Experiments in Mobile Spatial Audio-Conferencing: Key-based and Gesture-based Interaction

Christina Dicke, Shaleen Deo,
Mark Billinghurst, Nathan Adams
HIT Lab NZ, University of Canterbury
Christchurch, NZ
+64-3-364 2349
{firstname.lastname}@hitlabnz.org

Juha Lehikoinen
Nokia Research Center
Visiokatu 1, 33720 Tampere, Finland
+358 (0)7180 08000
juha.lehikoinen@nokia.com

ABSTRACT

In this paper we describe an exploration into the usability of spatial sound and multimodal interaction techniques for a mobile phone conferencing application. We compared traditional keypad based-interaction to that of a newer approach using the phone itself as a device to navigate within a virtual spatial auditory environment. While the traditional keypad interaction proved to be more straightforward to use, there was no significant impact on task completion times or number of interaction movements made between the techniques. Overall, users felt that the spatial audio application supported group awareness while aiding peripheral task monitoring. They also felt it aided the feeling of social connectedness and offered enhanced support for communication.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Audio output; H.5.2 [**User Interfaces**]: Input devices and strategies, interaction styles.

General Terms

Design, Experimentation, Human Factors

Keywords

Spatial Audio, Mobile HCI, Gesture Interaction

1. INTRODUCTION

With improvements in digital cellular networks, and greater pervasiveness of wireless communication, mobiles phones are starting to support multi-party calling. Due to limitations in the phone hardware, remote participants are presented using mono audio. So although there are multiple sound streams from the various people on the phone call, they are presented with a single audio output, making it difficult to distinguish between speakers. However, previous research has shown that spatial sound cues can be used to distinguish between multiple sound sources, improve speech perception, and facilitate speaker identification [6, 8, 9].

MobileHCI, September 2-5, 2008, Amsterdam.

Based on this earlier work spatial audio should also provide benefits to multi-party mobile calls. However, before this can be realised, research is required on how to best use spatial audio to support multi-party conferencing, handling of multiple streams of information, and appropriate interaction methods for such non visual tasks.

In a previous study [2], we showed how phone motion-tracking can be used to interact with mobile spatial audio content. Motion-tracking methods could be used to translate movement in the real world to orientation movements for navigating a virtual spatial audio space. Using phone and head tracking, we conducted a user study evaluating these techniques in a spatial audio environment. We found that spatial audio modes using head and mobile phone tracking enabled better discrimination between speakers than fixed spatial and non-spatial audio modes. Spatialised audio with mobile phone orientation tracking provided the same level of speech intelligibility as head-tracking. Our study suggested that phone tracking is a viable option for orienting speakers in mobile virtual spatial audio environments.

Our long term aim is to explore how spatial audio can enhance multi-party conversations with mobile devices, how motion tracking can provide navigational support, and what interface techniques can be used to easily create a multiparty audio space.
In this follow-on study, our objective was to further explore the use of spatial sound and multimodal interaction techniques for enhancing social connectivity with mobile devices. We wished to further explore the efficacy of well-known selection, manipulation, and monitoring interaction paradigms for visual interfaces and point and click devices, and how well these might translate into a spatialised auditory domain for the mobile user.

2. RELATED WORK

Several researchers have explored the uses of spatial audiovisual cues for stationary and mobile applications. For example, Crispien and Savidis [7, 15] designed an egocentric spatial interface for navigating in, and selecting from, a hierarchical menu structure. The interface is designed for aligning both non-speech and speech audio cues in a ring circling around the user's head. These auditory objects can be reviewed and selected by using 3D-pointing hand gestures or speech recognition input.
Kobayashi and Schmandt built an egocentric dynamic soundscape [13] to create a browsing environment for audio recordings. In this application, a speaker orbits the user's head as it reads out the audio data, thereby mapping advancing within the audio source to movements on the circular path. Using a touchpad, the user can interact with the system to either create a new speaker (to rewind

or fast forward) or switch to an existing speaker. Up to four speakers simultaneously playing different portions of the same audio stream are possible. One speaker is always in focus and is played louder.

Frauenberger and Stockman [10] positioned the user in the middle of a virtual room with a large horizontal dial in front of them. The menu items are presented on the edge of the dial facing the user while the rest of the dial disappears behind the wall. The user can turn the dial in either direction by using a gamepad controller. Only the item in front of the user can be selected or activated. All items are synthesised speech.

Sawhney and Schmandt [16] created one of the first mobile spatial audio interfaces - the so-called "Nomadic Radio", a spatial audio application based on a wearable computer. Nomadic Radio notified the user about current events such as incoming e-mails or voicemail, messages, and calendar entries. Confirmation, aborting, and status were also represented by sounds. The audio messages were positioned in a circle around the listener's head according to their time of arrival. User interaction with the Nomadic Radio was by means of voice commands and tactile input. Although an interesting interface, there were no long-term formal evaluation studies conducted on usability aspects of the Nomadic Radio.

Walker and Brewster [17] developed single-user spatial audio applications for PDAs and mobile phones. Their work showed how spatial audio can be used effectively for information display on a PDA to overcome the limitations of a small screen. They used spatial audio to convey time remaining on a file download using an audio progress bar. In a user study, they found that users were more accurate in monitoring download performance with the spatial audio progress bar compared to a traditional visual progress bar. Brewster et al. [3] also created a mobile system based on Audio Windows by Cohen and Ludwig [4]. They used spatialised auditory icons localised in the horizontal plane either around or in front of the user's head. By using head or hand gestures the user can select an auditory icon from the menu to trigger the corresponding event, for example, checking for traffic reports or weather. Brewster and colleagues showed that their auditory interface improves the usability of the wearable device.

