Investigating Background & Foreground Interactions Using Spatial Audio Cues

Yolanda Vazquez-Alvarez and Stephen Brewster
Glasgow Interactive Systems Group, Department of Computing Science
University of Glasgow, Glasgow, G12 8QQ, UK
{yolanda, stephen}@dcs.gla.ac.uk

www.gaime–project.org

CHI 2009, April 4 – April 9, 2009, Boston, MA, USA
ACM 978-1-60558-247-4/08/04.

Abstract

Audio is a key feedback mechanism in eyes-free and mobile computer interaction. Spatial audio, which allows us to localize a sound source in a 3D space, can offer a means of altering focus between audio streams as well as increasing the richness and differentiation of audio cues. However, the implementation of spatial audio on mobile phones is a recent development. Therefore, a calibration of this new technology is a requirement for any further spatial audio research. In this paper we report an evaluation of the spatial audio capabilities supported on a Nokia N95 8GB mobile phone. Participants were able to significantly discriminate between five audio sources on the frontal horizontal plane. Results also highlighted possible subject variation caused by earedness and handedness. We then introduce the concept of audio minimization and describe work in progress using the Nokia N95's 3D audio capability to implement and evaluate audio minimization in an eyes-free mobile environment.

Keywords

3D audio, evaluation, audio cues, background and foreground interactions, multiple audio streams.

ACM Classification Keywords
H5.2. User Interfaces: Interaction styles, evaluation.

Introduction

Auditory interaction methods are important in the context of non-visual interaction ("eyes-free"), where audio is the primary output. For instance, in a mobile environment users cannot always look at the screen of a device as they need to keep their attention on their surrounding environment [1]. Due to the increasing functionality of mobile phones, e.g. web browsers, there is a requirement for more complex audio-driven eyes-free interactions. Within such complex audio interfaces, spatial audio could help mitigate audio overload by differentiating sound sources to reproduce a visual display, teasing apart simultaneous sounds.

Central to this paper is the exploration of the usability of background and foreground interactions using audio cues. Buxton defines foreground interaction as "activities which are in the fore of human consciousness – intentional activities" and background interaction as "tasks that take place in the periphery – 'behind' those in the foreground" [2]. The application of Buxton's background and foreground model has been limited to sensing techniques and feature modeling and combination to detect true interactions and avoid spurious ones. In the section devoted to audio minimization, we will focus on a more traditional view of foreground and background perception where two audio streams, one offering a user-driven audio menu, and a second providing continuous streamed audio information, compete for attention. The audio menu is controlled by inputs such as gesture or more conventional key presses while at the same time managing a constant separate stream of audio information. Being able to maintain multiple interactive streams is attractive because it can potentially maintain coherence; offer increased efficiency of use, and potentially improve usability because it reflects the configuration of many everyday interactions.

Much of the previous work on mobile spatial audio interfaces in HCI has been done on laptops or more powerful devices but now, with increases in processing power, it is becoming possible to do spatial audio on mobile phones. However, it is not clear what level of localization accuracy can be achieved given the limitations of such devices. Spatial audio APIs are available on various platforms such as Vodafone (VFX Specification), NTT Docomo, JAVA JSR-234 Advanced Multimedia Supplements (AMMS), and Open SL ES. A set of Head Related Transfer Functions (HRTF's) [3] is typically used by 3D audio controls in these APIs to allow an accurate localization of sound in 3D space. If a monaural sound signal is passed through these filters and heard through headphones, the listener will hear a sound that seems to come from a particular location in space. However, these API's contain generic HRTF's (as measurements of the ears of all listeners is not possible) and this reduces the localization accuracy, potentially making auditory interfaces using them ineffective. To add to this problem, we lack any information on the type of HRTF data set used in these proprietary APIs, so we are not aware of the possible localization errors that originate from the implementation itself. Therefore, we need to investigate the possibilities of these mobile spatial audio systems to see if they are capable of supporting the types of interfaces that have been developed in the past and may be created in the future.

