Vision-Steered Beam Forming and Transaural Rendering for the Artificial Life Interactive Video Environment, (ALIVE)

Perceptual Computing Technical Report #352

Michael A. Casey, William G. Gardner, Sumit Basu
MIT Media Laboratory, Cambridge, USA
mkc,billg,sbasu@media.mit.edu

Abstract

This paper describes the audio component of an interactive video system that uses remote sensing to free the user from body-mounted tracking equipment. Position information is obtained from a camera and used to constrain a beam-forming microphone array, for far-field speech input, and a two-speaker transaural audio system for rendering 3D audio.

1 Vision-Steered Beam Forming

1.1 Introduction

The Media Lab's ALIVE project, *Artificial Life Interactive Video Environment*, is a testbed for research into remote-sensing, full-body, interactive interfaces for virtual environments, [4]. A plan view of the ALIVE space is shown in Figure 1. A camera on top of the large video projection screen captures images of the user in the active zone, these images are fed to a visual recognition system where various features are tracked. A mirror image of the user is projected onto the video screen at the front of the space and computer-rendered scenes are overlayed onto the users image to produce a combined real/artificial composite image. One of the applications in the ALIVE space is an autonomous agent model of a dog, called "Silas", which posseses its own system of behaviors and motor-control schemes. The user is able to give the dog visual commands by gesturing. The visual recognition system makes decisions on which command has been issued out of the twelve gestures that it can recognize, see Figure 2. This section describes the design and implementation of an audio component for the ALIVE system that allows the user to give the artificial dog spoken commands from the free field without the use of body-mounted microphones.

1.2 Free-field speech recognition

Speech recognition applications typically require near-field, i.e. $< 1.5m$, microphone placement for acceptable performance. Beyond this distance the signal to noise ratio of the incoming speech affects the performace significantly. Commercial speech-recognition packages typically break down over a $4 - 6dB$ range.

The ALIVE space requires the user to be free of the constraints of near-field microphone placement and the user must be able to move around the active zone of the space with no noticable degradation in performance.

As a result there are several potential solutions. One of these is to have a highly directional microphone that can be panned using a motorized control unit, to track the user's location. This requires a significant amount of mounting and control hardware, and is limited by the speed and accuracy of the drive motors. In addition, it can only track one user at a time. It is preferable to have a directional response that can be steered electronically. This can be done with the well-known technique of beamforming with an array of microphone elements. Though several microphones need to be used for this method, they need not be very directional and they can be permanently mounted in the environment. In addition, the signals from the microphones in the array can be combined in as many ways as the available computational power is capable of, allowing for the tracking of multiple moving sound sources from a single microphone array.

1.3 Adaptive Beamforming

Adaptive beamforming strategies account for movement in source direction by continuously updating the spatial characteristics of the array for an optimal signal to noise ratio. The behavior of the array is thus determined by the nature of the source signal(s). The problem with adaptive strategies for the ALIVE space is that there are typically many observers around the space and the level of ambient speech-like sound is very high as a result. Adaptive algorithms do not perform well when multiple sources arrive simultaneously from different directions [2].

1.4 Fixed Beamforming

Fixed array strategies optimize the microphone filtering for a particular direction and don't change with varying incident source direction. Thus the directional response of the array is fixed to a particular azimuth and elevation. However, if the target source is non-stationary, the signal enhancement performance is reduced as the source moves away from the steering direction. Spatial beamwidth constraints may be added to the fixed array such that the directionality of the response is traded for beam width to compensate for small movements in the source. As the beam width increases, so too does the level of ambient noise pickup. The active user zone in the ALIVE space allows movement over a large azimuth range thus fixed array formulations need to be modified in real-time in order to beamform in the direction of the user.

1.5 A Visually Constrained Beamformer

The ALIVE space utilizes a visual recognition system called *Pfinder*, short for *person finder*, developed at the Media Lab for tracking a person's hands, face or any other color-discriminable feature [4]. Pfinder uses an intensity-normalized color representation of each pixel in the camera image and multi-way Gaussian classifiers to decide which of several classes each pixel belongs to. Examples of classes are background, left hand, right hand and head. The background class can be made up of arbitrary scenery as long as the color value of each pixel does not vary beyond the decision boundary for inclusion in another class. The mean of each cluster gives the coordinates of the class, and the eigenvalues give the orientation. Pfinder provides updates on each class roughly 6 times a second. Further details on the visual recognition system can be found in [4].

The information from the visual recognition system is used to steer a fixed beamforming algorithm. Azimuth calculations are performed from the 3-space coordinate data provided by the mean of the head class. The re-calculation of weights for each new azimuth is a relatively low-cost operation since the weight update rate is 5 Hz.

The use of visual recognition techniques makes it possible to achieve both the optimal signal-enhancement performance of a fixed beamformer with narrow beam width and to get the spatial flexibility of an adaptive beamformer.

1.6 Fixed Beamformer Algorithms

In this section we describe a fixed beamformer algorithm and the different microphone arrangements that can be used with it. The geometry of the microphone array is represented by the set of vectors $\mathbf{r}_n$ which describe the position of each microphone n relative to some reference point (e.g., the center of the array), see Figure 3. The array is steered to maximize the response to plane waves coming from the direction $\mathbf{r}_s$ of frequency f_o. Then, for a plane wave incident from the direction $\hat{\mathbf{r}}_i$, at angle θ, the gain is:

$$G(\theta) = \begin{bmatrix} a_o & a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} F(\theta)e^{jk\mathbf{r}_o \cdot \hat{\mathbf{r}}_i} \\ F(\theta)e^{jk\mathbf{r}_1 \cdot \hat{\mathbf{r}}_i} \\ F(\theta)e^{jk\mathbf{r}_2 \cdot \hat{\mathbf{r}}_i} \\ F(\theta)e^{jk\mathbf{r}_3 \cdot \hat{\mathbf{r}}_i} \end{bmatrix} \tag{1}$$

where $a_n = |a_n| e^{-jk_o \hat{\mathbf{r}}_n \cdot \hat{\mathbf{r}}_s}$, $F(\theta)$ is the gain pattern of each individual microphone, and k $(2\pi f/c)$ is the wave number of the incident plane wave. k_o is the wave number corresponding to the frequency f_o of the incident plane wave.

