An Empirical Study of Hear-Through Augmented Reality: Using Bone Conduction to Deliver Spatialized Audio

Robert W. Lindeman[*1,2] Haruo Noma[†2] Paulo Gonçalves de Barros[‡1]

[1]Worcester Polytechnic Institute
Dept. of Computer Science
100 Institute Road
Worcester, MA 01609, USA

[2]ATR International
Knowledge Science Labs
2-2-2 Hikari-dai, Seika-cho, Souraku-gun
619-0288 Kyoto, JAPAN

ABSTRACT

Augmented reality (AR) is the mixing of computer-generated stimuli with real-world stimuli. In this paper, we present results from a controlled, empirical study comparing three ways of delivering spatialized audio for AR applications: a speaker array, headphones, and a bone-conduction headset. Analogous to optical-see-through AR in the visual domain, *Hear-Through AR* allows users to receive computer-generated audio using the bone-conduction headset, and real-world audio using their unoccluded ears. Our results show that subjects achieved the best accuracy using a speaker array physically located around the listener when stationary sounds were played, but that there was no difference in accuracy between the speaker array and the bone-conduction device for sounds that were moving, and that both devices outperformed standard headphones for moving sounds. Subjective comments by subjects following the experiment support this performance data.

Keywords: Augmented reality, audio, bone conduction

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems-Audio input/output; Artificial, augmented, and virtual realities

1 INTRODUCTION

Augmented reality (AR) is the mixing of computer-generated stimuli with real-world stimuli. While much work has been done for delivering mixed real-world (RW) and computer-generated (CG) stimuli in the visual domain, we focus here instead on the audio domain. Recently, we introduced two approaches for audio AR: *Hear-Through AR* and *Mic-Through AR* [8]. Our work on Hear-Through AR used either a speaker array or a bone-conduction headset (BCH) to deliver CG audio to a user, while RW sound was received through the unoccluded ear canals. Mic-Through AR captures RW sound using microphones mounted near the ears of the user, mixes it with CG sound in the computer, and delivers the resulting AR sound through standard headphones. In this paper, we present the first results from a formal, empirical study comparing subjects' sound-localization capabilities using both speaker-based and BCH-based Hear-Through AR, as well as Mic-Through AR. In order to gather baseline data, this study used only simple, well-controlled audio tones played at three frequencies. However, both static and moving tones were considered, while the head of the user was kept stationary.

Lindeman & Noma [7] present a scheme for classifying AR techniques for all the human sensory modalities by where the mixing of CG and RW elements takes place. They underscore the need to correctly match the attributes of CG and RW stimuli so that the user can easily fuse the two, thereby improving the realism of the resulting mixed reality. Two main characteristics differentiate RW and CG audio. Real-world audio is typically of higher fidelity than CG audio. Also, computationally expensive preprocessing of CG audio is required in order to subject CG audio to similar environmental effects to match the RW environmental effects. Where the mixing of these elements takes place can have a significant impact on this computational cost.

CG sound can be displayed using speakers placed within the real environment, allowing RW and CG sounds to mix in the environment before reaching the user. Alternatively, Mic-Through AR using two microphones and standard headphones (Figure 4) can be used [8]. Mixing using Mic-Through AR takes place either using an audio mixer or a computer. Hear-Through AR delivers CG sound through bone conduction [3], and RW sound through the unoccluded ear canals. Here, mixing takes place at the cochlea.

Figure 1: The AudioBone bone-conducting headset from Goldendance Co., Ltd.

Commercial bone-conducting devices have recently begun to emerge onto the market. Figure 1 shows the AudioBone MGD-01 produced by Goldendance Co., Ltd., Japan which was used in the current study. With this BCH, vibrational actuators are positioned

[*]e-mail: gogo@wpi.edu
[†]e-mail: noma@atr.jp
[‡]e-mail: pgb@wpi.edu

IEEE Virtual Reality 2008
8-12 March, Reno, Nevada, USA
978-1-4244-1972-2/08/$25.00 ©2008 IEEE

on the zygomatic (cheek) bones in front of the ear. The unit has a normal output of 30mW, a maximum of 70mW, a normal impedance of 8 ohms, a sound-pressure sensitivity of 80 dB/mW (dB 1.0 dyne), and a standard operating frequency of 50Hz-4kHz. The total weight of the unit is 60g. The headband wraps around the back of the head, and the ear loops rest on the tops of the pinnae. Another innovative use of bone conduction is for listening to music while swimming (http://www.finisinc.com/), because bone conduction does not require sound to pass through air.

In this paper, we explore how well people can successfully localize audio signals using a speaker array, a BCH, and standard headphones. The ability to localize audio in AR settings is especially useful in situations where social aspects are important, such as audio tour guides for museums [1], or where the user is otherwise engaged in workplace tasks [12]. Our longer-term goal is the creation and delivery of authentic audio AR stimuli. That is, we aim to produce a stimulus that combines computer-generated or mediated audio with environmental audio in such a way that users will not be able to distinguish between them.

