HEAD TRACKED 3-D AUDIO USING LOUDSPEAKERS

William G. Gardner

Machine Listening Group
MIT Media Lab, Rm. E15-401B
20 Ames St., Cambridge, MA 02139
billg@media.mit.edu

ABSTRACT

Existing loudspeaker 3-D audio systems suffer from a fixed listening location. This paper proposes using a head tracker to steer the equalization zone to the position of the tracked listener. Sound localization experiments show that this strategy greatly improves localization when the listener is displaced from the ideal listening location, and also enables dynamic localization cues.

1. INTRODUCTION

This paper describes a new approach to implementing a virtual acoustic display using a pair of conventional loudspeakers. The new idea is to use the tracked position of the listener's head to optimize the acoustical presentation, and thus produce a much more realistic illusion over a larger listening area than existing loudspeaker 3-D audio systems. By using a remote head tracker, for instance based on computer vision, an immersive audio environment can be created without donning headphones or other equipment.

The general approach to a 3-D audio system is to reconstruct the acoustic pressures at the listener's ears that would result from the natural listening situation to be simulated. To accomplish this using loudspeakers requires that 1) the ear signals corresponding to the target scene are synthesized by appropriately encoding directional cues, a process known as "binaural synthesis," and 2) these signals are delivered to the listener by inverting the transmission paths that exist from the speakers to the listener, a process known as "crosstalk cancellation."

Headphones are usually used for 3-D audio because they have excellent channel separation, they isolate the listener from external sounds and room reverberation, and the transmission paths from the transducers to the ears are easily inverted. Loudspeaker systems have none of these attributes. However, loudspeaker displays do have certain advantages over headphones. When synthesizing directional cues using non-individualized head-related transfer functions (HRTFs), headphones have difficulty reproducing externalized frontal images [1], whereas frontally placed loudspeakers do this naturally. Loudspeakers also have the advantage of not requiring the listener to don headphones.

There are a number of exisiting systems that render 3-D audio using loudspeakers by combining a binaural synthesizer (or a recorded source of binaural audio) with a symmetric crosstalk canceller [2-4]. However, their use is severely limited because the crosstalk cancellation only functions for a fixed listener position, called the "sweet spot" or "equalization zone." Consequently, when the listener moves away from the equalization zone, the 3-D illusion is lost. This paper describes crosstalk cancellers that can be steered to the location of a tracked listener. Steering the equalization zone to the listener preserves the 3-D illusion over a large listening volume, thus simulating a reconstructed soundfield, and also provides dynamic localization cues by maintaining stationary external sound sources during head motion.

We have constructed a head tracked loudspeaker 3-D audio system and have conducted sound localization experiments to test its performance under steered and unsteered conditions for both fixed and moving heads. With fixed heads, significant differences between steered and unsteered conditions are seen when the listener's head is displaced or rotated with respect to the ideal listening location. Experiments conducted with moving (rotating) heads show that improvement in the resolution of front-back confusions occurs when steering is enabled.

2. APPROACH

2.1. Theory of crosstalk cancellation

Binaural synthesis is accomplished by convolving an input signal with a pair of HRTFs:

$$\mathbf{x} = \mathbf{h}x \qquad \text{(EQ 1)}$$

Here, x is the input signal, $\mathbf{x} = \left[x_L\ x_R\right]^T$ is a column vector of binaural signals, and $\mathbf{h} = \left[H_L\ H_R\right]^T$ is a column vector of synthesis HRTFs, with the frequency variable omitted for brevity. This is a general specification of the binaural synthesis procedure. In order to deliver the binaural signals over loudspeakers, it is necessary to filter them with a 2x2 matrix $\mathbf{C}$ of transfer functions:

$$\mathbf{y} = \mathbf{C}\mathbf{x} \qquad \text{(EQ 2)}$$

where $\mathbf{y} = \left[y_L\ y_R\right]^T$ is the vector of loudspeaker signals, and $\mathbf{C}$ is the crosstalk canceller. The ear signals are related to the speaker signals by:

$$\mathbf{e} = \mathbf{A}\mathbf{y} \qquad \text{(EQ 3)}$$

where $\mathbf{e} = \begin{bmatrix} e_L & e_R \end{bmatrix}^T$ is the vector of ear signals, and **A** is a 2x2 matrix describing the electroacoustic transfer between the loudspeakers and the ears. **A** can be factored as follows:

$$\mathbf{A} = \mathbf{HS}$$

$$\mathbf{H} = \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix}, \mathbf{S} = \begin{bmatrix} S_L A_L & 0 \\ 0 & S_R A_R \end{bmatrix} \quad \text{(EQ 4)}$$

