(Figure 1). The vibrational actuators are positioned on the zygomatic (cheek) bones in front of the ear, instead of on the mastoid bones behind the ear [9]. In addition, the unit does not require an external amplifier, and uses a standard stereo 3.5mm audio connector, allowing it to be plugged into most consumer audio devices. It 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.

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 airborne 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

The ability to which individuals can determine the position of a sound source relative to their current head position and orientation is based on several factors which vary between individuals, including the shape of the torso, head, and ears [6, 4, 10]. Most recent psychoacoustic work done on the delivery of spatialized sound has focused on using head-related transfer functions (HRTFs) [11, 9]. HRTFs incorporate the individual differences outlined above, so a given HRTF might not work well for all listeners. However, Wenzel *et al*. showed that using an HRTF from a subject who exhibited a strong ability to localize audio allowed a large number of other subjects to properly identify spatialized audio sources, suggesting the utility of carefully chosen, non-individualized HRTFs [10]. 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) [6]. One open question is how well HRTFs can be applied to audio delivered using bone conduction.

Early bone-conduction 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. 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 [9]. These would allow audio designers to provide spatialized cues using bone conduction.

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. We now describe several ways that sound can travel from source to listener in AR environments.

4.1. Steps in the AR-Sound-Delivery Process

One of the main differentiators between the three types of audio AR studied here has to do with how the audio signals are altered on their way to the listener. 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 [5], 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 [10], but other effects, such as sound dampening and reflection from objects or structures in the environment, can also greatly affect sound believability [8].

Each of the three audio AR approaches we are studying starts with real-world (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 2a) and BCH-based techniques (Figure 2b), 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.

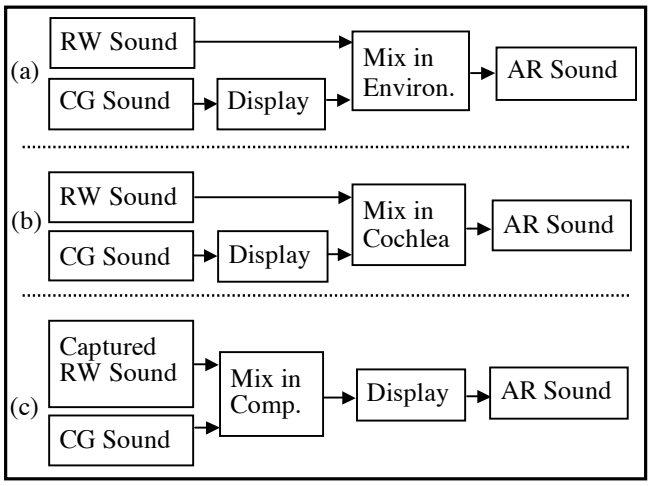

Figure 2: Path of sounds for (a) speaker-based, Hear-Through AR, (b) bone-conduction-headset-based, Hear-Through AR, and (c) headphone-based, Mic-Through AR

For the case where a set of speakers in the environment is used to deliver CG sounds (Figure 2a), 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 2b), 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.

For the Mic-Through approach (Figure 2c), 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 3). 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.

Figure 3: 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 [8]. 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.

5 Empirical User Study

We performed a user study to compare how well people can perceive spatialized audio using bone conduction versus traditional headphones or a speaker array. The study involved both stationary and moving sound cues of varying frequencies. This goal of the user study is to develop some baseline data using simple audio cues and a stationary listener. Later studies can then explore the use of Hear-Through and Mic-Through AR in more-realistic contexts with more complex audio.

5.1. Method

Twenty-four computer science students, 22 male and 2 female, ranging in age from 20 to 30 years, participated in the study. 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 1,000Hz (HIGH). Stimulus motion had two levels: Stationary (STAT) and Moving (MOV).

After signing an Institutional-Review-Board-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 4). 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.

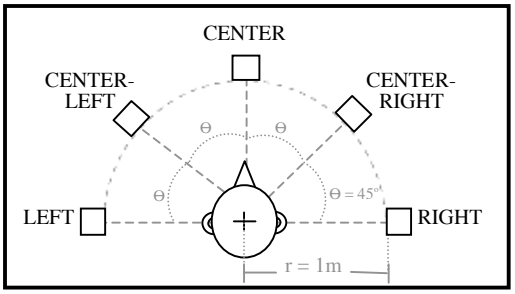

Figure 4: Reference tone locations. Each reference tone was either played through the corresponding speaker (S), or was captured from the speaker prior to the study, and replayed during the study (B & H).

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. 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.

5.2. Results

A 95% (α=0.05) confidence level was used to determine statistical significance. A summary of the statistically significant results can be found in Table 1. We can see a noticeable difference between subject accuracy with regard to STAT versus MOV in terms of audio device.

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

Table 1: 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)

An analysis of variance (ANOVA) of the mean accuracy values on the main effects for **STAT** showed 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 [$F(4,1071)=0.66$, $p=0.618$]. In addition, 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, subjects showed significantly better accuracy for LOW and HIGH, 75% (34%), than for MED, 68% (36%).

An ANOVA of the data on the main effects for **MOV** showed 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, 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%), or between B and H.

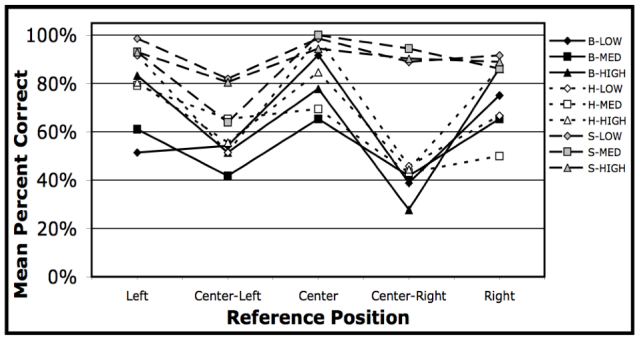

Figure 5: Graph of mean percent correct by audio device, frequency, and reference position

Figure 5 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 use squares, and HIGH use triangles.

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 great promise in the use of bone conduction. 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 [10]. 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 as well as situation awareness.

6 Conclusions and Future Work

Our results are encouraging, though much follow-on work is necessary. 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, 8]. 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.

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, pp. 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*, pp. 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] 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, pp. 113-120.

[6] 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.

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

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

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

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

[11] 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, pp. 111-123.