Note that there is also a ϕ dependence for F and G, but since we are only interested in steering in one dimension, we have omitted this factor. This expression can be written more compactly as:

$$G(\theta) = \mathbf{W}^T \mathbf{H} \tag{2}$$

where $\mathbf{W}$ represents the microphone weights and $\mathbf{H}$ is the set of transfer functions between each microphone and the reference point. In the formulation above, a maxima is created in the gain pattern at the steering angle for the expected frequency, since $\hat{\mathbf{r}}_i = \hat{\mathbf{r}}_s$ and the phase terms in $\mathbf{W}$ and $\mathbf{H}$ cancel each other. Note, however, that this is not the only set of weights that can be used for $\mathbf{W}$. For example, Stadler and Rabinowitz present a method of obtaining the weights with a parameter β that arbitrates high directivity and uncorrelated noise gain [7]. This method, when used to obtain maximum directivity, yields gain patterns that are slightly more directional than the basic weights described above.

The standard performance metric for the directionality of a fixed array is the *directivity index* which is shown in Equation 3, [7]. The directivity index is the ratio of the array output power due to sound arriving from the far field in the target direction, (ϕ_0, θ_0), to the output power due to sound arriving from all other directions in a spherically isotropic noise field,

$$D = \frac{|G(\phi_0, \theta_0)|^2}{(1/4\pi)\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} |G(\phi, \theta)|^2 \sin\theta d\phi d\theta}, \tag{3}$$

The directivity index thus formulated is a narrow-band performance metric; it is dependent on frequency but the frequency terms are omitted from Equation 3 for simplicity of notation. In order to assess an array for use in speech enhancement a broad-band performance metric must be used.

One such metric is the *intelligibility-weighted directivity index* [7] in which the directivity index is weighted by a set of frequency-dependent coefficients provided by the ANSI standard for the speech articulation index [1]. This metric weights the directivity index in fourteen one-third-octave bands spanning 180 to 4500 Hz [7].

1.7 Designing the Array

An important first consideration is the choice of array geometry. Two possible architectures are considered; endfire arrangement, Figure 3, and broadside arrangement, Figure 4. A second factor is the choice of microphone gain pattern for the individual microphone elements, $F(\theta)$. Since the gain pattern $F(\theta)$ can be pulled out of the $\mathbf{H}$ vector as a constant multiplier, the gain pattern for the array can be viewed as the product of the microphone gain pattern and an omnidirectional response where $F(\theta) = 1$. This is the well-known principle of *pattern multiplication* [3] [7]. For omnidirectional microphones, the gain patterns for the two layouts are identical but for a rotation. The gain patterns for an endfire array centered along the $\theta = 0$ axis with omnidirectional microphones steered at 15, 45, and 75 degrees is shown in Figure 5. The use of a cardioid response greatly reduces the large lobes appearing in the rear of the array. The corresponding responses for cardioid microphones are shown in Figure 6.

Cardioid elements for the broadside array are not as useful, since the null of the cardioid does not eliminate as much of the undesirable lobes of the gain pattern, Figure 7. Note the symmetry of the response about the $\theta = 0$ axis; the line containing the microphone array. This symmetry can be eliminated by nulling out one half of the array response using an acoustic reflector or baffle along one side of the microphone array. The reflector will effectively double one side of the gain pattern and eliminate the other, while the baffle will eliminate one side and not affect the other. Thus a good directional response can be achieved between 0 and 90 degrees using both cardioid elements and a baffle for the endfire configuration. The incorporation of a second array, on the other side of the baffle, gives the angles zero to -90 degrees. A layout of the ALIVE space with such an array/baffle combination is shown in Figure 8.

The response of each of the above arrays as measured by the directivity index of Equation 3 is given in Table 1. The integration was performed for a fixed ϕ across all θ.

Table 1: Directivity Index for Different Array Architectures

Array Architecture	15 degrees	45 degrees	75 degrees
Omni Broadside	0.0029	0.0022	0.0022
Omni Endfire	0.0022	0.0022	0.0029
Cardioid Broadside	0.0057	0.0038	0.0045
Cardioid Endfire	0.0036	0.0037	0.0045

1.8 Conclusion

In this section we have described a vision-steered microphone array for use in a full-body interaction virtual environment without body-mounted sensing equipment. A preliminary implementation for the ALIVE space has shown that the system works well for constrained grammars of 10-20 commands. The advantages of cross-modal integration of sensory input are paramount in this system since the desirable properties of fixed arrays are combined with the steerability of an adaptive system.

2 Visually Steered 3-D Audio

2.1 Introduction

This section discusses the 3-D audio system that has been developed for the ALIVE project at the Media Lab [4]. The audio system can position sounds at arbitrary azimuths and elevations around a listener's head. The system uses stereo loudspeakers arranged conventionally (at ± 30 degrees with respect to the listener). The system works by reconstructing the acoustic pressures at the listener's ears that would occur with a free-field sound source at the desired location. This is accomplished by combining a *binaural spatializer* with a *transaural audio system*. The spatializer convolves the source sound with the direction dependent filters that simulate the transmission of sound from free-field to the two ears. The resulting binaural output of the spatializer is suitable for listening over headphones. In order to present the audio over loudspeakers, the output of the spatializer is fed to a transaural audio system which delivers binaural signals to the ears using stereo speakers. The transaural system filters the binaural signals so that the crosstalk leakage from each speaker to the opposite ear is canceled.

Transaural technology is applicable to situations where a single listener is in a reasonably constrained position facing stereo speakers. If the listener moves away from the ideal listening position, the crosstalk cancellation no longer functions, and the 3-D audio illusion vanishes. As discussed earlier, the ALIVE system uses video cameras to track the position of the listener's head and hands. A goal of this work is to investigate whether the tracking information can be used to dynamically adapt the transaural 3-D audio system so that the 3-D audio illusion is maintained as the listener moves.

We will first briefly review the principles behind binaural spatial synthesis and transaural audio. Then we will discuss the 3-D audio system that has been constructed for the ALIVE project. Finally we will discuss how the head tracking information can be used, and give preliminary results.

2.2 Principles of binaural spatial synthesis

A binaural spatializer simulates the auditory experience of one or more sound sources arbitrarily located around a listener [9]. The basic idea is to reproduce the acoustical signals at the two ears that would occur in a normal listening situation. This is accomplished by convolving each source signal with the pair of head-related transfer functions (HRTFs)[1] that correspond to the direction of the source, and the resulting binaural signal is presented to the listener over headphones. Usually, the HRTFs are equalized to compensate for the headphone

[1] The time domain equivalent of an HRTF is called a head-related impulse response (HRIR) and is obtained via the inverse Fourier transform of an HRTF. In this paper, we will use the term HRTF to refer to both the time and frequency domain representation.

to ear frequency response [26, 17]. A schematic diagram of a single source system is shown in figure 9. The direction of the source (θ = azimuth, ϕ = elevation) determines which pair of HRTFs to use, and the distance (r) determines the gain. Figure 10 shows a multiple source spatializer that adds a constant level of reverberation to enhance distance perception.

