M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 373
Appears: Proceedings of IMAGE'COM '96, Bordeaux, France, May 1996

Vision-Steered Audio for Interactive Environments

Sumit Basu, Michael Casey, William Gardner, Ali Azarbayejani, and Alex Pentland
Perceptual Computing Section, The MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139 USA
{sbasu,mkc,billg,ali,sandy}@media.mit.edu

Abstract

We present novel techniques for obtaining and producing audio information in an interactive virtual environment using vision information. These techniques are free of mechanisms that would encumber the user, such as clip-on microphones, headphones, etc. Methods are described for both extracting sound from a given position in space and for rendering an "auditory scene," i.e., given a user location, producing sounds that appear to the user to be coming from an arbitrary point in 3-D space. In both cases, vision information about user position is used to guide the algorithms, resulting in solutions to problems that are difficult and often impossible to robustly solve in the auditory domain alone.

1 Introduction

In the design and development of interactive environments, we have strived to allow free and natural interaction with a synthetic world. A vision system (such as the one described in a section below) that can track a user, locate individual body parts, and recognize gestures allows such interaction to occur in the visual domain. However, for truly natural interaction, the system must be able to localize audio information coming from the user and produce audio information that appears to be coming from different regions of the synthetic environment. Of course, these problems are easily solved if the user is fit with a wireless microphone and headphone set. However, using such cumbersome hardware to solve the problem constrains a user in an unnatural way, just as special clothing or motion sensors would for the analagous vision problem. The objective is not for the user to have to adapt to the environment, but for the environment to adapt to the user. The user should not have to change her appearance or carry special equipment in order to interact with the environment.

In this paper, we present techniques for both obtaining and producing audio information that adapt to the user's position using vision information. The first problem we approach with a phased array of microphones; the latter with binaural spatialization and transaural rendering.

2 Overview of the Vision System

In order to frame our discussion, we first present a brief overview of Pfinder (Person finder), a real-time vision system for tracking and interpretation of people used in our interactive environment (for a more detailed account of the system, please refer to [20] and [10]). Pfinder has the capability to accurately determine the 3-D locations of the user's head and other features in real-time at a frame rate of 10Hz and an accuracy of 10cm. With two cameras (stereo Pfinder), the accuracy can be refined to 1.5cm. The audio techniques described in the rest of the paper depend on this to steer their respective responses/outputs.

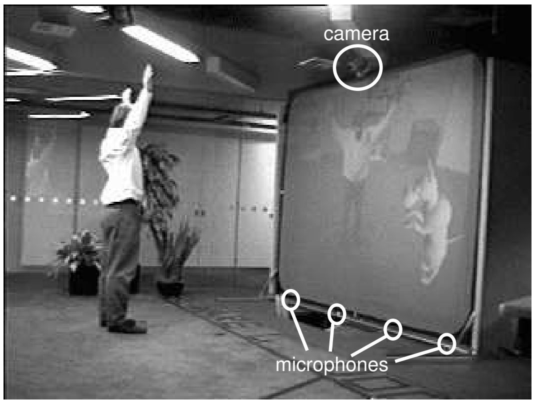

Figure 1: Location of the camera and microphone array in the virtual environment

In our setup, a camera facing the user is mounted on the video screen displaying the virtual environment (see Figure 1). The system uses a statistical model of color and shape to segment a person from a background scene and then to find and track body parts in a wide range of viewing conditions. It has performed reliably on thousands of people in many different physical locations.

Pfinder models the human as a connected set of blobs. Each blob has a spatial and color Gaussian distribution, and a *support map* that indicates which image pixels are members of each blob. The combination of these support maps segments the input image into the various blob classes.

The statistics of each blob are recursively updated to combine information contained in the most recent measurements with knowledge contained in the current class statistics and the priors. Because the detailed dynamics of each blob are unknown, we use approximate models derived from experience with a wide range of users. For instance, blobs that are near the center of mass have substantial inertia, whereas blobs toward the extremities can move much faster.

3 Obtaining Audio Information

Our original motivation for seeking directed audio input from the environment was for speech recognition. We desired to have agents in the environment react to speech from the user while allowing the user to move about freely. A task like speech recognition requires the high signal to noise ratio of a near-field (i.e., clip-on or noise-cancelling) microphone. However, we were unwilling to encumber the user with such devices, and thus faced the problem of getting high quality audio input from a distance.

