase workload (possibly because participants found the movement distracting "and felt it was more of an adjustment to pay attention to the podcast when it moved"), and a trend for participants preferring the spatially fixed and minimized conditions over the other simultaneous condition. Further work is required to investigate these effects. A study will be carried out where the continuous audio stream does not require active monitoring, for example listening to music. Removing the factor of cognitive load could produce contrasting results, and, for many applications, may be closer to a realistic design. When in the design phase of an auditory interface, the cognitive load caused by divided attention should be assessed. If the decrease in performance is unacceptable (and the interaction allows it), we would recommend the use of an alternative approach such as a buffer and catch-up method [12], thus avoiding simultaneous presentation of audio streams.

Addressing our research questions, we can state the following. Users are able to attend to simultaneous audio streams but they will experience a rise in perceived workload and a drop in performance, which although larger than we initially expected, is not so large as to make the approach unusable. Spatial audio and movement of the perceived location may be a promising approach when making simultaneous audio presentation more usable but not so effective when the user is under high cognitive load. In addition, sudden movement of audio streams may be distracting, and simultaneous presentation can affect performance even after the simultaneous presentation is complete. Further work is required to establish how 3D audio techniques might be effectively applied within an application, but results suggest these approaches may still offer advantages for certain designs.

The techniques presented in this paper have been suggested in previous research but, to our knowledge, they have never been evaluated formally against each other. The results presented here are important for the community to better understand the affordances of different delivery mechanisms for simultaneous presentation in auditory multitasking scenarios, and they quantify the performance of different potential approaches. We have shown that the use of 3D audio techniques in interfaces requires care and understanding of the users' cognitive load.

5. ACKNOWLEDGMENTS

This work was jointly funded by Nokia and EPSRC research grant EP/F023405.

6. REFERENCES

[1] Li, K. A., Baudisch, P. and Hinckley, K. 2008. Blindsight: eyes-free access to mobile phones. In *Proc. CHI 2008* (Florence, Italy, 2008). ACM Press, 1389-1398.

[2] Cherry, E. C. 1953. Some experiments on the recognition of speech, with one and with two ears. Journal of Acoustical Society of America 25(5), (1953) 975-979.

[3] Begault, D.R. 1994. 3D Sound for Virtual Reality and Multimedia. Academic Press Cambridge.

[4] Schmandt, C., Mullins, A. 1995. AudioStreamer: exploiting simultaneity for listening. In *Proc. CHI 1995*, ACM Press (1995), 218-219.

[5] Walker, A., Brewster, S.A., McGookin, D. and Ng, A. 2001. Diary in the sky: A spatial audio display for a mobile calendar. In *Proc. BCS IHM-HCI 2001*, Springer (2001), 531-540.

[6] Stifelman, L.J. 1994. The cocktail party effect in auditory interfaces: a study of simultaneous presentation. MIT Media Laboratory Technical Report.

[7] Ihlefeld, A. and Shinn-Cunningham, B. G. 2008. Spatial release from energetic and informational masking in a divided speech identification task. *Journal of the Acoustic Society of America,* 123 (2008) 4380-4392.

[8] Vazquez-Alvarez, Y. and Brewster, S.A. 2009. Investigating Background & Foreground Interactions Using Spatial Audio Cues. In *Proc. CHI 2009*, ACM Press (2009), 3823-3828.

[9] www.nseries.com/index.html#l=products,n95_8gb

[10] http://theoreticlabs.com/dev/api/jsr-234/javax/microedition/amms/package-summary.html

[11] Hart, S. G. and Staveland, L. E. 1988. Development of Nasa Tlx (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload* (1988), 139-183.

[12] Dietz, P.H. and Yerazunis, W.S. 2001. Real-Time Audio Buffering for Telephone Applications. In Proceedings of UIST: ACM Symposium on User Interface Software and Technology (Orlando FL USA, November 2001), ACM Press, 193-194.