Other researchers have explored spatial audio for conferencing applications. Billinghurst et al. [1] describe a wearable conferencing space using spatial audio to disambiguate between speakers. Kan et al. [12] present a laptop-based system using GPS to give spatial audio cues based on the actual location of the speakers relative to the listener. More recently, Goose et al. [11] have developed the Conferencing3 3D application that runs on PDAs. In this case, they combined VoIP software with spatial sound rendering and 3D graphics in a PDA client-server application. Spheres depicted on the PDA represent remote collaborators. The spatial audio is generated by a server PC, based on the position of the speakers relative to the user location in the conferencing space. The final audio stream is sent wirelessly to the PDA for playback. In addition, they developed an archive client, which can be used to playback previously recorded conference sessions. This interface is the first example of a spatial audio conferencing application on a PDA. However, no evaluations of the usability of the system were undertaken.

Our work is unique because it is based around the mobile phone form factor with additional inertial tracking. Given the imperative to develop non-visual interaction methods of navigation through virtual communication spaces while mobile, we propose motion-tracking to be a viable candidate. Also, unlike these previous studies, we have conducted rigorous user evaluations of the spatial conferencing application.

3. EXPERIMENTAL DESIGN

In our previous study, we described a user evaluation comparing spatial and non-spatial audio conditions, with phone- and head-tracking methods for orienting in a spatialised multi-party audio environment. Given the model of a user surrounded by a sound space, the goal of this experiment is to determine which of two methods is best (as measured by time performance, error rates, task transitioning, and user preference measures) for allowing the users to navigate through the soundspace "around" them.

One of the novelties of this soundspace is the translation of the "foreground-background"-metaphor heavily used in visual interfaces like Microsoft Windows or Apple OS into the auditory realm. In these GUIs the user can minimize, maximize or tile windows depending on their focus of interest. Minimized windows may not deliver a constant stream of information however they can be set to notify about users' change of status, incoming messages, etc. Therefore, although the user may have her main focus on a word processor she still has a sense of awareness of – in this case – her social network as represented by the messaging application. To support focusing of attention and monitoring of "background" events, we created an auditory interface consisting of two concentric user-centric horizontal rings. By using the metaphor of distance, we enabled the user to push items they are not currently focusing on "away" to the outer ring, but without depriving them of the option of monitoring and easily switching between different streams of information.

We were particularly interested in user interaction with sound sources in this 3D environment, and in particular with using the phone itself as an input device through gesture tracking for navigation. More specifically, our questions of interest were:

1. With no visual feedback, to what extent are users able to navigate and transition between several different audio streams of differing complexity?
2. In what ways do users utilize the perception of distance in the spatial sound environment to support task focus?
3. Which of the interaction methods did users prefer and why? How did these methods affect task performance with respect to time and degree of input required?

We used a conceptual model of a user surrounded by a sound space, and developed a prototype interface to evaluate the virtual spatial audio environment containing various types of audio items, including a simulated group conversation such as in a multiparty call. To do this we used a standard computer, headphones, and a phone mock-up device (shown in figure 1). This enables rapid application development and testing on the PC. The phone device had an Intersense[1] 3-DOF orientation sensor for motion direction tracking, in order to explore gesture-based interaction, and sent tracking data via USB to the PC-based application.

[1] www.isense.com

Figure 1. The mobile phone mockup device equipped with an Intersense Inertia Cube3 for motion tracking

The interaction metaphor for the experiment is push/pull and pan:

- Panning causes both rings (inner and outer) to rotate clockwise or counterclockwise. Items positioned on the rings rotate accordingly. A swooshing sound was played as feedback to a successful panning movement and the item in focus was announced. From the initial position rotating the rings counterclockwise would announce "Music" to be in focus, then "Podcast", then "Group Conversation" then again "Music" and so forth.
- Pushing the item in focus to the next farther ring (lower volume/farther distance)
- Pulling the item in focus to the inner ring (louder volume/closer distance to user's head)
- Activating the item in focus by pulling beyond the inner ring

The sound source directly in front of the user (that is, 0° azimuth) has focus and is played at a slightly louder volume than the other items on the same ring. A sound source can only be moved out/in to a ring when it is in focus.

The two interaction methods used in the study were:

Buttons: the left and right button input was use to pan the rings counterclockwise or clockwise; down to pull an item closer or activating it, and up to push an item away or deactivating it. *Gestures*: rotation of the phone left or right to pan the rings (shown in figure 2a); vertical gesture upwards/towards the user for pulling the item in focus closer, and a vertical movement downwards/farther away from the user to push the item in focus away (shown in figure 2b).

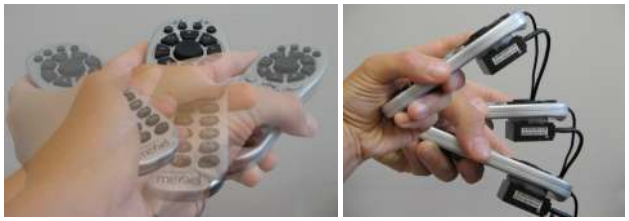

Figure 2a,b. Gestural interaction techniques: panning (left) and pitching (right)

3.1 Audio Content

The audio content consisted of different types of audio to simulate both sporadic and continuous audio streams, group conversation, and system notifications.