Not only are there significant differences in the implementation of spatial audio on different devices, there is

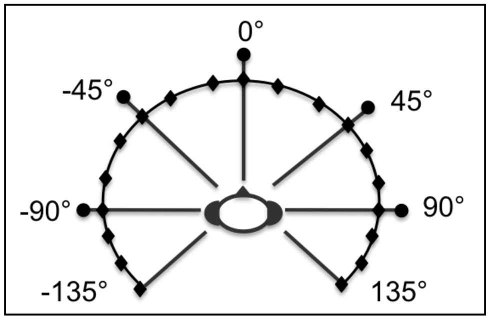

Figure 1. Evaluation setup. The black filled circles represent the different azimuth locations of the static sources placed at 100 mm from the listener. The inner circle with diamonds shows the trajectory of the acoustic pointer placed at 85 mm from the listener.

also evidence that users may experience spatial audio differently depending on their left/right auditory dominance (earedness). Many previous studies have found a hemispherical asymmetry in human perception and localization of spatial sounds, indicating a dominant role of the right hemisphere auditory areas [4]. In some people, left and right dominance may be reversed or may even be in both hemispheres. Also, hand dominance (handedness) is reported to be the opposite from the dominant hemisphere, i.e. left-handers are right-hemisphere dominant. However, this correlation varies. This hemispherical dominance variation could affect the localization accuracy of sounds.

Before we can proceed to test how effective and usable spatialized foreground and background interactions can be, we first carried out an evaluation of the positional 3D audio controls on the Nokia N95 8GB to determine if these controls would be accurate enough for the implementation of the audio minimization study.

Evaluation of localization accuracy

An evaluation of the AMMS 3D audio location controls on the Nokia N95 [5] was needed to investigate what level of localization accuracy listeners could achieve. We used the HRTFs and API of the AMMS 3D audio location controls [6] to position sounds at arbitrary points around the user. In addition, we also controlled for differences between noise and speech and auditory dominance, as this could have a critical effect on spatial audio perception.

Experimental design

An auditory pointer adjustment program, which allowed listeners to adjust an auditory pointer to the same direction of a static auditory source, was developed using the 3D audio capabilities offered in the AMMS API. The methodology used in this study replicates the one from Pulkki and Hirvonen [7] to evaluate an apparatus for auditory pointer adjustment and its localization accuracy in an eight-channel and 5.1 loudspeaker setups. This method will help us test to what extent listeners are able to discriminate the auditory sources as originating from different locations. It was found that humans generate errors and bias when interpreting auditory perception with any method [3]. However, when listeners compare two auditory perceptions, and adjusting the auditory pointer direction until there is no perceived difference in the direction between the pointer and the static sources, fewer errors and biases occur.

Evaluation setup

Twelve listeners matched the auditory pointer direction with single static sources in directions [0° (directly in front of the nose), 45°, 90°, -45°, -90°] and elevation 0° (see Figure 1). All static sources were placed in the front 180°, as it has been found to be the area of most accurate perception of direction [8]. All five directions used in this evaluation formed part of Pulkki and Hirvonen's study and so we will be able to compare the results from both studies. The experiment consisted of a training session followed by two different conditions. In one of the conditions, the static sources emitted pink-noise (a noise signal that contains all frequencies with equal energy per octave, commonly used to test loudspeakers [9]). The pink-noise source was 500 ms with a 50 ms fade-in and fade-out. In the other condition, the static sources emitted recorded speech, using the phrase "One head-line in Britain today", taken from a BBC podcast. This is the type of audio source we will use in the minimization study. The speech source was 1500 ms long. Both pink-noise and speech static

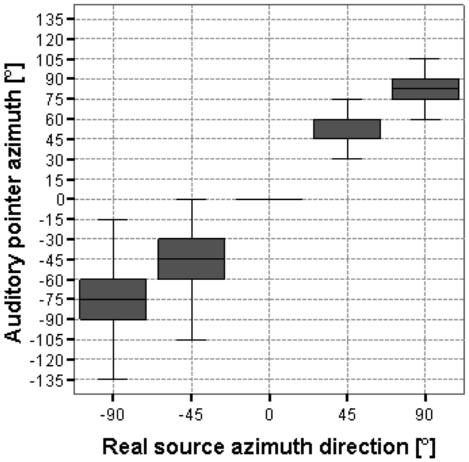

Figure 2. Box plots showing the localization accuracy achieved in the evaluation study.

sources were mono, 16-bit and sampled at 16 kHz. The order of the conditions was randomized per participant to control for ordering effects.

The acoustic pointer was a source placed closer to the listener (85 mm) and the static sources were placed further away (100 mm), as in Pulkki and Hirvonen's study. The acoustic pointer was always identical to the corresponding static source per trial for the given condition, be it pink-noise or speech. A 250 ms gap was inserted between the target sound and the pointer sound. The participants were able to move the pointer in 15° increments by using the left or right keys on a Nokia N95 8GB mobile phone.