2 Bone Conduction

The human auditory system is comprised of structures in the outer, middle, and inner ear, which transform sound waves and stimulate nerves that cause the brain to make sense of the sound. While most audio displays to date use signals delivered through air (air conduction) to stimulate our sense of hearing, the recent emergence of consumer-grade, bone-conduction audio devices now makes it feasible for us to deliver some signals through the bony structure of the skull. These signals bypass the outer and middle ear, leaving the ear canals unobstructed to receive signals from the real world. Signals arriving at the inner ear from these two channels are then mixed, and the resulting sound is what is delivered to the brain.

The effect of bone conduction is apparent to anyone who has heard a playback of his or her own voice. Because the human voice box produces both audible sound that leaves our mouths and arrives at our own ears through the air, as well as vibrations that reach our inner ear through our skulls, what we hear from the recording (only the air-conducted sounds) is very different from what we hear when speaking. How effective this bone-conduction channel is for spatialized audio is the focus of our current work.

3 Background and Related Work

For air-conducted audio, spatial properties of audio cues have successfully been captured using Head-Related Transfer Functions (HRTFs) [18, 15], which incorporate such individual differences as the shape of the torso, head, and ears [11, 4], as well as the position of sounds with regard to the listener [17]. However, due to the individualized nature of HRTFs, those captured for one person might not be effective for all listeners. Raykar *et al*. showed that the contribution of the different environmental factors (*e.g.*, head, torso, pinnae) can be identified using an HRTF along with its corresponding time-domain head-related impulse response (HRIR) [11].

Early bone-conduction audio work focused mainly on applications for individuals with outer- or middle-ear impairments, employing actuators placed either on the surface of the mastoid bone behind the ear, or attached to a surgical implant anchored to the skull bone. Recently, there has been significant work on the use of bone-conduction devices for non-clinical applications, as the technology for delivering the signals to users has become available in consumer-grade equipment, which tends to be less expensive than devices designed for clinical applications. Fukumoto and Tonomura describe a novel interface for cellular phones, where a wrist-worn, bone-conducting actuator passes audio to the ear of the listener when the listener puts his or her finger into the ear canal [3]. They describe three alternatives for placement of the actuator, and address the usability and sociological implications of each.

Walker and Lindsay have proposed the use of bone-related transfer functions (BRTFs) for use with bone-conduction devices [15]. These could potentially allow audio designers to provide spatialized cues using bone conduction. In an augmented reality environment, the fact that the ear canals are left unobstructed means we can better take advantage of high-fidelity real-world audio, which was not a viable option for a general solution to augmented audio reality in the past.

One problem with the use of bone conduction has to do with the potential for crosstalk. This occurs because vibrations from the transducer on one side of the head not only reach the near cochlea, but also the cochlea of the far ear. However, MacDonald *et al.* (2006) showed that the interaural level difference and the interaural time delay of signals displayed using bone conduction are sufficient for users to achieve localization performance similar to standard headphones [9]. The study by MacDonald *et al.* used individualized HRTFs for spatialized headphone and bone-conduction audio, and measured how well four listeners could judge the source of a sound played at one of eight locations equally spaced around the listener in the horizontal plane of the head. Their results showed that performance was similar for both devices.

4 Perceived Audio

Precisely controlling the perceived stimulus displayed to any sensory modality is a very difficult task, as many factors can alter a signal along the path from the computer to the user. Similar to work proposed by Niwa *et al.* in the haptic domain [10], we can define a transfer function J that maps from an input state *t* to an output state *s*, where *t* is the control signal to the audio device and *s* is the perceived sound at the cochlea of the user (Figure 2a). In a "perfect" system, the stimulus presented to the user would be identical to the one perceived ($s = t$), so J would simply apply *t* to the cochlea. In reality, as sound travels from an output source to the cochlea, it undergoes changes due to interactions with materials in the environment (*e.g.*, occluders, reflectors) and characteristics of the listener's body (*e.g.*, head shape, shoulder slope). In addition, some materials affect certain frequencies of sound differently than others, so the effects are not uniform with regard to sound frequencies. Considering these effects, we can decompose J into two components, G and H, and model the process of how a given audio device allows us to perceive sound as shown in Figure 2b.

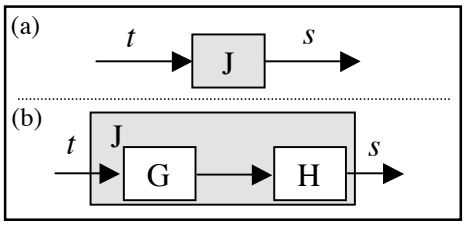

Figure 2: Transfer functions for audio (a) modeled as a single function, and (b) modeled as a series of functions affected by various factors.

An input state variable for the transfer function G is the sound wave signal for the audio display, and the output state variable represents the resulting sound characteristics, such as the result of calculating how the signal interacts with occluders in the environment, or the result of applying a set of HRTFs. Variable factors are mainly the properties of the physical environment and user: geometry location and head shape, for example.

The output state variable for the transfer function H is the auditory sensation *s*: how the user hears the sound. The transfer function H is related to the method of display, as well as the mental and physical condition of the user. The properties and type of speakers/headphones/BCH used to display the sound will alter what stimulus the user actually *hears*, as opposed to what she was *meant* to hear. If we could correctly encode and deliver a stimulus directly to the auditory nerve, the state of the sound would correspond one-to-one to the state of the applied stimulus, as illustrated in Figure 2a.