H is a 2x2 matrix of HRTFs measured relative to the free-field response at the center of the head with no head present, where H_{XY} is the response from speaker X to ear Y. **S** is a diagonal matrix that accounts for the frequency response of the speakers and the air propagation, where S_X is the response of speaker X, and A_X is the response of the air propagation from speaker X to the center of the head with no head present.

In order to deliver the binaural signals to the ears, we choose C to be the inverse of A:

$$\mathbf{C} = \mathbf{A}^{-1} = \mathbf{S}^{-1}\mathbf{H}^{-1}$$

$$= \begin{bmatrix} S_L^{-1}A_L^{-1} & 0 \\ 0 & S_R^{-1}A_R^{-1} \end{bmatrix} \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix}^{-1} \quad \text{(EQ 5)}$$

In practice, the speaker inverse filters are usually ignored. The air propagation inverse filters compensate for unequal path lengths from the speakers to the listener, and are critical for off-axis listening positions. The heart of the crosstalk canceller is the inverse head transfer matrix, which can be written [5]:

$$\mathbf{H}^{-1} = \begin{bmatrix} H_{LL}^{-1} & 0 \\ 0 & H_{RR}^{-1} \end{bmatrix} \begin{bmatrix} 1 & -ITF_R \\ -ITF_L & 1 \end{bmatrix} \frac{1}{1 - ITF_L ITF_R} \quad \text{(EQ 6)}$$

where:

$$ITF_L = \frac{H_{LR}}{H_{LL}}, ITF_R = \frac{H_{RL}}{H_{RR}} \quad \text{(EQ 7)}$$

are the interaural transfer functions (ITFs) that describe head shadowing. Crosstalk cancellation is effected by the *-ITF* terms in the off-diagonal positions of the righthand matrix. These terms predict the crosstalk and send an out-of-phase cancellation signal into the opposite channel. The scalar term compensates for higher order crosstalks. The lefthand diagonal matrix, which we call "ipsilateral equalization," associates an ipsilateral inverse filter with each speaker output.

2.2. Frequency-dependent implementation

Our approach to the transmission path inversion problem is to use different strategies for low and high frequencies. This is justified because the large intersubject variation in high frequency head response means that it is impossible to invert the head response using a non-individualized head model. Furthermore, inversion at high frequencies becomes critically sensitive to positional errors, and when the cancellation fails due to either a mismatch in head response or a position error, interaural cues may be seriously degraded.

Therefore, crosstalk cancellation is only attempted at low frequencies [6], where intersubject variation in head response is small. Because of the dominance of low frequency interaural time delay (ITD) cues for localization [7], this strategy is very effective at synthesizing extreme lateral sources. Bandlimited crosstalk cancellation is achieved by associating a lowpass filter with the off-diagonal terms of the head transfer matrix of Eq. 4 prior to inversion. This is equivalent to associating a lowpass filter with each ITF term in Eq. 6.

At high frequencies, the bandlimited crosstalk canceller will implement the diagonal ipsilateral equalization matrix (Eq. 6), and thus the speakers will emit free-field equalized binaural signals. We modify this situation by associating scaling gains with each channel. The gains are determined by solving a power transfer analog to Eq. 5 in an attempt to deliver the proper high frequency powers to each ear [8]. When an exact solution is not possible, the proper total high frequency power is delivered without regard to interaural cues. This strategy is implemented by applying a high frequency shelving filter to each channel of each binaural source, where the shelving gains are dependent on the location of the source and the orientation of the listener. When the listener is facing forward, this high frequency model provides a modest improvement in the high frequency interaural level difference (ILD) cue. When the listener's head is rotated, the model helps to align the high frequency ILD cue with the low frequency ITD cue.

