Experiments in Spatial Mobile Audio-Conferencing

Shaleen Deo
HIT Lab NZ
University of Canterbury
Christchurch, New Zealand
+64-3-364-2349
shaleen.deo@hitlabnz.org

Mark Billinghurst
HIT Lab NZ
University of Canterbury
Christchurch, New Zealand
+64-3-364-2349
mark.billinghurst@hitlabnz.org

Nathan Adams
HIT Lab NZ
University of Canterbury
Christchurch, New Zealand
+64-3-364-2349
nathan.adams@hitlabnz.org

ABSTRACT

With improvements in digital cellular networks, and greater pervasiveness of wireless communication, mobile phones begin to support multi-party calling. Due to hardware limitations in current models, speakers are presented in a single audio stream, causing confusion and decipherability issues. The case for using spatialised audio is strong from other domains, and here we present an interaction approach to realising spatialised audio for multi-party calls in mobile contexts using motion-tracking. Our study suggests that phone tracking is a viable option for orienting to speakers in virtual spatial audio environments on mobile devices.

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

1. INTRODUCTION

While audio-conferencing, it can be difficult to listen to several speakers at the same time. In face-to-face conversations, listeners are able to take advantage of binaural difference cues that occur when sounds are located at different positions relative to the listener’s head. This effect can be reproduced electronically through the use of bi- or multi-channel audio streams modified with the application of digital filters (head related transfer functions, or HRTFs). Thus, spatial sound cues can be used to distinguish between multiple sound sources, improve speech perception, and facilitate the identification of speakers [4, 5].

With improvements in digital cellular networks, and greater pervasiveness of wireless communication, mobile phones are starting to support multi-party calling. However, due to limitations in the phone hardware, speakers are presented using mono audio. So although there are multiple sound streams from the various speakers, they are presented with a single audio output, making it difficult to distinguish between speakers.

We propose that spatial audio capability will provide commensurate benefits to multi-party mobile calls also, in future devices. However, before this can be realised, research is required on how to best use spatial audio to support multi-party conferencing, and appropriate interaction methods for such non-visual tasks. Reliance solely on visual interactions methods in mobile contexts of use proves challenging, if not at times impossible. Exploration of alternative modal spaces and complimentary interaction methods is therefore necessary.

We suggest motion tracking methods to translate movement in the real world to orientation movements for navigating a virtual spatial audio space. Using phone and head tracking, we describe a user study evaluating these techniques in a spatial audio environment. Our long term aim is to explore how spatial audio can enhance multi-party conversations with mobile devices, how motion tracking can provide suitable navigational support, and what interface techniques can be used to easily create a multi-party audio space.

2. RELATED WORK

Researchers such as Walker and Brewster have explored the use of spatial audio on mobile devices, such as for addressing problems of visual clutter in mobile device interfaces [9]. Similarly Swahney [8] describes using spatial audio to manage phone messages and provide contextual cues in a mobile environment. Most work, however, has focused on single user interfaces, rather than enhancement of mobile conferencing.

There are several prior examples of using spatial cues in a conferencing setting. In collaborative virtual environments Benford [2] has shown that spatial cues can combine with audiovisual cues in a natural way to aid communication. The well-known “cocktail-party” effect shows that people can easily monitor several spatialised audio streams at once, selectively focusing on those of interest. Even a simple virtual avatar representation and spatial audio model of other users in the collaborative space enables discrimination between multiple speakers, as shown in the FreeWalk interface [7].

These systems use fixed computers or audio conferencing systems. Communication tasks in mobile contexts are arguably more challenging, as navigating physical environments competes for the user’s attention, without further burdens of confusion, frustration and communication losses occurring as a result of speaker indistinguishability and miscomprehension in a conference call.

Several researchers have explored the uses of spatial audiovisual cues for mobile contexts. Billinghurst et. al.[3] describe a wearable conferencing space using spatial audio to disambiguate between speakers. Kan et. al. [6] present a laptop-based system using GPS.
to give spatial audio cues based on the actual location of the
speakers relative to the listener.

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. To this end, we describe a user
evaluation comparing spatial and non-spatial audio conditions,
with phone-based and head-based tracking methods for orienting in
a spatialised multi-party audio environment.

3. EXPERIMENTAL DESIGN

We developed a prototype interface to evaluate a virtual spatial
audio environment with ubiquitous multi-party calls. To do this
we used a standard computer, stereo headset, and a custom-built
mobile phone mock-up device that has a small color LCD screen
(260 x 234 pixel resolution) and a phone keypad (see figure 1).
The phone mock-up is connected via USB and video cables to the
PC that runs the sample mobile application and shows the output
on the LCD display. This enables rapid application development
and testing on the PC. Both the stereo headset and the phone had
Intersense\textsuperscript{1} tri-axis digital gyroscopes for motion direction
tracking to explore gesture-based interaction.