4.1. Steps in the AR-Sound-Delivery Process

One of the main differentiators between the three types of audio AR studied here have to do with how the audio signals are altered on their way to the listener (H in Figure 2b). In the visual AR domain, CG objects placed within the context of the user's view of the real world should attempt to mimic the lighting and environmental effects present in the real world. For example, if a CG object is lit from the opposite direction as the real objects, or if a shadow cast into the scene does not affect the lighting of the CG object [6], believability is sacrificed. A similar situation exists for audio, where real-world audio emanating from a particular location undergoes certain transformations on its way to the listener. If CG audio does not take into account these same environmental effects, confusion in the listener may occur. The most straightforward effects that can be incorporated are distance and lateralization cues [17], but other effects, such as sound dampening and reflection from objects or structures in the environment, can also greatly affect sound believability [14].

Each of the three audio AR approaches we are studying starts with RW and CG sounds, processes them, and mixes them in some fashion before the resulting AR sound can be interpreted by the brain. We first consider the Hear-Through approaches, followed by the Mic-Through approach. In both the speaker-based (Figure 3a) and BCH-based techniques (Figure 3b), RW sounds follow the same path. Sounds emanate from a source and interact with environmental objects and the listener's body on their way to the outer, middle, and, finally, inner ear. The cochlea is thus stimulated, and sends sound signals along the auditory nerve to the brain.

For the case where a set of speakers in the environment is used to deliver CG sounds (Figure 3a), the CG sound is preprocessed to apply effects, such as HRTFs and cross-talk cancellation [4], before being delivered into the physical environment through the speakers. At this point, the CG sound mixes with the RW sound, and follows the same path to the listener's brain.

For the BCH system (Figure 3b), the CG sound is again preprocessed to apply effects, such as HRTFs and reverb, before being delivered to the cheekbones of the listener through the BCH device. The skull vibrations in turn stimulate the cochlea, where the mixing with RW sounds takes place. The combined AR sound is then delivered to the brain along the auditory nerve.

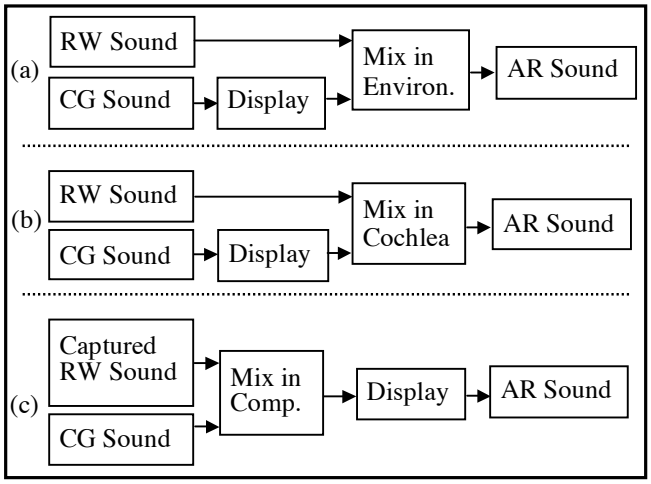

Figure 3: Path of sounds for (a) speaker-based, Hear-Through AR, (b) bone-conduction-headset-based, Hear-Through AR, and (c) headphone-based, Mic-Through AR

For the Mic-Through approach (Figure 3c), the RW sound goes through a more-complex set of steps before being delivered to the listener. In our current configuration, two channels of RW audio are captured using two microphones, each positioned at the opening of the ear canal of the listener (Figure 4). Similar to the RW sound path for the Hear-Through approaches, RW sounds interact with the environment and the listener's body before reaching the microphones. The audio can then (optionally) be post-processed to, for example, adjust the loudness of the signal, perform noise cancellation, and the like, before being mixed with the CG sound in the computer, and displayed through headphones. The mixed RW and CG sound then passes through the ear canal and middle ear to the inner ear, stimulating the cochlea, and reaching the brain through the auditory nerve.

As has been done in the visual domain [6], the captured RW audio could also be analyzed in order to extract information about the environment, which could then be used to guide transformations of the CG sound. For example, if the left channel of the captured sound is louder than the right, then the CG sound could be adjusted to place CG crowd noise off to the left as well, under the assumption that people are grouped together. We are unaware of any work done in this area for audio, but think it is a rich area of research.

Figure 4: Microphones for capturing audio clips (mounted on ear-bud headphones).

4.2. Analysis of the Three Approaches

An advantage of audio mixed in the environment, for example using speakers, is the fact that both the CG and RW audio interact directly with the physical environment. This reduces the computational cost incurred when considering how CG audio is transformed by modeling the geometry of the physical space.

An advantage of mixing the audio in the computer (Mic-Through using headphones) is that the system has complete control over the entire sound experience. This would allow, for example, captured RW audio to be further processed to account for things like virtual occluders (CG geometry placed in the physical environment), or virtual surfaces that reflect sound in different ways than objects physically present in the scene [14]. A disadvantage is the additional computational cost needed to achieve these effects. Also, because the environmental audio is captured (as opposed to synthesized), the possible transformations are more limited than for CG audio, as it is more difficult to, for example, identify and manipulate individual parties from the captured stream. Once the RW audio has been captured, however, it can also be transmitted to a remote site, and used as additional spatialized audio channels for remote collaboration applications.