The simplest implementation of a binaural spatializer uses the measured HRTFs directly as finite impulse response (FIR) filters. Because the head response persists for several milliseconds, HRTFs can be more than 100 samples long at typical audio sampling rates. The interaural delay can be included in the filter responses directly as leading zero coefficients, or can be factored out in an effort to shorten the filter lengths. It is also possible to use mimimum phase filters derived from the HRTFs [15], since these will in general be shorter than the original HRTFs. This is somewhat risky because the resulting interaural phase may be completely distorted. It would appear, however, that interaural amplitudes as a function of frequency encode more useful directional information than interaural phase [16].

There are several problems common to headphone spatializers:

- The HRTFs used for sythesis are often a generic set and not the specific HRTFs of the listener. This can cause localization performance to suffer [24, 25], particularly in regards to front-back discrimination, elevation perception, and externalization. When the listener's own head responses are used, their localization performance is comparable to natural listening [26].

- The auditory scene created moves with the head. This can be fixed by dynamically tracking the orientation of the head and updating the HRTFs appropriately. Localization performance and realism should both improve when dynamic cues are added [24].

- The auditory images created are not perceived as being external to the head, but rather are localized at the head or inside the head. Externalization can be improved by using the listener's own head responses, adding reverberation, and adding dynamic cues [12].

- Frontal sounds are localized between the ears or on top ot the head, rather than in front of the listener. Because we are used to seeing sound sources that are in front of the head, it is difficult to convince the perceptual system that a sound is coming from the front without a corresponding visual cue. However, when using the listener's own HRTF's, frontal imaging with headphones can be excellent.

2.3 Implementation of binaural spatializer

Our implementation of the binaural spatializer is quite straightforward. The HRTFs were measured using a KEMAR (Knowles Electronics Mannequin for

Acoustics Research), which is a high quality dummy-head microphone. The HRTFs were measured in 10 degree elevation increments from -40 to +90 degrees [13]. In the horizontal plane (0 degrees elevation), measurements were made every 5 degrees of azimuth. In total, 710 directions were measured. The sampling density was chosen to be roughly in accordance with the localization resolution of humans. The HRTFs were measured at a 44.1 kHz sampling rate.

The raw HRTF measurements contained not only the desired acoustical response of the dummy head, but also the response of the measurement system, including the speaker, microphones, and associated electronics. In addition, the measured HRTFs contained the response of the KEMAR ear canals. This is undesirable, because the final presentation of spatialized audio to a listener will involve the listener's own ear canals, and thus a double ear canal resonance will be heard. One way to eliminate all factors which do not vary as a function of direction is to equalize the HRTFs to a diffuse-field reference [15]. This is accomplished by first forming the diffuse-field average of all the HRTFs:

$$|H_{DF}|^2 = \frac{1}{N} \sum_{i,k} |H_{\theta_i,\phi_k}|^2 \tag{4}$$

where H_{θ_i,ϕ_k} is the measured HRTF for azimuth θ_i and elevation ϕ_k. $|H_{DF}|^2$ is therefore the power spectrum which would result from a spatially diffuse soundfield of white noise excitation. This formulation assumes uniform spatial sampling around the head. The HRTFs are equalized using a minimum phase filter whose magnitude is the inverse of $|H_{DF}|$. Thus, the diffuse-field average of the equalized HRTFs is flat. Figure 11 shows the diffuse-field average of the HRTFs. It is dominated by the ear canal resonance at 2-3 kHz. The low-frequency dropoff is a result of the poor low-frequency response of the measurement speaker. The inverse equalizing filter was gain limited to prevent excessive noise amplification at extreme frequencies. In addition to the diffuse-field equalization, the HRTFs were sample rate converted to 32 kHz. This was done in order to reduce the computational requirements of the spatializer. The final HRTFs were cropped to 128 points (4 msec) which was more than sufficient to capture the entire head response including interaural delays.

The spatializer convolves a monophonic input signal with a pair of HRTFs to produce a stereophonic (binaural) output. The HRTFs that are closest to the desired azimuth and elevation are used. For efficiency, the convolution is accomplished using an overlap-save block convolver [20] based on the fast Fourier transform (FFT). Because the impulse response is 128 points long, the convolution is performed in 128-point blocks, using a 256-point real FFT. The forward transforms of all HRTFs are pre-computed. For each 128-point block of input samples (every 4 msec), the forward transform of the samples is calculated, and then two spectral multiplies and two inverse FFTs are calculated to form the two 128-point blocks of output samples. In addition to the convolution, a gain multiplication is performed to control apparent distance.

It is essential that the position of the source can be changed smoothly without introducing clicks into the output. This is easily accomplished as follows. Every 12 blocks (48 msec) the new source position is sampled and a new set of HRTFs is selected. The input block is convolved with both the previous HRTFs and the new HRTFs, and the two results are crossfaded using a linear crossfade. This assumes reasonable correlation between the two pairs of HRTFs. Subsequent blocks are processed using the new HRTFs until the next position is sampled. The sampling rate of position updates is about 20 Hz, which is quite adequate for slow moving sources.

2.4 Performance of binaural spatializer

The binaural spatializer (single source, 32 kHz sampling rate) runs in realtime on an SGI Indigo workstation. Source position can be controlled using a MIDI (Musical Instrument Digital Interface) controller that has a set of sliders which are assigned to control azimuth, elevation, and distance (gain). A constant amount of reverberation can be mixed into the final output using an external reverberator as shown in figure 10.

The spatializer was evaluated using headphones (AKG-K240, which are diffuse-field equalized [23, 18]). The input sound, usually music or sound effects, was taken from one channel of a compact disc player. The spatializer worked quite well for lateral and rear directions for all listeners. As expected, some listeners had problems with front-back reversals. Elevation control off the medial plane was also good, though this varied considerably among listeners. All listeners experienced poor externalization of frontal sounds. At zero degrees elevation, as the source was panned across the front, the perception was always of the source moving through the head between the ears, or sometimes over the top of the head. Externalization was far better at lateral and rear azimuths. Adding reverberation did improve the realism of the distance control, but did not fix the problem of frontal externalization. Clearly a problem of using non-individualized HRTFs with headphones is the difficulty of externalizing frontal sources.