We have implemented both symmetric and asymmetric bandlimited crosstalk cancellers [8]. The symmetric implementation, where $H_{LL} = H_{RR}$ and $H_{LR} = H_{RL}$, is based on a shuffler filter topology [4] and uses infinite impulse response (IIR) filters to model head shadowing. The equalization zone can be steered front-back by changing the ITD parameter of the head shadowing model, and steered left-right by suitably delaying and attenuating an output channel. The symmetric implementation can not compensate for head rotations and does not implement the high frequency shelving filters described above.

The general asymmetric implementation allows dynamic head motion and rotation. Although efficient recursive implementations for the asymmetric crosstalk canceller exist [8], we used a very simple approach. For each head orientation, speaker spread angle, and source location, we computed a pair of 128-point finite impulse response (FIR) filters containing the synthesis HRTFs, crosstalk cancellation filters, and high frequency shelving filters. Filters were computed for horizontal sources, at head rotations from 0 to ±45 degrees, at speaker spread angles from 55 to 70 degrees, all in 5 degree increments, for a total of 2880 filter pairs. The air propagation inverse filters are implemented by delaying and attenuating the speaker closer to the listener. Head tracking is accomplished using a Polhemus device which must worn on the head, though we ultimately plan to use a visual head tracker [9]. The tracked head position and source position are used to index a pair of synthesis filters and a simple interpolation scheme prevents clicks during updates.

Both implementations use non-individualized HRTFs measured from a KEMAR [10] for both binaural synthesis and crosstalk cancellation. The crossover frequency between low and high frequency regions is 6 kHz., and the sampling rate is 32 kHz.

Table 1: RMS errors between ideal responses and mean lateral angle judgements for various experiments are given in first row. Front-back reversal percentages are given in lower rows. Pairs of values are for (unsteered, steered) conditions.

	baseline	25 cm. rear	20 cm. right	40°left	head rotating
RMS error	10°	(14°, 10°)	(27°, 18°)	(47°, 26°)	(13°, 11°)
front->back	28%	(23%, 34%)	(37%, 34%)	(28%, 30%)	(4%, 2%)
back->front	36%	(60%, 39%)	(57%, 39%)	(48%, 50%)	(100%, 53%)

3. LOCALIZATION EXPERIMENTS

A number of sound localization experiments have been conducted using these systems to test the idea of steering the equalization zone to the listener [8]. The experiments were conducted in a small studio with a 500 Hz. reverberation time of 0.23 sec, using speakers placed at ±30 degrees relative to the ideal listening location, at a distance of 30 in. (76 cm.) from head center. The experiments were conducted sequentially as various features were added to the audio system. In order, the experiments were:

1. listener in ideal position, fixed head.
2. listener displaced to the front 16 cm. and to the rear 25 cm., fixed head.
3. listener displaced to the right 10 cm. and 20 cm., fixed head.
4. listener's head rotated 20 degrees left and 40 degrees left, fixed head.
5. listener's head rotating right during stimulus presentation.

Experiment 1 is a baseline localization experiment, and the others compare localization under steered and unsteered conditions. Experiments 1-3 used the symmetric crosstalk canceller, and 4-5 used the asymmetric crosstalk canceller. Head position was tightly controlled using a sighting apparatus for experiments 1-4, and experiment 5 used the head tracker.

Eight paid adult volunteers served as subjects (6 male, 2 female). All subjects reported normal hearing, and none had prior experience with virtual acoustic displays. The stimulus for experiments 1-4 was a set of five pink noise bursts, 250 msec in duration with 10 msec linear onset and offset ramps, with 500 msec gaps between bursts. Experiment 5 used a single pink noise burst. Some of the experiments also contained trials where the pink noise was lowpass filtered at 6 kHz. The stimulus was processed by the 3-D system and presented to the subject over the loudspeakers at a level of approximately 64 dBA SPL. Spectral randomization across trials was not done.

Subjects were instructed to report the perceived location of the sound. This was done by verbally reporting the azimuth and elevation angles and distance from the head. This process was aided by descriptive charts placed in front of the subject. Distance responses were given in a body-centered coordinate system (in head, on head, shoulder, elbow, arm's length, beyond reach). The set of target locations and the responses that were recorded varied per experiment. For all experiments, each combination of stimulus, conditions, and target location was only presented once to each subject, in random order.