Mixing at the cochlea using the BCH has the advantage of using the simplicity and high-fidelity of RW sound, together with the privacy provided by headphones. Like headphones, others cannot hear the CG sound when played through the BCH, so the audio is private, allowing different users to be presented with different CG audio. Unlike headphones, however, the ear canals are free to receive the high-fidelity RW sound, providing collocated users with the shared experience of the real world at no extra cost. Anecdotally, the BCH does not seem to provide as rich of an audio experience as standard headphones or speakers, and is especially weak in the low frequency (*i.e.*, bass) range. The BCH does, however, seem to be well suited for voice audio.

5 Empirical User Study

We performed a user study to compare how well people can localize audio using bone conduction versus traditional headphones or a speaker array. The study involved both stationary and moving sound cues of varying frequencies. We hypothesize that sound-localization accuracy will be highest with the speaker array, as this method most-closely resembles how human hearing ability has evolved. The larger dynamic range of headphones, coupled with the lack of familiarity of people with the BCH, will make accuracy with the headphones better than with the BCH. It is unclear how the various frequencies will impact sound localization accuracy. The goal of the user study is to develop some baseline data using simple audio cues with a stationary listener. Later studies can then explore the use of Hear-Through and Mic-Through AR in more-realistic contexts, using more complex audio and moving users.

5.1. Subject Demographics

Twenty-four uncompensated students, 22 male and 2 female, in a university computer science department participated in the study. We collected some basic demographic information prior to the study. Subject ages ranged from 20 to 30 years (mean: 24 years, 7 months), and all reported having "normal" (13) or "good" (11) hearing ability. Ten subjects reported daily use of headphones, while six reported weekly use, and eight reported infrequent headphone use.

5.2. Method

This was a 3×3×2 within-subjects, factorial design, with the independent variables being audio device, stimulus frequency, and stimulus motion. Audio device had three levels: BCH (B), Headphones (H), and Speaker array (S). Stimulus frequency had three levels: 200Hz (LOW), 500Hz (MED), and 1kHz (HIGH), and each was a square wave with a 44.1kHz sampling rate and 32-bit depth. The frequencies were chosen for their attenuation predictability within the range of 150Hz-1kHz [16]. Stimulus motion had two levels: Stationary (STAT) and Moving (MOV).

After signing an Institutional Review Board (IRB) approved human-subjects consent form, subjects performed 63 trials three times, once under each audio-device condition, B, H, and S. Each trial consisted of either a stationary or moving tone at one of the frequency levels. Stationary tones emanated (either physically or virtually) from one of five equally-spaced locations around the subject (Figure 5). Moving tones emanated from each of the five locations in sequence from either left to right or right to left. For both STAT and MOV, total stimulus time was one second. After the tone was played, the user responded with either "*Left*", "*Center-Left*", "*Center*", "*Center-Right*", or "*Right*" for STAT, and "*Left-to-Right*" or "*Right-to-Left*" for MOV, depending on the perceived position of the tone, or direction of movement.

There were 21 possible combinations of position/direction (7) and frequency (3), and each combination was presented three times, making up the 63 trials. Each subject's responses on the three repetitions for each combination were averaged to give an accuracy percentage for each combination of audio device, frequency, and position (direction), for a grand total for all the subjects of 1,512 data points. Subjects were blindfolded during each condition, though for each condition, the display devices (speakers, BCH, or headphones) were seen prior to donning the blindfold. The order of presenting the conditions (B, H, and S) was varied for each subject, and the order of the trials was randomized for each condition run in order to minimize confounds to validity.

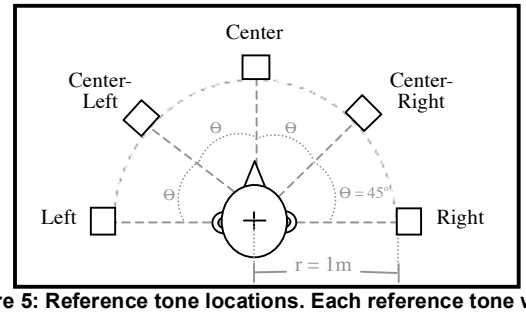

Figure 5: Reference tone locations. Each reference tone was either played through the corresponding speaker (S), or was captured from the speaker prior to the study, and replayed during the study (B & H).

5.3. Apparatus

For the S condition, the tone for a given trial was played through one of five speakers (Excel Sound ESP-757PW) equally spaced around the subject at a distance of 1 meter from the center of the body, and a -10º elevation from the ears. For the trials with a moving stimulus, a manual switch-box was used to "move" the signal from speaker to speaker. For the H condition, the tone was played through a pair of headphones (Grado SR-60), and for the B condition, the AudioBone was used. No HRTFs were used in this study. Other than the experimental sounds, the experimental

environment was kept silent. In preparing the experiment, 21 sound samples (7 positions/directions and 3 frequencies) were captured in a soundproof room from the speaker array using a pair of powered omnidirectional microphones (Sony ECM-C115) anchored to the outside of inactive ear-bud headphones (Figure 4) worn by one of the investigators. Each captured sample was adjusted in order to reduce the variability in volume of all the samples. The resulting samples were then used for both the B and H conditions, while the original tones were used for S.