In order to increase the number of sources, or to add integral reverberation, the performance of the spatializer would need to be improved. Several things could be done:

- Reduce the filter size by simple cropping (rectangular windowing).
- Reduce the filter size by factoring out the interaural delay and implementing this separately from the convolution.
- Reduce the filter size by using minimum phase filters.
- Model the HRTFs using infinite impulse response (IIR) filters.

Many of these strategies are discussed in [15]. To obtain the best price to performance ratio, commercial spatializers attempt to be as efficient as possible, and usually run on dedicated DSPs. Consequently, the filters are modeled as efficiently as possible and the algorithms are hand-coded. In addition, there are usually serious memory constraints which prevent having a large database of HRTFs, and thus parameterization and interpolation of HRTFs is an important issue. Lack of memory is not a problem in our implementation.

2.5 Principles of transaural audio

Transaural audio is a method used to deliver binaural signals to the ears of a listener using stereo loudspeakers. The basic idea is to filter the binaural signal such that the subsequent stereo presentation produces the binaural signal at the ears of the listener. The technique was first put into practice by Schroeder and Atal [22, 21] and later refined by Cooper and Bauck [10], who referred to it as "transaural audio". The stereo listening situation is shown in figure 12, where $\hat{x_L}$ and $\hat{x_R}$ are the signals sent to the speakers, and y_L and y_R are the signals at the listener's ears. The system can be fully described by the vector equation:

$$\mathbf{y} = \mathbf{H}\hat{\mathbf{x}} \tag{5}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_L \\ y_R \end{bmatrix}, \mathbf{H} = \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix}, \hat{\mathbf{x}} = \begin{bmatrix} \hat{x}_L \\ \hat{x}_R \end{bmatrix} \tag{6}$$

and H_{XY} is the transfer function from speaker X to ear Y. The frequency variable has been omitted.

If $\mathbf{x}$ is the binaural signal we wish to deliver to the ears, then we must invert the system transfer matrix $\mathbf{H}$ such that $\hat{\mathbf{x}} = \mathbf{H}^{-1}\mathbf{x}$. The inverse matrix is:

$$\mathbf{H}^{-1} = \frac{1}{H_{LL}H_{RR} - H_{LR}H_{RL}} \begin{bmatrix} H_{RR} & -H_{RL} \\ -H_{LR} & H_{LL} \end{bmatrix} \tag{7}$$

This leads to the general transaural filter shown in figure 13. This is often called a crosstalk cancellation filter, because it eliminates the crosstalk between channels. When the listening situation is symmetric, the inverse filter can be specified in terms of the ipsilateral ($H_i = H_{LL} = H_{RR}$) and contralateral ($H_c = H_{LR} = H_{RL}$) responses:

$$\mathbf{H}^{-1} = \frac{1}{H_i^2 - H_c^2} \begin{bmatrix} H_i & -H_c \\ -H_c & H_i \end{bmatrix} \tag{8}$$

Cooper and Bauck proposed using a "shuffler" implementation of the transaural filter [10], which involves forming the sum and difference of x_L and x_R, filtering these signals, and then undoing the sum and difference operation. The sum and difference operation is accomplished by the unitary matrix $\mathbf{D}$ below, called a shuffler matrix or MS matrix:

$$\mathbf{D} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{9}$$

It is easy to show that the shuffler matrix $\mathbf{D}$ diagonalizes the matrix $\mathbf{H}^{-1}$ via a similarity transformation:

$$\mathbf{D}^{-1}\mathbf{H}^{-1}\mathbf{D} = \begin{bmatrix} \frac{1}{H_i+H_c} & 0 \\ 0 & \frac{1}{H_i-H_c} \end{bmatrix} \tag{10}$$

Thus, in shuffler form, the transaural filters are the inverses of the sum and the difference of H_i and H_c. Note that $\mathbf{D}$ is its own inverse. This leads to the transaural filter shown in figure 14. The $1/\sqrt{2}$ normalizing gains can be commuted to a single gain of $1/2$ for each channel, or can be ignored.

In practice, the transaural filters are often based on a simplified head model. Here we list a few possible models in order of increasing complexity:

- The ipsilateral response is taken to be unity, and the contralateral response is modeled as a delay and attenuation [21].
- Same as above, but the contralateral response is modeled as a delay, attenuation, and lowpass filter [2].
- The head is modeled as a rigid sphere [10].
- The head is modeled as a generic human head without pinna.

At high frequencies, where pinna response becomes important (> 8 kHz), the head effectively blocks the crosstalk between channels. Furthermore, the variation in head response for different people is greatest at high frequencies [19]. Consequently, there is little point in modeling pinna response when constructing a transaural filter.

2.6 Implementation of transaural filter

Our transaural filter is based on a simplified head model suggested by David Griesinger. The ipsilateral response is taken to be unity and the contralateral response is modeled as a delay, attenuation, and a lowpass filter:

$$H_i(z) = 1, H_c(z) = gz^{-m}H_{LP}(z) \tag{11}$$
$$H_{LP}(z) = \frac{1-a}{1-az^{-1}}$$

[2]David Griesinger, personal communication

where $g < 1$ is a broadband interaural gain, m is the interaural time delay (ITD) in samples, and $H_{LP}(z)$ is a one-pole, DC-normalized, lowpass filter that models the frequency dependent head shadowing. For a horizontal source at 30 degrees azimuth, typical contralateral parameters might be g = 0.85 (-1.5 dB broadband attenuation), m = 7 (ITD of 0.2 msec at 32 kHz) and a lowpass filter cutoff of 1000 Hz (the frequency where head shadowing becomes significant). These parameters were not in fact calculated but were established through a calibration procedure discussed below.

Using this simplified head model the transaural filter in shuffler form is given by:

$$\mathbf{H}^{-1}(z) = \mathbf{D} \begin{bmatrix} \frac{1}{1+gz^{-m}H_{LP}(z)} & 0 \\ 0 & \frac{1}{1-gz^{-m}H_{LP}(z)} \end{bmatrix} \mathbf{D} \tag{12}$$

This filter structure is efficiently implemented using only two delays and two lowpass filters.