3.1. Results

We will focus our attention on the efficacy of steering the equalization zone to a displaced or rotated listener. This was tested in experiments 2-4 using horizontal target locations at 30 degree increments under steered and unsteered conditions. The unsteered condition is equivalent to steering the equalization zone to the ideal listening location. Results are analyzed in terms of the judged lateral angle, i.e. the angle between the judged azimuth and the median plane. Table 1 lists the root mean square (RMS) error in degrees between the ideal responses and the mean lateral judgements. The judged lateral angle is unaffected by front-back reversals, which are counted separately [1]. Table 1 lists the percentage of total judgements considered to be front-back reversals. Not shown is that the pattern of front-back reversals varies considerably per individual.

Figure 1a shows the results for experiment 2, when the head was displaced 25 cm. to the rear. For each target azimuth, the lateral judgement averaged across all subjects is plotted. The solid line is the mean judgement for the steered condition (including ±1 standard deviation errorbars) and the dashed line is for the unsteered condition. The circled points indicate target azimuths that yielded significant differences under the two conditions, according to a two-tailed matched pairs t test at a 5% significance level. The arrowheads indicate the location of the speakers relative to the subject.

Because a front-back displacement doesn't change the relative distance to the speakers, we wouldn't expect the low frequency crosstalk cancellation to be affected much by small displacements. In fact the results for steered and unsteered conditions are similar, and there are only significant differences for targets near the median plane.

Figure 1b shows the results for experiment 3, when the head was displaced right by 20 cm. We now expect the crosstalk cancellation to be seriously degraded by this displacement, and we see that the unsteered results look like a conventional stereo pan between the two speaker positions [3]. With steering enabled, extreme lateral sources are more easily synthesized. However, the maximum lateral angle is smaller on the right hand side and there is more variance in the responses, probably due to the frontal location of the right hand speaker. The frontal location of the right speaker decreases the high frequency ILD cue that can be generated for right hand targets.

Figure 1c shows the results for experiment 4, when the head was rotated 40 degrees to the left. Note that both the target locations and the responses are relative to the rotated head orientation. With steering disabled, the responses are mostly confined to the right hemisphere because both speakers are on the right hand side of the head. When steering is enabled, it is possible to generate left sounding images,but these images are not localized very far to the left and there is a great deal of variation in the responses. It is not possible to synthesize left side images containing only high frequencies, because the crosstalk cancellation is limited to low frequencies. The

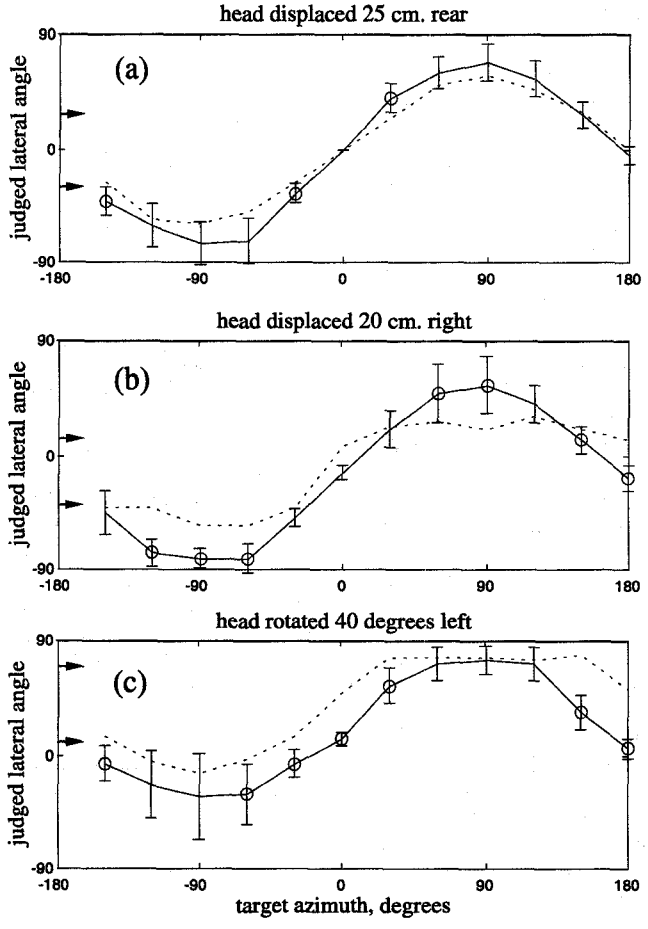

Figure 1: Comparison of steered (solid line) and unsteered (dashed line) results for various localization experiments.