5.4. Experimental Procedure

The basic process for the study was for subjects to listen to a one-second tone, and to indicate the location of the source of the tone in space. This was repeated 63 times for each audio device. Prior to the B condition, a brief explanation of bone conduction was given, in order to familiarize the subject with the technology. At the start of each condition, the subject was seated, and the appropriate headset was donned (B and H), along with the blindfold. A calibration procedure was performed prior to each condition, whereby each tone was played at each speaker position, and the subject was asked to adjust it to a comfortable volume, as well as to equalize the loudness across samples. Based on verbal commands from the subject, the volume level for each individual sample was manipulated by the experimenter, and subsequently used during playback of that sample for that condition. In addition, the subject adjusted the positioning of the headphones and BCH during this period. This procedure also allowed the subject to experience each spatialized audio sample prior to performing the actual experiment.

Following calibration, a random sequence of the 63 trials was presented. The subject was instructed to look straight ahead, and to keep his/her head as still as possible during each condition. Each one-second reference sample was output from the headphone jack of a Macintosh PowerBook G4 at the volume level determined during calibration for that sample. After each reference tone, the subject answered with the location or direction that best described the signal. After the answer had been recorded, the experimenter triggered the next trial. Upon completion of the condition, the subject was told to take off the blindfold and headset (B and H), and rest until ready to continue with the next condition. After all conditions had been completed, an informal interview was conducted to collect subject impressions and answer questions about the relative effectiveness of the three devices. A complete session lasted between 45 and 60 minutes.

5.5. Results

The results of the main effects are shown in Table 1 and Table 2. These tables respectively show the mean percentage of trials where subjects correctly identified the position (STAT) or direction (MOV) of the reference tone, sorted by device and frequency. The standard deviation for each value is shown in parentheses. A 95% (α=0.05) confidence level was used to determine statistical significance for all of our analyses. There were no significant differences based on audio-device presentation order, reported hearing ability, or headphone usage.

5.5.1. Stationary Tones

A total of 1,080 data points (5 positions × 3 frequencies × 3 devices × 24 subjects) made up the trials for STAT. An analysis of variance (ANOVA) of the mean accuracy values on the main effects showed statistically significant differences in accuracy for both device [$F(2,1071)=79.43$, $p<0.0001$] and frequency [$F(2,1071)=5.77$, $p<0.004$], and no interaction effects [$F(4,1071)=0.66$, $p=0.618$]. For device, a TukeyHSD post-hoc analysis, used to divide main effects with more than two levels into groups with statistically different means, showed that subject accuracy was statistically different for each of the three devices, with subjects performing best with S, mean=90% accuracy (sd=23%), second best with H, 68% (36%), and worst with B, 61% (36%). In terms of frequency, a TukeyHSD revealed significantly better accuracy with LOW and HIGH, 75% (34%), than for MED, 68% (36%).

Table 1: Summary results of mean percent correct (std. dev.) by audio device and frequency for stationary tones.

		Frequency			
		LOW (200Hz)	MED (500Hz)	HIGH (1kHz)	Total
Audio Device	B	62.2 (36.90)	55.0 (34.20)	65.3 (37.02)	60.8 (36.22)
	H	70.8 (35.26)	61.4 (36.41)	70.3 (34.52)	67.5 (35.57)
	S	91.9 (18.83)	87.5 (27.35)	89.4 (22.45)	89.6 (23.15)
	Total	75.0 (33.73)	68.0 (35.69)	75.0 (33.54)	72.7 (34.46)

5.5.2. Moving Tones

A total of 432 data points (2 directions × 3 frequencies × 3 devices × 24 subjects) made up the trials for MOV. An ANOVA of the data on the main effects showed statistically significant differences in accuracy for device [$F(2,432)=6.46$, $p<0.002$], but not frequency [$F(2,432)=0.68$, $p=0.510$], and there were no interaction effects [$F(4,432)=0.55$, $p=0.698$]. For device, a TukeyHSD analysis showed that subject accuracy was statistically better for S, 100% (0%), than H, 94% (21%), but that there was no statistical difference between S and B, 97% (11%), nor between B and H.

Table 2: Mean percent correct (std. dev.) by audio device and frequency for moving tones.

		Frequency			
		LOW (200Hz)	MED (500Hz)	HIGH (1kHz)	Total
Audio Device	B	98.6 (6.73)	97.2 (11.67)	95.8 (14.79)	97.2 (11.49)
	H	95.83 (17.70)	91.67 (25.26)	95.14 (18.18)	94.2 (20.61)
	S	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
	Total	98.1 (10.99)	96.3 (16.31)	97.0 (13.61)	97.1 (13.80)

5.6. Discussion

A summary of the statistically significant results can be found in Table 3. Levels within a given circle were **not** statistically different from others within the same circle (*i.e.*, those levels came from the same population). We can see a noticeable difference between subject accuracy with regard to STAT versus MOV in terms of audio device. For STAT, subjects approached 90% accuracy when the speaker array was used, but only 68% and 61% with the headphones and BCH, respectively (Table 1). For MOV, there was no difference between performance with the speaker array and BCH, or between the BCH and the headphones, though using the speaker array improved subject accuracy by about 6% over the use of headphones (Table 2).