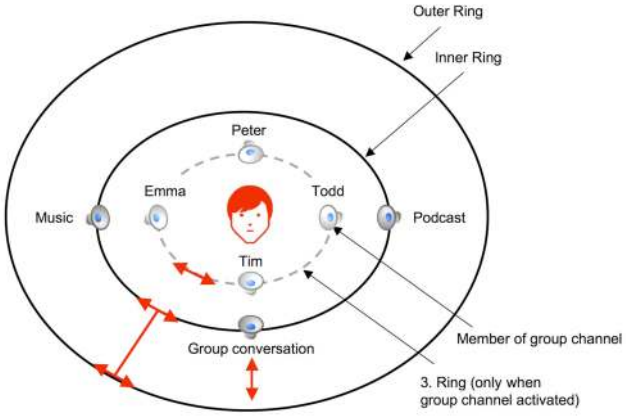

Figure 3. Layout of the soundspace, showing sound cues and interaction methods

The audio content used (as depicted in figure 3) was:

- Group channel (semi-continuous stream): four speakers having a pre-recorded conversation (sometimes overlapping in speech). Each speaker has a separate audio stream. The group-level (combined) audio was initially placed directly in front of the user, on the inner ring.
- Music (continuous stream): to the speaker's right at 90 degrees.
- Podcast (continuous stream): individual audio stream representing speech-based audio content of interest to the user. Initially placed to the speaker's left at 90 degrees.
- Notification beeps (sporadic audio items): representing generic events, such as incoming phone call, or calendar event. These occurred at random intervals in the final task only. They were played in stereo and not spatialised.

3.2 Group Interaction

Our goal was to gauge the effectiveness of our transitioning mechanism between individual speakers and a group of speakers, and then to be able to identify, isolate, and direct communication to an individual from a group. To this end, we included one individual speaker audio channel in the set-up as a Podcast, as well as one four-speaker "synchronized" group-level channel, in which a pre-recorded group conversation was taking place.
If the group-level channel was located on the inner ring, pulling closer created a temporary ring composed of the group's individual speakers, spatially separated around the user. While the group channel was activated, the Podcast and music were muted and the user could only hear the group conversation. Activating the Podcast or the music by pulling the item from the inner ring towards the user muted all other sounds. While activated they had no spatial effect but were played in stereo. All sounds could be

deactivated by pushing them away (onto the inner ring), which restored the previous layout.

3.3 "Whispering" to an Individual Speaker

The group members are, by default, heard at the same volume level by the user. In order to isolate a particular member, the user has to rotate the ring until that speaker is at the 0° azimuth position. In order to "whisper" (that is, communicate on a dedicated channel) to that person, the user pulls him or her closer. While whispering to a person all other group members are muted. For the purposes of our study, in which we wished to firstly examine the efficacy of spatial sound for disambiguation in listening to and isolating various group members as per communication needs, this functionality is currently intended as proof of concept. The emphasis was on identifying the high-level interaction concerns.

3.4 3D Spatial Sound Representation

The hardware set-up is as follows (see figure 4). A Mavael Keiboard equipped with an Intersense Inertia Cube3 (IC3) for 3-DOF gesture tracking and button interaction has been used. For this experiment, the sound sources were pre-setup. To guarantee an authentic distance effect each individual sound stream was pre-recorded very close to the speaker and at the same time from a distance of about three meters. We chose to use binaural processing on stereo headphones, with the application of HRTFs from the fmod sound library[2]. A pre-study showed that the distance effect in the recorded files was not perceived to be very distinct. Paying particular attention to preservation of intelligibility we post-processed attenuation and reverberation to achieve a clear, recognizable distance effect.

Figure 4. Experimental setup

All four speakers of the group conversation were recorded separately. These four tracks were initially played from the same position. If the group conversation was activated these four tracks were spatially fanned out as can be seen in figure 3. Special care has been put into the recording procedure to ensure authentic reproduction of each speaker's individual characteristics. In addition, Yankelovich et al. [19] have shown that audio quality has a strong influence on the effort required to understand the meaning of sentences and on the perceived sense of presence in a teleconferencing environment. Thus, we tried to minimize confounding variables caused by low audio quality. During the testing users wore Sony MDR-V700 adjustable headphones.

3.5 Experimental Procedure

In this experiment we wanted to simulate the experience of a multi-party group conversation in the presence of concurrent audio information and tasks. To this end, we used pre-recorded audio streams to have complete control over the audio content.

We conducted a within-subjects study. The first two factors were the input methods (namely, input via buttons on the keypad or using motion tracking of the phone for gesture input), and the dependent variables were task completion time, number of clicks respectively gestures made, and number of missed notifications (only for task four) across the four tasks in the study. We also evaluated the results of an extensive post-study questionnaire on user interaction satisfaction, perception of sounds and sound localization, and personal preferences for an interaction technique.

34 people participated in the study. The mean age was 33.9 years spanning from 16 to 57 years and approximately half of the participants were below thirty years of age. 47 percent of the participants were female, and 53 percent were male. The majority regarded English as their first language (70.6 percent). The majority of the participants (73.5 percent) reported using a computer more than 30 hours per week. 32.4 percent of the participants regarded themselves as quite experienced mobile phone users, with another 23.5 percent regarding themselves as expert. Just over half of the participants felt themselves to be novices with auditory interfaces (58.8 percent). Two of the participants reported to have minor hearing difficulties but were not excluded from the study because of the negligibility of their impairments.