presence of low frequencies in the stimulus allows the low frequency ITD cue to dominate the conflicting high frequency ILD cue [7]. Experiments conducted using 6 kHz lowpass filtered pink noise bursts showed much less variation in the responses for left side targets, but the responses did not become more lateral [8].

Experiment 5, which test localization during dynamic head motion, differs from the other experiments. The subject was seated in the ideal listening location, the head tracker was donned and then calibrated using the sighting apparatus. Subjects were instructed to turn their head to face the left speaker and to rotate their head to face the right loudspeaker. This motion triggered the stimulus presentation. The triggering was designed so that the 250 msec stimulus would play while the listener was subtending -10 to +10 degrees azimuth, although a range of angular velocities was allowed to avoid frustrating the subjects. After stimulus presentation, the subject reported the azimuth judgement relative to frontal orientation. When steering was enabled, head motion updated both the crosstalk cancellation and the binaural synthesis to maintain a stationary external source. The total latency from head motion to audible change was 75 msec maximum. When steering was disabled, head motion produced no change in the audio processing.

Table 1 gives preliminary results with 5 subjects tested. With steering disabled, almost all targets are heard as frontal, with 100% of rear targets reversed to the front. Enabling steering reduced the back-to-front reversals to 53%, while the front-to-back reversals stayed negligibly small. The pattern of reversals is very subject specific: one subject heard all sources in front regardless of steering; two subjects heard all steered sources without reversals, having heard all unsteered sources as frontal; the other subjects had mixed results.

4. CONCLUSIONS

We have presented a loudspeaker 3-D audio system that steers the equalization zone to a listener whose position is tracked by some means. The steerable crosstalk canceller uses different strategies for low and high frequencies. Results from sound localization experiments show that steering the equalization zone greatly improves horizontal localization performance when the listener's head is laterally displaced or rotated with respect to the ideal position. Tracking the head also enables dynamic localization cues that are useful for resolving front-back reversals. It is difficult to synthesize consistent images on one side of the head when both loudspeakers are on the opposite side, due to the problem of inverting the high frequency transmission paths. This would suggest the use of wider spaced loudspeakers, or additional rear loudspeakers, but these solutions may not be practical for desktop computer interfaces.

REFERENCES

1. Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, 94(1), 111-123.

2. Schroeder, M. R., and Atal, B. S. (1963). "Computer simulation of sound transmission in rooms," *IEEE Int. Conv. Record*, 7, 150-155.

3. Damaske, P. (1971). "Head-related Two-channel Stereophony with Loudspeaker Reproduction," *J. Acoust. Soc. Am.*, 50(4), 1109-1115.

4. Cooper, D. H., and Bauck, J. L. (1989). "Prospects for Transaural Recording," *J. Audio Eng. Soc.*, 37(1/2), 3-19.

5. Moller, H. (1992). "Fundamentals of Binaural Technology," *Applied Acoustics*, 36, 171-218.

6. Cooper, D. H., and Bauck, J. L. (1990). "Head Diffraction Compensated Stereo System," U.S. Patent no. 4,893,342.

7. Wightman, F., and Kistler, D. (1992). "The dominant role of low-frequency interaural time differences in sound localization", *J. Acoust. Soc. Am.*, 91(3), 1648-1661.

8. Gardner, W. G. (1997). "3-D Audio Using Loudspeakers," Ph.D. Dissertation, MIT Media Lab, Cambridge, MA.

9. Oliver, N., Pentland, A., and Berard, F. (1997). "LAFTER: Lips and Face Real Time Tracker," *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition.*

10. Gardner, W. G., and Martin, K. D. (1995). "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, 97(6), 3907-3908.