Table 3: Summary of statistical significance of the main effects. Circles enclose levels shown to come from the same population (*i.e.*, no statistical difference). (*ns* = not significant)

	Stationary	Moving
Audio Device	(S) (H) (B)	(S (B) H)
Frequency	(HIGH LOW) (MED)	*ns*
Interaction	*ns*	*ns*

In terms of frequency, subjects seemed to have less trouble with tones at the 200Hz and 1kHz frequencies, compared to tones at 500Hz on STAT (Table 1), though there was no clear difference for MOV (Table 2). The latter could be due to a ceiling effect, as overall accuracy was quite high.

We saw no differences with regard to the direction of movement for the MOV trials (Table 4). However, looking more deeply at the errors committed during STAT (Table 5), we can see that by far, subjects had greater difficulty when presented with Center-Left and Center-Right than the other reference positions across all device types and frequencies. Figure 6 shows the reference position accuracy percentages by device type and frequency. The values for B are shown using solid black lines and icons, the values for H using short stipple and white icons, and the values for S using long stipple and grey icons. The values for LOW have diamond-shaped icons, MED uses squares, and HIGH uses triangles.

Table 4: Mean percent correct (std. dev.) by audio device and direction for moving tones.

		Reference Direction		
		Left-to-Right	Right-to-Left	Total
Audio Device	B	97.2 (10.83)	97.2 (12.19)	97.2 (11.49)
	H	94.0 (23.29)	94.4 (17.69)	94.2 (20.61)
	S	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
	Total	97.1 (14.96)	97.2 (12.55)	97.1 (13.80)

Table 5: Mean percent correct (std. dev.) by audio device and position for stationary tones.

		Reference Position					
		Left	Center-Left	Center	Center-Right	Right	Total
Audio Device	B	65.3 (32.35)	49.1 (38.75)	78.2 (29.16)	36.1 (32.50)	75.5 (29.60)	60.8 (36.22)
	H	83.8 (25.02)	57.4 (38.85)	84.3 (26.22)	44.4 (34.94)	67.6 (34.02)	67.5 (35.57)
	S	94.9 (14.44)	75.5 (34.48)	97.7 (8.53)	91.2 (21.66)	88.9 (21.67)	89.6 (23.15)
	Total	81.3 (27.77)	60.6 (38.84)	86.7 (24.46)	57.3 (38.70)	77.3 (30.07)	72.7 (34.46)

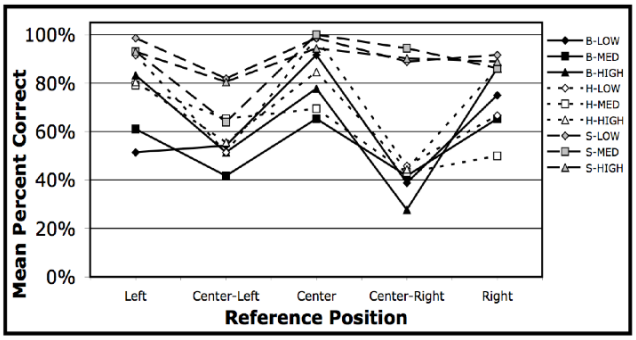

Figure 6: Graph of mean percent correct by audio device, frequency, and reference position.

Table 6 shows the mean **relative** error for each reference position. As in MacDonald *et al.* [9], we calculate localization error as the angular distance between the perceived and actual (reference) locations in degrees. The relative error shows the direction of the error, with negative numbers indicating errors to the left of the reference tone. It is clear from this data that when in doubt, subjects tended to err away from the Center position, and towards the Left or Right extremes. MacDonald *et al.* (2006) refer to this as subjects being able to correctly *lateralize* the sounds, but not necessarily *localize* the sounds [9]. Lateralization means subjects knew from which side of the head the sounds came, but had difficulty being more precise than that. While lateralization information is useful, it is much less expressive than localization information.

Table 6: Mean *relative* angular error in degrees by reference position.

		Reference Position				
		Left (-90°)	Center-Left (-45°)	Center (0°)	Center-Right (45°)	Right (90°)
Audio Device	B	19.4	-7.9	1.3	14.4	-12.9
	H	8.1	-14.6	-8.5	3.3	-19.2
	S	2.3	-11.0	0.6	4.0	-5.0

Table 7: Mean *absolute* angular error in degrees by reference position.

		Reference Position					
		Left (-90°)	Center-Left (-45°)	Center (0°)	Center-Right (45°)	Right (90°)	Overall
Audio Device	B	19.4	23.8	14.6	29.4	12.9	20.0
	H	8.1	19.6	9.4	27.1	19.2	16.7
	S	2.3	11.0	1.0	4.0	5.0	4.7

Table 7 shows the mean **absolute** error for each of the reference locations. Our overall error rates for B and H (20.0 and 16.7, respectively) are in line with those found by MacDonald *et al.* (17.0 and 21.7), though we found H to be better than B. We assume this is due to the fact that their work presented reference tones in both the front and rear hemispheres around the user, leading to increased front-back or left-right reversals, as reported

in the paper. Since our work only presented reference tones in the front hemisphere, we had fewer reversals. As hypothesized, the mean localization error for S (4.7) is considerably lower than the other two conditions.