The transaural filter is calibrated as follows. A standard stereo listening setup was constructed with speakers at ±30 degrees with respect to the listener. Several stereo test signals are sent through the transaural filter and presented to the listener. The signals include stereo uncorrelated pink noise, left only pink noise and right only pink noise, and commercial binaural recordings made with dummy head microphones. During playback, the listener can continuously adjust the three transaural parameters (g, m, and the lowpass cutoff frequency) using a MIDI controller. The calibration procedure involves adjusting the parameters such that single sided noises are located as close as possible to their corresponding ears and the stereo noise is maximally enveloping [11]. The interaural delay parameter has the most effect of steering the signal and changing the timbre, provided the gain parameter is sufficiently close to 1. The lowpass cutoff has the most subtle effect. Interestingly, while it is possible to steer the single sided noise close to the corresponding ear, this often has the effect of moving the opposite sided noise closer to its corresponding speaker. Consequently a compromise has to be reached. In general, the final parameters one obtains via the calibration procedure agree with the physics of the situation.

Listening to the binaural recordings through the transaural system is an enjoyable experience. The speakers vanish and are replaced by an immersive auditory scene. Sounds can be heard from all directions except directly behind the listener. The so-called "sweet spot" is readily apparent. When one moves outside of the sweet spot, the sensation of being surrounded by sound is lost, and is replaced by the the sensation of listening to a conventional stereo setup. Within the sweet spot, the transaural system is tolerant of head motion, particularly front-back translation, less so for left to right translation, and least tolerant of turning side to side. Turning to face either loudspeaker is sufficient to destroy the illusion.

2.7 Performance of combined system

The binaural spatializer and transaural filter were combined into a single program which runs in realtime on an SGI Indigo workstation.

Listening to the output of the binaural spatializer via the transaural system is considerably different than listening over headphones. Overall, the spatializer performance is much improved by using transaural presentation. This is primarily because the frontal imaging is excellent using speakers, and all directions are well externalized. The drawback of transaural presentation is the difficulty in reproducing extreme rear directions. As the sound is panned from the front to the rear, it often suddenly flips back to a frontal direction as the illusion breaks down. Most listeners can easily steer the sound to about 120 degrees azimuth before the front-back flip occurs. It is easier to move the sound to the rear with the eyes closed.

Elevation performance with transaural presentation is not as good as with headphone presentation. However, because the sounds are more externalized with the speakers, changing either the azimuth or elevation induces more apparent motion than with headphone presentation. Many listeners reported that changing the elevation also caused the azimuth to change. For instance, starting the sound directly to the right and moving it up often causes the sound to move left towards center before it reaches overhead.

All the performance evaluation discussed is completely informal. It would be useful to have an efficient procedure for evaluating the performance of such systems, one that does not require lengthy training sessions or experimentation.

2.8 Adding dynamic tracking

We now discuss efforts underway to extend this technology by adding dynamic head tracking capability. The head tracker should provide the location and orientation of the head. For simplicity, we will only consider situations where the head is upright and in the horizontal plane of the speakers. There are two possible uses for the tracking information:

- Update the parameters of the transaural filter based on the tracked head position and orientation. Thus, as the listener moves about, the transaural filter parameters are adjusted to move the "sweet spot" along with the listener.

- Update the positions of sound sources according to the tracked head position, so that the auditory scene remains fixed as the listener moves, rather than moving with the listener.

As discussed earlier, the ALIVE project uses video cameras to track people using the system. The *Pfinder* program can track various features of the human body, including the head and hands [4]. With a single camera viewing a standing person, the distance between the camera and the person is calculated by finding the position of the feet relative to the bottom of the image. *Pfinder* assumes the person is a vertical plane, and thus the head and hands are assumed to be equidistant from the camera. Features in the plane of the person are determined from the feature positions in the 2-D video image and the calculated distance of the person. Another approach uses stereo cameras and two *Pfinder* programs to calculate distances to objects by considering parallax. Neither of these systems are currently able to reliably estimate the orientation of the head, because the facial features are too small. However, orientation of the head can be estimated from a closeup view of the face. This is accomplished by obtaining templates for the eyes, nose and mouth, recovering these feature positions via normalized correlation, and assuming an elliptical head shape [8].

2.9 Preliminary results

In order to experiment with head tracking in the context of transaural 3-D audio, we are currently using a Polhemus tracking system. This system returns the position and orientation of a sensor with respect to a transmitter (6 degrees of freedom). The sensor can be easily mounted on headphones or a cap to track head position and orientation. The head position and orientation can be used to update the parameters of the 3-D spatializer and transaural audio system. Results are preliminary at this time.

The strategy used to update transaural parameters based on head position and orientation obviously depends greatly on the head model used for the transaural filter. We used the simple head model given in equation 11. The following points were observed:

- For front-back motions, the symmetrical transaural filter can be used, and the interaural delay can be adjusted as a function of distance between the speakers and the listener. This has been tested and seems to be effective.

- For left-right motions and head rotations, the symmetrical transaural filter is no longer correct. The general form of the transaural filter (equation 7) may be used instead, but at much greater computational cost. It may be better to abandon the simplified IIR model and use an FIR implementation based on a more realistic head model [21].

- To compensate for head rotations, the general form of the transaural filter (equation 7) was implemented with the simplified head model (equation 11). The resulting dynamic filter compensated for the changing path lengths between the speakers and the ears in order to keep the cancellation signals aligned properly. However, this filter did not function effectively. In part this was due to the audible artifacts from the various linearly interpolated delay lines used in the implementation. However, it may be that small time adjustments are in fact unnecessary, judging from the insensitivity to small head rotations with a static transaural system. This is also suggested in the literature on equalizing dummy-head recordings for loudspeaker reproduction [14].

Using the static, symmetrical transaural system described earlier, the head tracking information was also used to update the positions of 3-D sounds so that the auditory scene remained fixed as the listener's head rotated. This gives the sensation that the source is moving in the opposite direction, rather than remaining fixed. There is a good reason for this. Using a static transaural system, the position of rendered sources remains fixed as the listener changes head orientation (provided that the change in head orientation is small enough to maintain the transaural illusion). This is contrary to headphone presentation, where the auditory scene moves with the head, even for small motions. Thus, the transaural presentation doesn't require compensation for small head motions, and if the compensation is provided, it is perceived as motion in the opposite direction. We hoped that this form of head tracking would provide dynamic localization cues to improve rear localization, but this is inconclusive. Despite the fact that head orientation should be decoupled from the positions of rendered sources, it may be important to compensate for listener position, in order that the listener can walk past virtual sources and sense the direction of the source changing.

2.10 Conclusions

We have discussed a single source transaural spatializer that runs in realtime on an SGI Indigo workstation. Despite a simple implementation, the informal performance results are quite good. We are currently working to improve this basic system by adding dynamic head tracking so that veridical 3-D audio cues are maintained as the listener moves in the space.

References

[1] ANSI. S3.5-1969,*American National Standard Methods for the Calculation of the Articulation Index*. American National Standards Institute, New York, 1969.