Participants were presented with each of the four tasks in the same order, with either the buttons or gesture-based interaction methods first in alternating order. Before starting the tasks participants were asked to familiarize themselves with the technology until they felt to have a good understanding of the interface and the interaction methods.

Each subject was then presented with the following four tasks, in the given order:

T1. Please move the music to the position 0° azimuth (right in front of you) and push it to the outer ring.

T2. You want to monitor all sounds but the group channel is distracting you. What do you do?

T3. You would like to concentrate on the group conversation. What do you do?

T4. Please identify the member of the group channel who is still working. Once you've done so please open a whisper/private channel to this member by pulling him or her closer. While you are doing so, please hit SPACEBAR whenever you hear a notification.

T4 was designed to be the longest and most complex task.

Data was logged on task times, and lateral/vertical movements (both via buttons or gestures). For T4, the number of notifications given, and the number of these responded to by pressing the

[2] www.fmod.org

spacebar was also logged. The virtual configuration of sound sources was only reset prior to task four.

After data was collected for each task the subjects also filled out a subjective survey giving responses on a number of questions such as how easily they were able to navigate between audio streams, how intuitive each interface was, which interaction method they preferred, as well as whether they would consider using the application for group-based communication in daily life.

4. RESULTS

The following section describes the detailed results of the tasks, including time and number of movement differences between the two interaction methods (that is, the use of buttons or phone-based gestures) to re-position sound sources, system notifications missed, and participant responses to each interface on the whole.

4.1 Performance with Buttons vs. Gestures

We found that, on average, users used the following number of movements for solving the tasks, where a movement is either a left/right/up/down click of a button on the keypad or a lateral/pitch movement of the phone in the case of gestures.

Task	Mean number of button clicks	Mean number of phone gestures
T1	4.2	4.0
T2	6.9	6.2
T3	5.0	7.3
T4	10.1	10.0

Table 1. Mean number of movements made per task for both interaction techniques

The minimum number of movements required per task was 2 for each of T1, T2, and T3 and a minimum of three movements for T4. Table 1 gives an overview of mean numbers of movements per task and table 2 gives the equivalent for task completion times.

Task	Mean task completion time for buttons (sec.)	Mean task completion time for gestures (sec.)
T1	7.1	12.5
T2	20.7	16.4
T3	12.1	22.9
T4	76.2	82.5

Table 2. Mean task completion times per task for both interaction techniques

For T1 we found no significant difference ($t(30)=0.16$, $p=.88$) in the number of movements for buttons ($M=4.2$, $SD=3.5$) and gestures ($M=4.0$, $SD=2.5$). We did find a trend ($t(30)=1.89$, $p=.069$) in the time taken for this task between the two conditions of buttons ($M=7.1$, $SD=8.3$) and gestures ($M=12.5$, $SD=12.3$) towards a faster task completion time when using buttons.

For T2 and T3, we observed that people usually chose between two different approaches. The approach chosen influences the number of movements made. To recap, T2 involved monitoring all sounds whilst finding that the group channel was distracting, and therefore finding a way to deal with that. The approaches in response to this task were:

1. Pull music back on inner ring, push group conversation on outer ring
2. Leave music on outer ring, push group conversation on outer ring

T2	N	Buttons Movmts	Buttons Time (sec.)	N	Gestures Movmts	Gestures Time sec.)
A1	17	$M=5.3$ $SD=9.0$	$M=26.1$ $SD=55.1$	19	$M=6.5$ $SD=4.8$	$M=17.1$ $SD=19.0$
A2	13	$M=7.9$ $SD=2.5$	$M=8.9$ $SD=7.7$	11	$M=5.2$ $SD=4.8$	$M=11.1$ $SD=11.1$

Table 3. T2, for both approaches: Participants chosen an approach, mean number of movements, and task completion time for both interaction techniques

We found no significant difference in the task completion time ($t(28)=1.115$, $p=.274$) nor in the mean number of movements made ($t(28)=1.266$, $p=.216$) using either interaction technique, between participants who opted for the first approach and those who chose approach 2 (cf. table 3 for details on means and standard deviations).

Breaking down approach 1 further, we found no significant difference in times ($t(10)=-0.868$, $p=.406$) or number of movements made ($t(10)=-1.005$, $p=.339$) using either interaction technique. Due to insufficient numbers of participants opting for the second approach, the same comparison between interaction techniques using this approach could not be made.

T3 involved concentrating on the group channel. The approaches in response to T3 were:

1. Pushing music and Podcast on the outer ring, bringing group conversation on the inner ring (this allowed monitoring the two other sound sources while focusing on the group conversation)
2. Pulling group conversation to the inner ring, then activate group conversation (which muted all other sound sources).