It is clear from our results that this particular study showed a marked superiority of S for allowing subjects to localize reference tones, especially for STAT. Because this study concentrated on producing mostly base-line data, using stationary subjects and simplistic, synthetic audio tones, as opposed to using more-realistic sounds, we nevertheless see promise in the use of the bone-conduction device. Most sounds we hear in the real world are more complex than those used in our current study, providing many more cues listeners use for localization, such as distance attenuation [17]. Furthermore, listeners typically do not keep their heads still, so the positive results obtained from MOV lead us to believe that Hear-Through AR could provide a good balance between CG-audio expressiveness and computational cost. Compared to S, the BCH device has much broader applicability, as it is a wearable solution, providing both audio privacy and greater mobility.

6 Comments from Subjects

Upon completing all the conditions, each subject was asked for any comments about 1) the speaker array, 2) the headphones, and 3) the BCH. The comments were generally in line with our findings for subject performance (Figure 6).

Of the 24 subjects, 11 mentioned having problems identifying the Center-Left and Center-Right tones for B, while only eight commented on the same problem with H, and only two for S (some subjects mentioned this problem for both B and H, or for all three conditions).

Nine subjects self-ranked the conditions from "liked-best to liked-least" or "easiest to hardest" as S-H-B, while one subject ranked them B-H-S. In terms of frequency, four subjects commented that the high-frequency tones were hardest, while two felt that low-frequency tones were easiest. Three subjects commented that moving tones were easier than static tones.

7 Conclusions

Our results show that subjects could successfully lateralize sound using the BCH and headphones, but had considerable difficulties localizing the sound. In this study, we opted to use captured spatial audio rather than use HRTFs. While we feel that captured audio should be used for RW sounds, in order to leverage the quality of the sample, we also feel that HRTFs should be used for CG sound, as others have found that localization performance using HRTFs to be indistinguishable comparing headphones to BCH devices [9].

Another conclusion has to do with the frequency response of the BCH unit used in our study. While the unit we used has a higher frequency response range than units used in previous studies (AudioBone=50Hz-4kHz vs. TEMCO-HG17 = 300Hz-3kHZ), this is still significantly lower than typical headphones (*e.g.*, Grado-SR60=20Hz-20kHz). Several companies have recently released bone-conduction headsets with specifications more in line with traditional headphones. For example, the AudioBone Aqua (Figure 7) provides a frequency response of 50Hz-12kHz, and the TEAC HP-F100 Filltune unit boasts a frequency response range of 25Hz-25kHz. These new units will better support user studies that employ complex, real-world sound, including both background and voice audio, and we plan to use them in follow-on studies.

Figure 7: The AudioBone Aqua with improved frequency response.

Another possible confound to our study has to do with the captured sound itself. As described in Section 4, there are many factors that influence a signal on its way from the source to the brain. Each of our sound samples was generated by a computer, sent to a speaker array, interacted with the capture environment, captured by ear-worn microphones, and stored onto a computer. Each of these steps influences the spatial parameters of the captured sound, and hence should be handled in a controlled way. AR environments, on the other hand, are notoriously uncontrolled and chaotic, so it is difficult to control all of these. For example, pre-processing of CG audio to take into account reflection of objects or structures in the real environment requires a computer representation of that environment. As AR users are often mobile and moving about in dynamic scenes, precise knowledge of the surrounding environment is not available. Therefore, general approximations could be applied in place of perfect knowledge. For example, audio played in indoor scenes could apply softening filters to CG sounds. Another approach would be to analyze audio captured at runtime for properties of people with whom the user is currently engaged in conversation, in order to extract spatial information, such as lateralization information.

Cross-modal analysis could also be employed. Since many graphical AR systems use vision-based tracking, GPS, and digital compasses to extract location and orientation information, the same information could be used to better determine current surroundings for audio rendering.

If Mic-Through AR is to be used for a given application, then choosing the combination of capture and display hardware is important. Because both supra-aural and circum-aural headphone designs directly interfere with the sound waves reaching the pinnae, microphones mounted on ear-bud headphones, such as those used by Härmä *et al*. [5], should allow more of the spatial components contributed by the pinnae to be captured [11], thereby improving localization ability.

8 Future Work

We can envision several interesting follow-on studies to tease out the most applicable situations for BCH-based Hear-Through AR, as well as to continue to compare it to speaker-based solutions and Mic-Through AR.

In order to support a moving user (or audio sample), Hear-Through AR using a speaker array or a head-tracked user wearing a BCH could be compared with a Mic-Through AR setup using headphones. The CG audio could be suitably manipulated to

account for the movement of the listener or objects in the scene [2, 14]. By using more-realistic sounds, we can gauge the applicability of the BCH for speech and non-speech audio, as well as compare how sound-elevation cues can be perceived using the various approaches to audio AR.

Another option is to use the same set of signals for all conditions, but pass the signal through a set of HRTFs for H, and BRTFs for B [15]. This approach would take advantage of the capabilities of modern sound cards, which support the application of HRTFs for generating spatial sounds.