![Figure 1: Mobile phone mockup hardware](image1)

The prototype hardware enabled us to display both spatial and
non-spatial sounds. For the spatial audio conditions, sound
sources were placed in the horizontal circular plane, with the user,
or ‘host’ of the call, at the centre. We chose to use binaural
processing on stereo headphones, with the application of HRTFs
from the fmod sound library\textsuperscript{2}. The PC was used to generate simple
visuals shown on the LCD screen representing the conferencing
space. When a call came in a numbered square appeared on the
screen representing the caller. As the user turned his or her head
in the head tracked case a triangle rotated showing them the
relative orientations of the users to them. Figure 2 shows the
hardware configuration for user testing.

![Figure 2: Experimental apparatus configuration](image2)

3.1 The Audio Conditions

Using the hardware described above we conducted a user study
involving the following five audio modes; mono (control), fixed
spatial audio, head tracked spatial binaural audio, and phone-
tracked binaural spatial audio. These are described further below:

\textbf{A: Mono}

This mode gave the listener all mono-sound sources originating
from the same spot, so that the voices were ‘on top of each other’.
In this case the sound was not spatialised.

\textbf{B: Fixed spatial}

In this spatial mode, multiple sound sources are arranged within
the user’s audio sphere. The audio plane is centred on the user, and
sound sources are arranged radially and equidistant from the user
(see figure 3). There were two configurations: 5 speakers or 3
speakers. In the 5-speaker condition, the sources were placed at the
angles: -90/-45/0/+45/+90 positions. With 3 speakers, audio
sources were placed at angles: -45/0/+45.

![Figure 3: Spatial Audio Locations](image3)

\textbf{C: Head tracking}

Multiple sources were spatially arranged in a semi-circle in front of
the user, in the same configuration as in the fixed spatial mode. A
head-mounted inertial tracker mapped the user’s head rotation to
rotation in the sound space.

\textsuperscript{1} www.intersense.com

\textsuperscript{2} www.fmod.org
D: Phone tracking

The phone tracking condition was similar to the head tracked condition, except an inertial compass was mounted on the phone and the phone rotation used as input into the sound space. The purpose of this was to explore the impact of removing the act of rotation from the user’s head, which is the normal frame of reference for hearing, while retaining spatial input.

E: Volume filter

In this condition rather than using a spatial sound model we selectively filtered competing audio streams by boosting the volume of the desired stream. The users head was tracked with an inertial compass, but the tracker information was used to select the audio stream to play at full volume (while reducing the volume of the others), rather than implementing a full spatial audio model. We included this mode in order to gauge the extent of the effect volume had on the particular task to be performed.

3.2 Experimental Procedure

In this experiment we wanted to simulate the experience of a multi-party phone call. To do this we used pre-recorded audio streams so that we could have complete control over the audio content. These audio streams were fixed in palace for the purposes of this study. The listener’s rotation about the vertical axis, via either the head or phone tracker, is translated to the corresponding binaural directional change for the user to hear the spatial sound.

The experimental task was to identify a word in the middle of a spoken phrase spoken. There were 4 topics, with 9 phrases each – differing only in the target key word. Each phrase was recorded as spoken by five different speakers, some of whom were non-native English-speakers. When the phrases were played back they were all played at the same time.

Twelve participants (nine male and three female) were presented with each of the five audio modes in random order, and either the 5-speaker condition first or the 3-speaker case. Before each condition participants were asked to familiarise themselves with the technology for the condition. They were told to listen to one particular speaker and trained on their voice so they could identify the voice. All the speakers had different accents and so were easy to identify.

Each subject was then asked to listen to the group of 3 or 5 speakers, identify their target speaker, and the key word spoken. The users hit a key to say they were ready for the task, then all the speakers spoke at once, the subject had to say the word aloud, and it was recorded by the experimenters. The time taken to understand the keyword was logged via the software, with control from the experimenters. Immediately following this, the subject was asked to identify the perceived angle (shown as squares on the PC monitor and phone screen) of the target speaker. This was recorded, and the process repeated three times for each condition. After data was collected for each condition the subjects also filled out a subjective survey giving responses on a number of questions such as how easily they were able to understand the speakers and how intuitive the interface was.

4. RESULTS

There was a significant difference between the number of correctly understood speakers in each condition for 5 speaker conferences and nearly significant for 3 speakers. Figure 4 shows the average (out of 3 phrases), of correctly understood speakers.

![Figure 4: Correct responses across conditions (3 & 5 speakers)](image)

Using a one factor ANOVA we found for 3 speakers, F(4,55) = 2.05, p = 0.1) and for 5 speakers, F(4,55) = 7.11, p< 0.001) From figure 4 it is clear that the number of correct responses is much lower in the non-spatial case (Condition A), especially for five speakers, while the spatial audio cases provide almost the same number of correct answers.