[2] H. Cox. "Robust Adaptive Beamforming" *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1365-1376, 1987.

[3] F. Khalil, J.P. Jullien, and A. Gilloire. "Microphone Array for Sound Pickup in Teleconference Systems". *Journal of the Audio Engineering Society*, 42(9):691-699, 1994.

[4] P. Maes, T. Darrell, B. Blumberg, and A. Pentland. "The ALIVE System: Full-body Interaction with Autonomous Agents". *Proceedings of the Computer Animation Conference*, Switzerland, IEEE Press, 1995.

[5] R.J. Mailloux. *Phased Array Antenna Handbook.* Artech House, Boston, London, 1994.

[6] W. Soede, A.J. Berkhout, and F.A. Bilsen. "Development of a Directional Hearing Instrument Based on Array Technology". Journal of the Aoustical Society of America, 94(2):785-798, 1993.

[7] R.W. Stadler and W.M. Rabinowitz. "On the Potential of Fixed Arrays for Hearing Aids". *Journal of the Acoustical Society of America*, 94(3):1332-1342, 1993.

[8] Ali Azarbayejani, Thad Starner, Bradley Horowitz, and Alex Pentland. Visually controlled graphics. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):602–605, June 1993. (Special Section on 3-D Modeling in Image Analysis and Synthesis).

[9] Durand R. Begault. *3-D Sound for Virtual Reality and Multimedia.* Academic Press, Cambridge, MA, 1994.

[10] Duane H. Cooper and Jerald L. Bauck. "Prospects for Transaural Recording". *J. Audio Eng. Soc.*, 37(1/2):3–19, 1989.

[11] P. Damaske. "Head-Related Two-Channel Stereophony with Loudspeaker Reproduction". *J. Acoust. Soc. Am.*, 50(4):1109–1115, 1971.

[12] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. "On the Externalization of Auditory Images". *Presence*, 1(2):251–257, 1992.

[13] W. G. Gardner and K. D. Martin. "HRTF measurements of a KEMAR". *J. Acoust. Soc. Am.*, 97(6):3907–3908, 1995.

[14] D. Griesinger. "Equalization and Spatial Equalization of Dummy-Head Recordings for Loudspeaker Reproduction". *J. Audio Eng. Soc.*, 37(1/2):20–29, 1989.

[15] J. M. Jot, Veronique Larcher, and Olivier Warusfel. "Digital signal processing issues in the context of binaural and transaural stereophony". In *Proc. Audio Eng. Soc. Conv.*, 1995.

[16] Keith D. Martin. A computational model of spatial hearing. Master's thesis, MIT Dept. of Elec. Eng., 1995.

[17] Henrik Moller, Dorte Hammershoi, Clemen Boje Jensen, and Michael Fris Sorensen. "Transfer Characteristics of Headphones Measured on Human Ears". *J. Audio Eng. Soc.*, 43(4):203–217, 1995.

[18] Henrik Moller, Clemen Boje Jensen, Dorte Hammershoi, and Michael Fris Sorensen. "Design Criteria for Headphones". *J. Audio Eng. Soc.*, 43(4):218–232, 1995.

[19] Henrik Moller, Michael Fris Sorensen, Dorte Hammershoi, and Clemen Boje Jensen. "Head-Related Transfer Functions of Human Subjects". *J. Audio Eng. Soc.*, 43(5):300–321, 1995.

[20] Alan V. Oppenheim and Ronald W. Schafer. *Discrete Time Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, 1989.

[21] M. R. Schroeder. "Digital simulation of sound transmission in reverberant spaces". *J. Acoust. Soc. Am.*, 47(2):424–431, 1970.

[22] M. R. Schroeder and B. S. Atal. "Computer simulation of sound transmission in rooms". *IEEE Conv. Record*, 7:150–155, 1963.

[23] G. Theile. "On the Standardization of the Frequency Response of High-Quality Studio Headphones". *J. Audio Eng. Soc.*, 34:956–969, 1986.

[24] E. M. Wenzel. "Localization in Virtual Acoustic Displays". *Presence*, 1(1):80–107, 1992.

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

[26] F. L. Wightman and D. J. Kistler. "Headphone simulation of free-field listening". *J. Acoust. Soc. Am.*, 85:858–878, 1989.