T3	N	Buttons Movmts	Buttons Time (sec.)	N	Gestures Movmts	Gestures Time (sec.)
A1	13	$M=6.8$ $SD=4.1$	$M=14.4$ $SD=12.7$	11	$M=9.5$ $SD=4.1$	$M=35.3$ $SD=36.4$
A2	19	$M=4.0$ $SD=2.4$	$M=10.4$ $SD=15.1$	21	$M=5.8$ $SD=4.0$	$M=16.0$ $SD=19.6$

Table 4. T3, for both approaches: Participants chosen an approach, mean number of movements, and task completion time for both interaction techniques

For T3, we found no significant difference in the task completion time ($t(30)=-0.783$, $p=.44$) using either interaction technique, between participants who used the first approach and those who used the second. However, in comparing the number of movements made on average between the approaches for this task, we found that those who used the first approach made more clicks (in the use of the keypad) ($t(28)=-2.334$, $p=.027$), as well as more movements (in the use of the phone gestures) ($t(29)=-2.396$,

p=.023, cf. table 4 for details on mean values and standard deviations).

On this basis, approach 1 appears to have necessitated more by way of input from the users without affecting their task completion times significantly. This apparent contradiction is reinforced with the only moderate correlation ($r = 0.536$) between times and number of movements made in this task as can be seen in table 5.

We had insufficient data again to further analyze between the interaction techniques for those who preferred approach 1 for this task. For approach 2 we found no significant difference in task completion time ($t(15)=-1.248$, $p=.231$) between the two techniques but a trend towards less movements being made when using buttons ($t(14)=-2.082$, $p=.056$).

For T4, we found no significant difference between the techniques, either in terms of the number of movements ($t(29)=0.52$, $p=.601$), or in task completion times ($t(28)=0.67$, $p=.499$).

The following table presents the strength of the correlation (using Pearson Correlation coefficients) between movements made and time for task completion, across all approaches and tasks.

Task	Buttons: Correlations between movements and task completion time	Gestures: Correlations between movements and task completion time
1	N=32, $r = 0.827$, $p<.01$	N=33, $r = 0.683$, $p<.01$
2, A1	N=17, $r = 0.903$, $p=.01$	N=17, $r = 0.909$, $p<.01$
2, A2	N=3, insufficient number of cases	
3, A1	N=14, $r = 0.536$, $p<.05$	N=14, $r =0.752$, $p<.01$
3, A2	N=8, insufficient number of cases	
4	N=31, $r = 0.461$, $p=.01$	N=32, $r =0.570$, $p<.01$

Table 5. Correlations per task between the interaction techniques

As shown in table 5 there were - as expected - moderate to strong correlations found between the number of movements made and task completion time across the conditions, where there was sufficient data.

4.1.1 Sequence effects

When subjects commenced the study using the gestures interaction method, they rated using buttons ($M=4.5$, $SD=0.6$) significantly higher ($t(32)=2.553$, $p<.05$) for straightforwardness of use, than participants who started by using buttons ($M=3.9$, $SD=0.8$).

An impact was also observed of sequence on task completion time and the number of movements made. Participants were faster and used fewer movements in both T2 and T4 when their first condition was gestures. For T2, using buttons was significantly faster ($M=7.4$, $SD=5.6$, $t(30)=-2.036$, $p=.05$) and significantly fewer clicks were made ($M=4.5$, $SD=2$, $t(29)=-2.057$, $p<.05$) than participants starting with buttons (task completion time: $M=36$, $SD=57.8$; movements: $M=9.5$, $SD=9.8$).

For T4, the same learning effect could be observed. Again, for the condition "buttons" participants were significantly faster ($M=63.6$, $SD=12.8$, $t(28)=-2.81$, $p<.01$) and used significantly fewer movements ($M=8.3$, $SD=5.1$, $t(30)=-2.214$, $p<.05$) when they started with gestures. Starting with buttons had a negative effect on the mean scores for the "button" condition as participants were slower ($M=91.4$, $SD=39.3$) and took more clicks ($M=13.4$, $SD=7.9$).

Participants performing in the sequence of first using gestures and then using buttons to complete the tasks showed a strong learning effect for solving tasks by pressing buttons. This could be due to knowledge and experience gained while using gestures. The same effect could not be observed for using gestures. We assume that the observed difficulties with the gesture recognition forced some participants to concentrate more on the interaction technique than on performance and hence overlaid the learning effects we could observe for button interaction.

4.1.2 Missed Notifications

For T4, participants were asked to perform a secondary task of listening for system "notification" beeps, conceptually representing the arrival of email or incoming calls, etc, whilst conducting the primary task of identifying a particular group member and whispering to them. On average, users missed 0.6 notifications, or approximately fifteen percent of the notifications that each participant heard. No significant difference ($t(30)=-0.226$, $p=.823$) could be found for missed notifications between the two interaction conditions.

4.2 User Satisfaction

Tables 6 and 7 present a summary of participant responses to the post-study questionnaire, comparing each interaction technique against several criteria including navigability between sounds sources, whether the system functioned in a straightforward manner (thus aiding ease of learning), and the overall user preference. We also evaluated the overall satisfaction with the interface. Questions included:

Are you satisfied with the accuracy of the system?

Not at all < 1 - 2 - 3 - 4 - 5 > very much so

Is the system easy to use?

Not at all < 1 - 2 - 3 - 4 - 5 > very much so

Using the application is/feels:

difficult < 1 - 2 - 3 - 4 - 5 > easy

frustrating < 1 - 2 - 3 - 4 - 5 > satisfying

dull < 1 - 2 - 3 - 4 - 5 > fun

Ratings of 1 or 2 were grouped as negative responses, 3 regarded as undecided, and 4 and 5 as positive responses.