Evaluation procedure

Participants were seated on a chair, holding the mobile phone in an upright position and wearing a pair of DT770 PRO – 250 OHM Beyerdynamic headphones. The participants were mostly students at Glasgow University, ten males and two females aged between 23 and 35 who were paid £3 for their participation. All participants were asked to report their dominant hand and ear, right, left or mixed, answering simple handedness and earedness questions [10]. None of the participants was excluded based on their handedness or earedness results and none of them reported a hearing deficiency.

The static source and the pointer signal were played once, one after another, every time a key was pressed to move the acoustic pointer left or right. Once the listener adjusted the pointer to the same direction as the static source, the central navigation key on the phone was pressed to indicate the adjustment was complete. After this, the location of the auditory pointer was recorded and a spoken prompt saying 'next' was played to introduce the next stimulus. The auditory pointer's starting position was initially random but for the following trial it was set to be the last position recorded in the previous trial.

The test was organized so that both the pink-noise and the speech condition contained a total of 15 trials (five azimuth directions x 3 repetitions of each stimulus type) with 3 trials of each stimulus type in the training session. Each trial took approximately one minute. Sessions took less than 30 minutes in total and participants were allowed to rest between conditions. The trials were presented in randomized order for each session.

Results

The deviation of the acoustic pointer adjustment from the direction of the target source was recorded. A three-way between-subjects ANOVA was performed comparing the different static source azimuth directions, type of stimuli and earedness. The results showed a significant main effect for the different static source azimuth directions (F (4,340)= 317.753, p<0.001). *Post hoc* Tukey HSD comparisons indicated that static source azimuth direction -90° (M=-81.00), -45° (M=-51.55), 0° (M=1.07), 45° (M=53.00) and 90° (M=85.71) were all perceived as being significantly different locations, (p<0.001). Figure 2 presents the acoustic pointer data across participants.

There was a main effect for the different stimuli type: speech and pink-noise (F (1,340)=4.065, p<0.05) showing that participants were better at localizing pink-noise than speech, especially on the left side of 0°; earedness, i.e. left and right ear dominance (F (1,34)= 3.889, p<0.05) showing that right-eared participants were more accurate than left-eared ones; and a two-

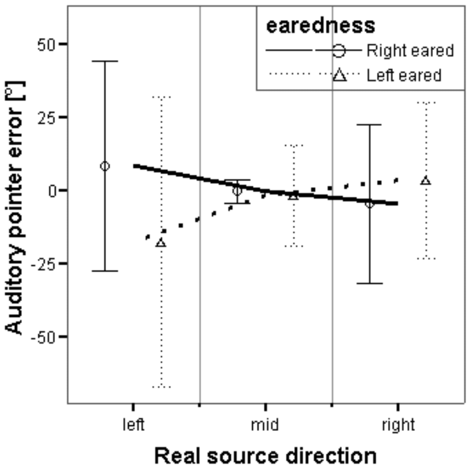

Figure 3. Signed error per static source direction and ear dominance. Error bars show ± 1.0 SD.

way interaction between earedness and the different static source azimuth directions (F (4,340)=5.469, p<0.001). We could conclude from these results that both ear dominance and the type of stimuli would be important factors influencing spatial audio localization.

However, the results from our only left-handed participant contained a high number of outliers. If this subject is removed earedness, stimuli type and the interaction, stop being significant. Figure 3 shows the different signed error means by earedness grouped by left (azimuths -90°, -45°), mid (azimuths 0°), left (azimuths 90°, 45°). The data suggests that right-eared participants tend to perceive sources as being more central than participants whose right ear is not dominant.

Discussion

The participants in this evaluation were successfully able to use the 3D audio system on the Nokia phone to identify targets at 45° intervals. As in Pulkki and Hirvonen's study, the deviation was considerably larger on the left than on the right side of azimuth 0°, but based on the results from this study we can assume that the AMMS 3D audio location controls will be appropriate for a 3D auditory interface. However, discriminative locations greater than five seem unlikely unless headtracking is used to allow more 'active' listening.