Figure 1: Target and Ambient Sound in the ALIVE Space

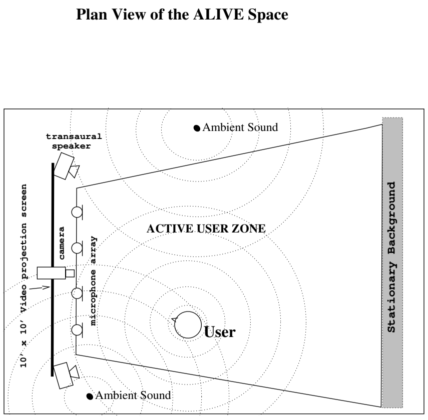


Figure 2: Gesturing the "Beg" Command to Silas

Figure 3: Endfire Microphone Array Geometry and Notation

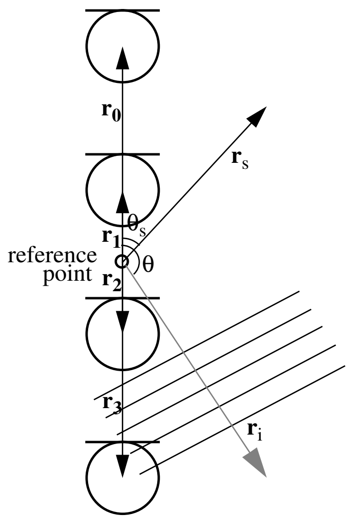


Figure 4: Broadside Microphone Geometry

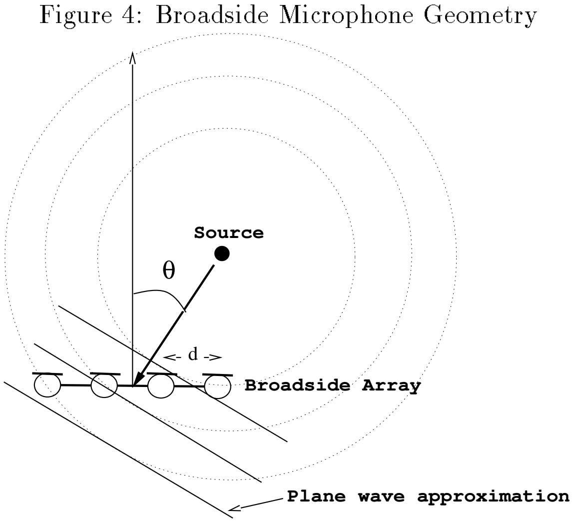
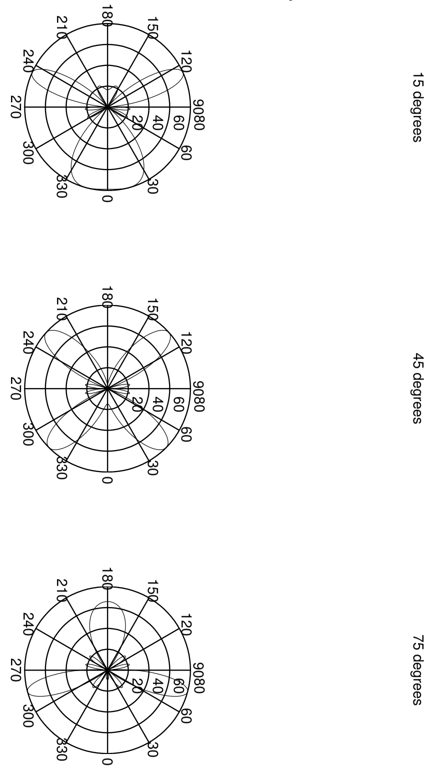


Figure 5: Directivity Pattern of Endfire Array with Omnidirectional Elements



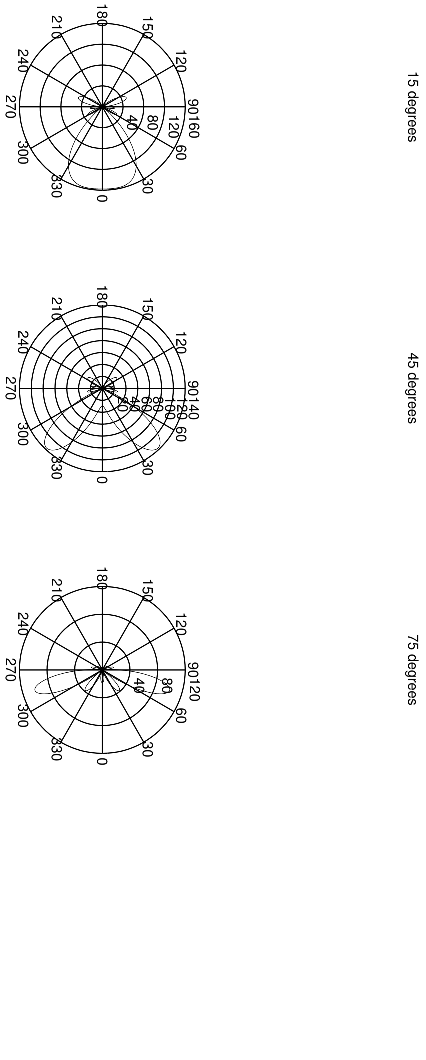


Figure 6: Directivity Pattern of Endfire Array with Cardioid Elements

Figure 7: Directivity Pattern of Broadside Array with Cardioid Elements

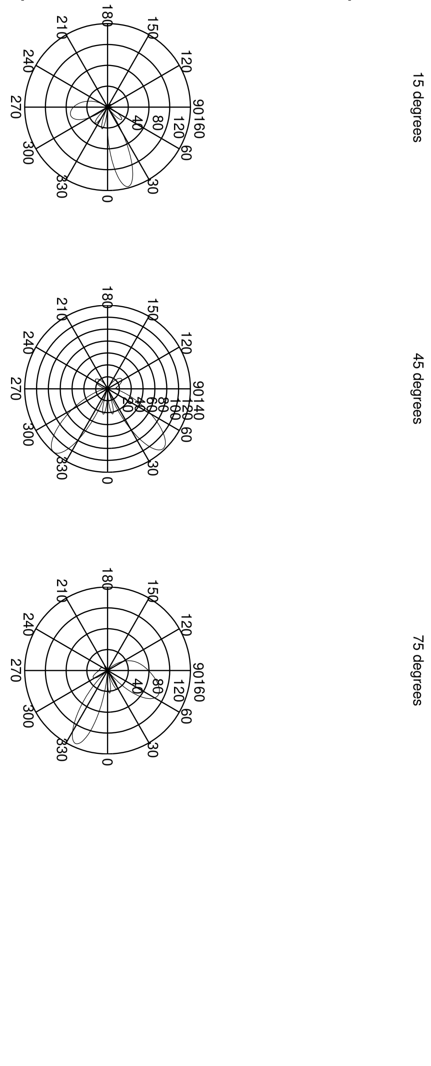
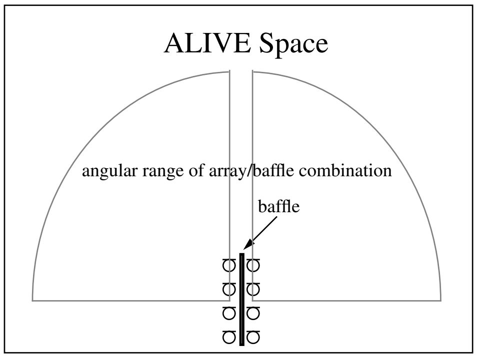


Figure 8: Dual Endfire Arrays with Acoustic Baffle



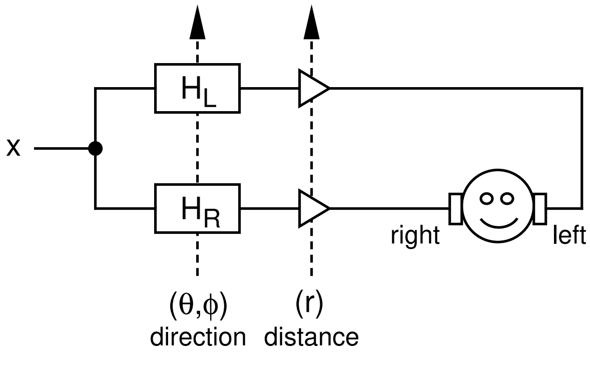


Figure 9: Single source binaural spatializer.