Interface aspect	Buttons, % pos. response	Gestures, % pos. response
Learning to navigate the system is easy.	73.5	65.2
All given tasks can be performed straightforward.	85.3	52.9
My location within the system at any given time is apparent.	67.7	64.7

I am satisfied with the accuracy of the interaction technique.	100	61.8
I liked using buttons better than gestures.	61.7	
I liked using gestures better than buttons.	32.7	

Table 6. Participants' ratings of interaction methods

The results for these preference ratings can be summarized as follows: In learning to navigate, buttons (M=4.3[3], SD=1.0) were deemed to be significantly easier ($t(33)$=3.419, p<.01) than using phone gestures (M=3.5, SD=1.2).

On straightforwardness of use, buttons were rated to be significantly more straightforward ($t(33)$=3.316, p<.01). The average rating for buttons was M=4.2 (SD=0.8), and for gestures M=3.6 (SD=1.1).

In accuracy of interaction technique, buttons were deemed significantly more accurate ($t(33)$=4.509, p<.01), with average scores of M=4.6 (SD=0.5) for buttons and M=3.5 (SD=1.4) for gestures. We interpret these results as a consequence of the prototypical implementation of the gestural interaction and not the gestural interaction per se. There was a significant overall preference for buttons as the interaction method for this application ($t(33)$=2.701, p=.01) with buttons: M=3.8 (SD=1.3), gestures: M=2.7 (SD=1.3) as mean scores.

There was no significant difference found in sense of location perceived within the system, as afforded by either technique. The following table presents the percentage of positive and negative reactions by the participants to various aspects of the system as a whole.

Aspect	% Positive Response	% Undec.	% Negative response
Interface			
Satisfaction with …			
System accuracy	79.2	-	29.6
Auditory nature	67.7	20.6	11.8
Audio layout	91.1	5.9	2.9
Easiness of Usage	76.4	17.6	5.9
Efficiency of System	70.6	17.6	8.8
User Interaction Satisfaction	% Left/Neg. (word of pairing)	% Undec.	% Right/ Pos.
Terrible - Wonderful	2.9	32.4	64.7
Difficult - Easy	23.5	5.9	70.6
Frustrating - Satisfying	5.9	32.4	61.7
Dull - Fun	2.9	5.9	91.2
Slow - Fast	8.8	38.2	52.9
Boring - Stimulating	2.9	14.7	82.4
Impersonal - Personal	8.8	20.6	70.6
Passive - Active	8.8	8.8	82.3
Awareness, distraction, etc.	% Pos. response	% Undec.	% Neg. response
Support for group awareness	58.8	26.5	14.7
Aids concentration on other tasks	29.4	26.5	44.1
Causes distraction from other tasks	55.8	20.6	23.5
Aids monitoring other tasks	67.7	20.6	32.4
Aids connectedness to social network	70.6	17.6	11.8
Sound	% Pos. response	% Undec.	% Neg. response
Sound identification	70.6	20.6	8.8
Overall quality of sound	91.2	8.8	0
Sound position identification	64.9	20.6	13.5
Helpfulness of spatial sound	85.3	11.8	2.9
Distance effect of sound	67.7 (rather good)	20.6	11.7 (rather poor)

Table 7. Participants' interaction satisfaction responses

In addition, in response to whether they would use this type of application for group awareness activities in daily life, 8.8 percent said never, 50.0 percent said occasionally, 32.4 responded often, and 8.8 percent responded always.

4.3 Gender Effect and Lab Affiliation

Our results show a strong influence of what could either be gender or affiliation with our lab on the overall rating of the interface. This indistinguishability is due to the fact that most male participants (except for three) were working at the lab and most female (except for four) were not. The variables representing gender and affiliation with the lab show a significant correlation (N=34, r=0.589, p<.01).

The results show no significant difference between genders/affiliation with the lab measured on task completion time, number of movements made, and missed notations. But for task T3, (using gestures) a significant relation $\chi 2$ (2, N=32) = 9.219, p<.01) between men/members and approach 2 was found. That is, men/members (N=15) expanded the group call and therewith muted all other sources more often than women/non-members (N=6). Nine women/non-members chose approach 1 which means they brought the group conversation to the inner ring and moved all other sound sources to the outer ring, in comparison, only two

[3] User ratings on a Likert Scale of 1= negative maximum and 5 = positive maximum rating.

men/members did the same. No significant effect could be found for T3 when using buttons.

Interface aspect	p	t	M / W[4]	SD	M / M[5]	SD
Accuracy of System	< .01	(32) 3.371	4.7[6]	0.6	3.9	0.8
Easiness of usage	.01	(32) 2.755	4.5	0.7	3.7	0.9
Efficiency of System	< .01	(31) 3.088	4.6	0.6	3.7	1.1
Respond in real time	< .01	(31) 3.002	4.6	0.6	4	0.6
Frustrating-Satisfying	< .01	(32) 3.039	4.2[7]	0.9	3.4	0.6
Dull-Fun	<. 01	(32) 3.088	4.7	0.5	4	0.8
Boring-Stimulating	< .05	(32) 2.453	4.4	0.5	3.8	0.9
Insensitive-Sensitive	< .05	(32) 2.593	4.3	0.8	3.4	0.8
Cold-Warm	< .01	(32) 2.843	4.1	0.7	3.4	1.1
Passive-Active	< .01	(32) 2.618	4.7	0.5	3.9	1.1
Aids monitoring other tasks	< .05	(32) 1.766	4.1	1	3.6	0.9
Sound pos. identification	< .05	(32) 2.102	3.7	1.1	3	1.3