Our approach to the study of audio AR is more applied than theoretical. By constructing systems that support various techniques, and then comparing them with sound captured from the real world, we hope to complement the more-fundamental work being done by others [16, 19, 13]. To this end, we are looking at novel ways of using and mixing environmental and CG sounds. For example, because BCH devices propagate sounds below 1kHz more predictably [16], one approach might be to use this channel for voice communication between individuals, and standard headphones for RW and non-speech-CG sound.

Another interesting study would be to combine a BCH with activated ear-bud headphones, and to compare different components of spatialized audio signals directed to each device separately. Using the results found by Raykar *et al*. [11], our next study is looking at how to pre-process CG audio in order to extract cues that relate to aspects of localization that might be best suited for delivery using a BCH.

To be successful, the subjective authenticity of voice signals delivered through BCH devices needs to be studied. One interesting study would be to display live and BCH spoken audio to blindfolded subjects, and ask them whether the voice is live or recorded. Informal tests have shown that these types of audio are almost indistinguishable, so we see promise in this line of study.

ACKNOWLEDGEMENTS

This research was supported in part by the National Institute of Information and Communications Technology of Japan.

REFERENCES

[1] Bederson, B. Audio Augmented Reality: A Prototype Automated Tour Guide, *Proc. of ACM CHI'95*, 1995, 210-211.

[2] Cook, P.R., Essl, G., Tzanetakis, G., Trueman, D. N >> 2: Multi-speaker Display Systems for Virtual Reality and Spatial Audio Projection, *Proc. of Int'l Conf. Auditory Display (ICAD)*, Glasgow, 1998.

[3] Fukumoto. M., Tonomura, Y. Whisper: A Wristwatch Style Wearable Handset, *Proc. of ACM CHI'99*, 112-119.

[4] Gardner, W.G. 3D Audio and Acoustic Environment Modeling, 1999. Retrieved May 29, 2007, from Harmony Central Web site: http://harmony-central.com/Computer/Programming/3d-audio.pdf

[5] Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Nironen, H., Vesa, S. Techniques and Applications of Wearable Augmented Reality Audio, *114th Conv. of Audio Eng. Soc.*, 2003.

[6] Jacobs, K., Nahmias, J., Angus, C., Reche, A., Loscos, C., and Steed, A. Automatic Generation of Consistent Shadows for Augmented Reality, *Proc. of the 2005 Conf. on Graphics interface*, 2005, 113-120.

[7] Lindeman, R.W., Noma, H. A Classification Scheme for Multi-Sensory Augmented Reality, *Proc. of ACM Virtual Reality Software and Technology (VRST) 2007*, 175-178.

[8] Lindeman, R.W., Noma, H., de Barros, P.G. Hear-Through and Mic-Through Augmented Reality: Using Bone Conduction to Display Spatialized Audio, *Proc. of Int'l Symposium on Mixed and Augmented Reality (ISMAR) 2007*.

[9] MacDonald, J.A., Henry, P.P., Letowski, T.R. Spatial Audio Through a Bone Conduction Interface, *Int'l J. of Audiology*, 2006, 45, 595-599.

[10] Niwa, M., Noma, H., Yanagida, Y., Hosaka, K., Lindeman, R.W. Controlling the Perceived Vibrational Frequency and Amplitude of a Voice-Coil-Type Tactor, *Proc. of the 14th Symp. on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, 2006, 55-56.

[11] Raykar, V.C., Duraiswami, R., Davis, L., Yegnanarayana, B. Exracting Significant Features from the HRTF, *Proc. of the 2003 Int'l Conf. on Auditory Display*, 115-118.

[12] Sawhney, N., Schmandt, C. Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments, *ACM Trans. On Computer-Human Interaction (TOCHI)*, 7(3), 2000, 353-383.

[13] Stanley, R.M., Walker, B.N. Lateralizaton of Sounds Using Bone-Conduction Headsets, to appear in *Proc. of the Annual Meeting of the Human Factors and Erg. Soc. (HFES2006)*.

[14] Takala, T., Hahn, J. Sound Rendering. *Proc. of SIGGRAPH '92*, 1992, 211-220.

[15] Walker, B.N., Lindsay, J. Navigation Performance in a Virtual Environment with Bonephones, *Proc. of the Int'l Conf. on Auditory Display (ICAD2005)*, 260-263.

[16] Walker, B.N., Stanley, R. Thresholds of Audibility for Bone-Conduction Headsets, *Proc. of the Int'l Conf. on Auditory Display (ICAD2005)*, 218-222.

[17] Wenzel, E.M., Wightman, F.L., Kistler, D.J. Localization with Non-Individualized Virtual Acoustic Display Cues, *Proc. of ACM CHI'91*, 351-359.

[18] Wenzel, E.M., Arruda, M., Kistler, D.J., Wightman, F.L. Localization Using Nonindividualized Head-related Transfer Functions, *J. Acoust. Soc. Am.*, 94(1), 1993, 111-123.

[19] Won, S.Y., Berger, J. Estimating Transfer Function from Air to Bone Conduction using Singing Voice, *Proc. of Int'l Computer Music Conf.*, 2005.