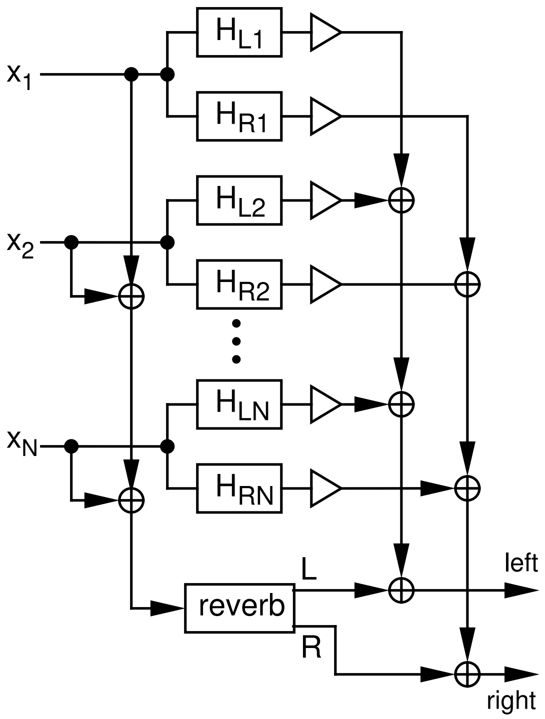


Figure 10: Multiple source binaural spatializer with reverberation.

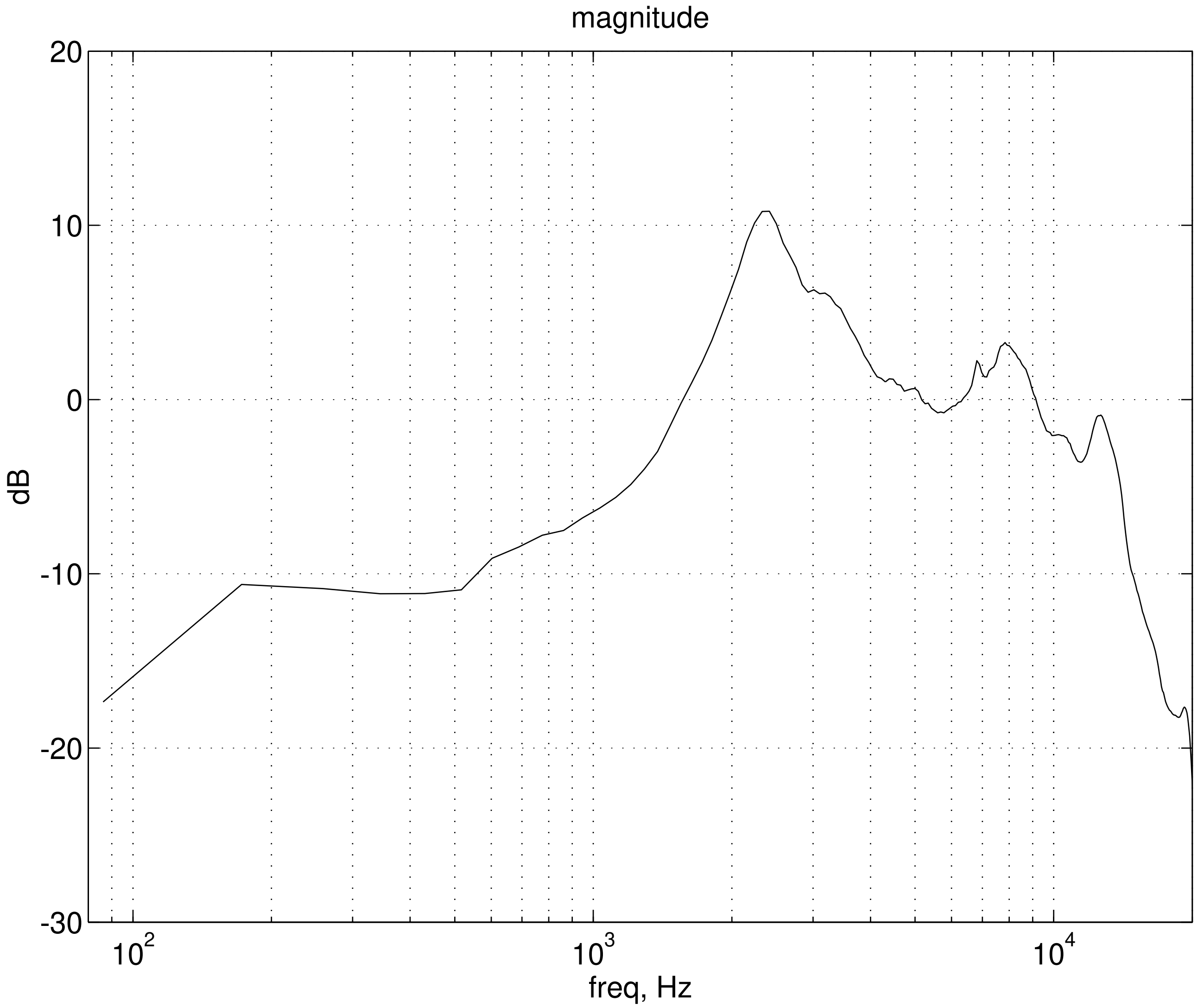


Figure 11: Diffuse-field average of KEMAR HRTFs.

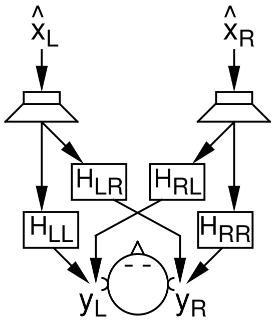


Figure 12: Transfer functions from speakers to ears in stereo arrangement.

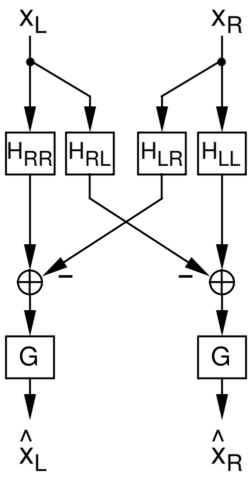


Figure 13: General transaural filter, where $G = 1/(H_{LL}H_{RR} - H_{LR}H_{RL})$.

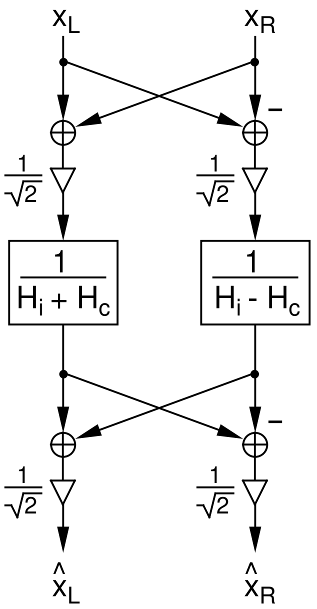


Figure 14: Shuffler implementation of transaural filter for symmetric listening arrangement.