Table 8. Satisfaction responses influenced by gender or affiliation with the HIT Lab NZ

Table 8 summarizes results from an independent sample t-test on general satisfaction with the interface. It shows that ratings from women/participants not from the lab were significantly higher for some items of the questionnaire. Observations during the study would rather support the interpretation that the actual influential factor is affiliation with the HIT Lab NZ. Participants not from the lab were more excited about the experiment itself and about using the interface. As they are not as experienced in dealing with new technologies and multimodal prototypes as participants from the lab were, the uniqueness of the whole experience might have influenced the ratings. However, the interface introduced in this paper is purely auditory and research suggests women have better hearing at frequencies above 2000 Hz (frequency range of speech is approx. between 150-5000 Hz) [5] [14]. Additionally, women are more likely to engage in the elaborative processing of the meaning of verbal (or verbally encoded) information [18]. These aspects suggest it may have been easier or more enjoyable for women to operate the system. However, further research for these variables is definitely indicated on the basis of our current results.

[4] Mean values for women/participants not from the lab.

[5] Mean values for men/participants from the lab.

[6] On a 5-point Likert Scale with 1 being the negative maximum and 5 being the positive maximum.

[7] On a 5-point Likert Scale with 1 representing the left word of pairing (e.g. frustrating) and 5 the right word of pairing (e.g. satisfying).

4.4 Discussion

The significantly longer performance times observed in T3 with using gestures rather than buttons was likely due to the observation that participants paused to listen more to the audio content during this task using gestures, perhaps to affirm their orientation.

Participants commented that they would have preferred to have more feedback for the pitch movements, corresponding to the level provided with the lateral movements (for which a "swooshing" sound was heard), and that it consequently took them more effort to develop their conceptual model of the system, with relation to navigation.

We also observed that it took participants longer to form an adequate mental model if they began the study with using gestures. It appeared that this interaction technique produced more errors and confusion initially, as participants continued to assimilate the proper gestures, despite the training. During this training period, the techniques were each explained and demonstrated. All participant questions were answered, and it was emphasized that the training could take as long as the participant desired to feel comfortable with the technique. However, the novelty of this interaction technique in this domain is subsequently reflected in the results where users who commenced with using gestures rated using buttons as more straightforward than users who started with buttons. Buttons similarly fared better in their estimation of the usefulness of the system to social networking, using that interaction method, and for T2 and T4, task completion times and number of movements made were also affected by this sequence effect with buttons taking less time and movements when preceded by gestures.

On the other hand, if the participants started with buttons we did not observe an effect of sequence on their use with gestures. This may be due to the quicker forming of a correct mental model without the confusion deriving from the unfamiliarity of gesturing with the phone as an interaction technique. Gaining a good understanding of the interface while using buttons may have compensated irritations produced by using gestures to interact with the system.

However, it was observed that once participants understood how to operate the interface they were getting much faster and more precise (that is, on the basis of making less unnecessary movements). Thus it appears that there was a learning effect when buttons were used first.

On the basis of the secondary listening task in T4, whereby participants monitored system notifications and responded to them, the interaction method used was not found to affect participants' response rates. Thus, it could be that using gestures for secondary audio-based tasks was not additionally distracting to the primary task, relative to using the traditional buttons on the keypad.

5. CONCLUSIONS

The analysis of our study into the usability of spatial sound, the metaphor of "distance", and multimodal interaction techniques with mobile devices provided interesting insight with regards to our research questions. These findings are summarized below.

1. Identifying, navigating, and transitioning between several different audio streams

- Our study used no visual feedback, and successfully used auditory feedback as well as the participants' own kinesthetic awareness through gesture movement, to aid in imparting a sense of the virtual 3D sound environment.
- User satisfaction ratings with the overall system were very high, particularly with respect to the elements of fun, stimulation and active participation. Users stated that they had no problems identifying single sound items. Item selection and manipulation could be easily accomplished.
- Approaches chosen for T2 and T3 indicate that applying a distance effect is a viable option to support background awareness and focus direction.

2. Mental model, awareness of system state, awareness of location within the system

- There were negligible rates of failure to complete the tasks. Once participants felt themselves to be acquainted with the application they had a correct understanding of the system and a good sense of their location within the system. Objects could be correctly identified and successfully manipulated.
- Almost all participants were able to maintain a sense of the spatial environment with both approaches. This suggests further avenues of research for the positive use of such a spatial auditory environment. Enhancements to the spatial effect and to the proximity feedback of various sound sources in the user's audio sphere would also be areas for future improvements to the system.

3. Interaction methods

- While most people preferred using directional keys on the device for this application, approximately one third of participants preferred using gestures. Users' comments suggest that they liked the playfulness and intuitiveness of the gesture-based interaction method. These findings suggest a positive pathway for the acquisition of more familiarity with the gestural interaction technique.
- A strong effect of either gender or familiarity and experience with new technologies was seen on general ratings of the interface. Further evaluation is needed to clarify which of these factors lead to a rather positive or negative perception and hence rating of the interface.
- By enabling participants to pursue their own approach to T2 and T3, we discovered that some participants utilized the notion of distance (with foreground/background) separation of focus with the spatial sound sources. Others preferred to focus more particularly on the single sound item in question.