There was also a significant difference in the time taken to complete the task depending on the audio interface. Figure 5 shows the average time for the users to identify all of the spoken phrases in each condition. As can be seen the subjects took longer in the phone-tracked case, especially when trying to distinguish between three speakers. A one factor ANOVA found for 3 speakers, F(4,55) = 4.74, p < 0.01), but no difference between the five speaker case (F(4,55) = 1.59, p = 0.19).

![Figure 4: Average time taken for task completion](image)

Figure 5 shows the average number of sectors deviated from target sector, for each set of 3 phrases in each of the 3 or 5 speaker spatial audio conditions. Subjects were asked to identify where they thought the speakers were relative to them. As can be seen subjects were more accurate in condition E using the volume filter, and accuracy decreased between 3 and 5 speakers. Most of the responses were, on average, less than half a sector off the target.
We also asked subjective questions about how easy it was to distinguish between speakers, how understandable the speakers were, and how much they felt that the speakers were really distributed around them in space. For each of these questions subjects rated response on a scale of 1 (very easy) to 5 (very difficult) or in the case of speaker distribution 1 (no spatial sense) to 5 (immersed in spatial environment). The table below shows the average scores for each of these questions.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinguishability</td>
<td>4.9</td>
<td>3.4</td>
<td>2.4</td>
<td>2.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Understandability</td>
<td>5.0</td>
<td>3.3</td>
<td>2.3</td>
<td>2.6</td>
<td>1.8</td>
</tr>
<tr>
<td>Sense of Space</td>
<td>1.1</td>
<td>2.8</td>
<td>4.0</td>
<td>3.4</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 1: Subjective survey scores.

In both the ability to distinguish between speakers and understand speakers the mono audio case (condition A) was ranked the worst, while the volume-filtered case (condition E) was rated highest. There was a significant difference between these results. For distinguishability ($F(4,55) = 20.7, p < 0.001$) and for understandability ($F(4,55) = 21.0, p < 0.001$). When asked about sense of space the mono case rated worst, and the head tracked condition (condition C) best. There was also a significant difference here ($F(4,55) = 26.4, p < 0.001$).

4.1 Discussion

These results show the value of spatial audio cues in a mobile conferencing application and in particular using a volume filtering technique in conjunction with head tracking (Condition E). Users reported that this was the easiest condition in which to distinguish between speakers and understand what they were saying, and they were more accurate correctly locating the speakers (figure 5).

The inherent limitations of listening to competing sounds sources showed a consistent effect over the spatial modes (fixed and tracking modes showed similar response times as mono sound). Volume-tracking effectively alleviated this by enabling a single source to be filtered, leading to the lower times.

Tracking the users phone rather than their head position provided some benefit (especially compared to no spatial cues, or fixed spatial (Condition B)), but in some cases it took users longer to complete the task and locate the speaker (figure 4). This may have been due to the extra effort required to move the phone rather than the head and map the phone direction to speaker location. In several cases users had to rotate the phone clockwise and ant-clockwise a few times to distinguish the speaker.

All of the tracked conditions (C,D,E) made the subjects feel that they were more immersed in a spatial sound environment than the non spatial and fixed spatial conditions. It is interesting that even though the volume filtered case does not have an accurate spatial sound model. The only disadvantage of the volume-filtered method is that it does not allow any peripheral awareness of other sound sources, and therefore the hearer would not detect changes to the audio environment.

5. CONCLUSIONS

We found spatial audio modes using head tracking and mobile phone tracking enabled better discrimination between speakers than fixed spatiality 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 phone tracking as a viable option for orienting to speakers in mobile virtual spatial audio environments, as well as raising a number of issues in employing spatial sound in this multi-party conversation context that will form the basis of future explorative studies.

We found motion tracking of the head and mobile device to be comparable for speech intelligibility with multi-party voices. The mapping of physical rotational movement to that in the audio sphere enables a mechanism for orienting in the soundscape, and potentially augmenting the perceived spatiality of the audio, without the need for specialised headphones. This could provide disambiguation of speakers in multi-party conferences, and affords further exploration for gesture-based spatial audio augmentation in mobile multi-party calling scenarios.

Our results also suggest that simple volume filtering cue can be as effective, if not more so, as using accurate spatial audio models and that users may still perceive themselves to be immersed in a spatial environment in this case.

In the future, we will study methods for user-controlled placement and adjustment of multiple sound sources within their audio sphere. This will focus on the mobile device, and will compare gesture-based placement and adjustment of speaker positions to standard techniques, such as keypad input with a visual display. We will also explore the combination of attenuation boosters with the spatial audio environment in providing the benefits noted, both in our study and in [1], with volume enhancement of the speakers of interest.

REFERENCES