The earedness results were inconclusive but there is a suggestion that left or right ear dominance might affect the perception of the relative positioning of azimuths without affecting discriminative ability. The effect of centralizing the sources for right-eared subjects might be connected to the right hemisphere dominance in spatialization (the right ear is more strongly connected to the left hemisphere). However, our single left-handed participant performed very differently from the rest but without more data it is not possible to say if this was caused by left-handedness alone. The earedness effect has not been examined for 3D audio interfaces before so a separate study will be run to look at the effects of ear dominance on spatial audio localization where ear and hand dominance will be balanced.

Audio minimization study

Current cable TV interfaces deal with the issue of presenting concurrent visual streams by minimizing the TV image when the user interacts with the television menu to change channels or just browse what is available in the different channels. In the same way, in a rich auditory interface we need to be able to minimize streams when we are busy and need to focus on something else (e.g. talking to someone, crossing the road). We also need to minimize the current sounds (using spatial audio and distance attenuation) to be able to interact with the auditory menus controlling our user interface. We believe audio minimization, as with minimization in visual systems, could act as an important component in any audio interface.

Our current work is investigating how we minimize sound sources in a 3D auditory interface in a simple and coherent way so that users can deal with the interaction on which they are focused but return to the original source easily and without confusing its location. Our next study will act as a baseline to investigate the requirements for audio minimization. Namely, limits of cognitive load, user acceptability and the extent to which simple spatialization will support a simultaneous streaming strategy. To focus on these core questions, this experiment will consist of two single point sources, one streamed and one user-activated, and basic button presses. The streamed source will be minimized when the listeners interact with the user-activated sources in the foreground. Based on our previous evaluation results, the minimization effect will be created by moving the streamed audio source to the right hand-side of the frontal horizontal plane, which showed less variation in the location perception by listeners. The user-activated sources will not be spatialized and will be located at the origin (0°) so we can control for minimization effects alone. Results from this experiment will be used as a basis for further experiments investigating spatial user activated menus and head tracking using a SHAKE (Sensing Hardware Accessory for Kinesthetic Expression, see [11]) wireless sensor pack for real-time recording of tilt and heading data. To investigate the limits of cognitive load when using minimization, participants will be asked to carry out a number of tasks while listening to a podcast, i.e. streamed source.

Our baseline experiment hypothesis is that we will obtain better usability and effectiveness results when spatialized background and foreground interactions are used compared to non-spatialized implementations.

Conclusions

In this paper we have presented the results of an evaluation of the Java AMMS 3D audio location controls supported on the Nokia N95 8GB mobile phone. Results showed that the spatial audio system on this device provided clear location discrimination for 5 sources in the front 180°. This suggests that the audio capabilities of mobile phones are now capable of running 3D audio interfaces that were previously only possible on laptops, allowing the design and evaluation of more practical and effective mobile spatial audio interactions.

Acknowledgements

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

References

[1] Sawhney, N. and Schmandt C. Nomadic Radio: speech & audio interaction for contextual messaging in nomadic environments. ACM Transactions on Computer-Human Interaction. Vol. 7, (2000) 353-383.

[2] Buxton, B. Integrating the periphery and context: A new model of telematics. In: Graphics Interface 95 May 17-19, Quebec, Quebec, Canada, (1995) 239-246.

[3] Blauert, J. Spatial Hearing: The psychophysics of human sound localization. The MIT Press, (1999).

[4] Kaiser, J., Lutzenberger, W., Preissl, H., Ackermann, H., Birbaumer, N. Right-hemisphere dominance for the processing of sound-source lateralization. Journal of Neuroscience (2000) 20:6631-6639.

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

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

[7] Pulkki, V. and Hirvonen, T. Localization of virtual sources in multichannel audio reproduction. In: IEEE Transactions on Speech and Audio Processing. Vol. 13, No. 1, (2005) 105-119.

[8] Marentakis, G. and Brewster, S.A. A comparison of feedback cues for enhancing pointing efficiency in interaction with spatial audio displays. In: 7th int. conf. on Human computer interaction with mobile devices and services, ACM Press. Vol. 111, (2005) 55-62.

[9] D'Appolio, J. Testing loudspeakers. Audio Amateur Press. Peterborough (1998).

[10] www.jackielam.net/handedness/

[11] Williamson, J., Murray-Smith, R., and Hughes, S. Shoogle: excitatory multimodal interaction on mobile devices. In: Proc. SIGCHI conference on Human factors in computing system, ACM Press (2007) 121–124.