Overall, the results of this study were positive for continuing to further explore multimodal interaction for purely auditory interfaces to compliment or substitute visual interfaces for mobile application scenarios. Responses were strong as regards the applicability to group communication and multitasking. We assume that the proposed interface would also be applicable for functional navigation or navigation in hierarchies. This does not only include handheld devices like mobile phones, music players, or digital assistants but also assistive technology for the visually impaired.

Our results also point towards an optimistic estimation for utilizing spatial sound to promote a feeling of greater connectedness with social networks. Currently most social network technology (websites, messengers, virtual worlds, etc.) is stationary and heavily based on visual cues. Using a constant but only sporadically used audio connection, as we have simulated in our experiment, may be a viable alternative for supporting group communication, awareness, and a feeling of social presence of both real and virtual social networks, not only in mobile scenarios. However these seem to be very fruitful areas for future research.

REFERENCES

[1] Billinghurst, M., Bowskill, J., and Morphett, J., Wearable communication space, in *British Telecommunications Engineering, vol. 16, 1998*, pp. 311-317.

[2] Billinghurst, M., Deo, S., Adams, N., & Lehikoinen, J., Motion-Tracking in Spatial Mobile Audio-Conferencing, presented at *Workshop on Spatial Audio for Mobile Devices (SAMD'07), Mobile HCI'07 (Singapore)*, 2007.

[3] Brewster, S., Lumsden, J., Bell, M., Hall, M. and Tasker, S., Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices, in *SIGCHI conference on Human Factors in Computing Systems, vol. 5 (1), 2003*, pp. 473-480.

[4] Cohen, M. and L. F. Ludwig, Multidimensional audio window management, in *International Journal of Man-Machine Studies, vol. 34(3), 1991*, pp. 319-336.

[5] Corso, J.F., Age and Sex Differences in Pure Tone Thresholds, in *Archives of Otolaryngology, vol 77, 1963*, pp. 385-405.

[6] Crispien, K. and Ehrenberg, T., Evaluation of the "cocktail party effect" for multiple speech stimuli within a spatial audio display, in *Journal of the Audio Engineering Society 43, 1995*, pp. 932–940.

[7] Crispien. K., Fellbaum. K., Savidis. A. and Stephanidis. C., A 3D-Auditory Environment for Hierarchical Navigation in Non-visual Interaction, in *Proceedings of the 3rd International Conference on Audio Display*, Palo Alto, USA, 1996, pp. 18-21.

[8] Drullman, R., & Bronkhorst, A.W., Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation, in *Journal of the Accoustical Society of America, vol. 107(4), 2000*, pp. 2224-2235.

[9] Ericson, M. A., and McKinley, R. L., The intelligibility of multiple talkers separated spatially in noise, in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Erlbaum, Mahwah, NJ), 1997, pp. 701–724.

[10] Frauenberger. C. and Stockman. T., Patterns in Auditory Menu Design, in *Proceedings of the 12th International Conference on Auditory Display,* London, UK, 2006, pp. 141-147.

[11] Goose, S., Riedlinger, J. and Kodlahalli, S., Conferencing3: 3D audio conferencing and archiving services for handheld wireless devices, in *International Journal of Wireless and Mobile Computing, vol. 1(1)*, 2005, pp. 5-13.

[12] Kan, A., Pope, G., Jin, C., and van Schaik, A., Mobile Spatial Audio Communication System, in *Proc of ICAD 04-Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, 2004*.

[13] Kobayashi, M. and Schmandt, C., Dynamic Soundscape: mapping time to space for audio browsing., in *Human Factors in Computing Systems, CHI*, 1997, pp. 194-201.

[14] Pearson J.D., Morrell C.H., Gordon-Salant S., Brant, L.J., Metter, E.J., Klein, L.L. and Fozard, J.L., Gender differences in a longitudinal study of age-associated hearing loss, in *Journal of the Acoustic Society of America ,1997*, pp.1196-1205.

[15] Savidis A., Stephanidis. C., Korte. A., Crispien. K. and Fellbaum. K., A Generic Direct-Manipulation 3D-Auditory Environment for Hierarchical Navigation in *Non-visual Interaction, in Proceedings of Assets'96. New York, USA, ACM*, 1996, pp. 117-123.

[16] Sawhney, N., and Schmandt, C., Nomadic Radio: Speech and audio interaction for contextual messaging in nomadic environments, in *ACM Transactions on Computer-Human Interaction, vol. 7, 2000*, pp. 353-383.

[17] Walker, A. and Brewster, S. A, Extending the Auditory Display Space in Handheld Computing Devices, in *Proceedings of the Second Workshop on Human Computer Interaction with Mobile Devices, Edinburgh, 1999*, Available at www.dcs.gla.ac.uk/~mark/research/workshops/mobile99/

[18] Wirth, M., H. Horn, Koenig, T., Stein, M., Federspiel, A., Meier, B., Michel, CM. and Strik, W., Sex Differences in Semantic Processing: Event-Related Brain Potentials Distinguish between Lower and Higher Order Semantic Analysis during Word Reading, in *Cerebral Cortex, vol. 17(9), 2006*, pp. 1987-1997.

[19] Yankelovich, N., J. Kaplan, Provino, J., Wessler, M. and DiMicco, J. M.. Improving Audio Conferencing: Are Two Ears Better than One? *CSCW, ACM*, *2006,* p. 334-342.