This leaves 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 not only requires a significant amount of mounting and control hardware, it is also 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.

3.1 The Beamforming Approach - with a Twist

This goal can be achieved with the well-known technique of beamforming with an array of microphone elements. The signals from several omnidirectional or partially directional (i.e., cardioid) microphones are combined to form a more directional response pattern. 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 by a single microphone array. The setup of the array used in our implementation is shown in Figure 1 and Figure 2.

Beamforming is formulated in two flavors: fixed and adaptive. In fixed beamforming, it is assumed that the position of the sound source is both known and static. An algorithm is then constructed to combine the signals from the different microphones to maximize the response to signals coming from that position. This works quite well, assuming the sound source is actually in the assumed position. Because the goal is to have a directional response, this method is not robust to the sound source moving significantly from its assumed position. In adaptive beamforming, on the other hand, the position of the source is neither known nor static. The position of the source must continuously be estimated by analyzing correlations between adjacent microphones, and the corresponding fixed beamforming algorithm must be applied for the estimated position. This does not tend to work well whenever there are multiple sources of sound, since there are high correlations for multiple possible sound source positions. It is difficult and often impossible to tell which of these directions corresponds to the sound of interest, e.g., the voice of the user.

Our solution to this problem is a hybrid of these two flavors and a twist from another domain. Instead of using the audio information to determine the location of the sound source(s) of interest, we use the vision system, which exports the 3-D position of the user's head. Using this information, we formulate the fixed beamforming algorithm for this position to combine the outputs of the microphone array. This algorithm is then updated periodically (5 Hz) with the vision information. As a result, we have the advantages of a static beamforming solution that is adaptive through the use of vision information.

Beamforming is a relatively old techique; it was developed in the 1950's for radar applications. In addition, its use in microphone arrays has been widely studied [6, 9, 17, 18]. We certainly do not claim to have developed the "optimal" beamforming strategy for an interactive environment: we leave that task to the audio engineering community. In fact, our approach to beamforming is among the simplest possible. However, this is sufficient to greatly improve the signal to noise ratio to the point where the speech recognizer can correctly process the signal, i.e., close to the level of a near-field microphone.

3.2 Theoretical Formulation of the Phased Array

In this section, we present a brief theoretical overview of the beamforming algorithms for a phased array of microphones. Further details for the system we have implemented can be found in [4]; further details on beamforming in general can be found in [11].

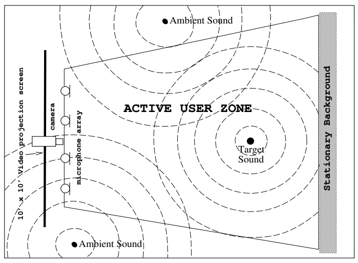

Figure 2: Target and Ambient Sound in our Virtual Environment

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.

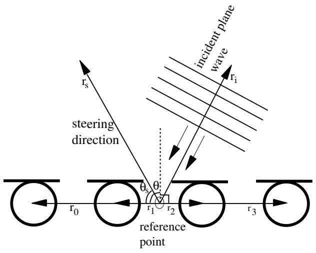

Figure 3: Broadside Microphone Array Geometry and Notation

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 $a_n = |a_n| \, e^{-jk_o \mathbf{r}_n \cdot \hat{\mathbf{r}}_s}$ and $F(\theta)$ 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 that there are a variety of ways of optimizing the $|a_n|$ values in $\mathbf{W}$.

The standard performance metric for the directionality of a fixed array is the *directivity index* which is shown in Equation 3 [18]. 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:

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

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* [18] 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 [18].

3.3 Designing the Array

An important first consideration is the choice of array geometry. Two possible architectures were considered; endfire (not shown) and broadside Figure 3. 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* [9] [18]. For omnidirectional microphones, the gain patterns for the two layouts are identical but for a rotation. In our implementation, cardioid microphones were used and were placed in a broadside arrangement due to space constraints (see Figure 2). The polar response patterns for this arrangement are shown in Figure 4.

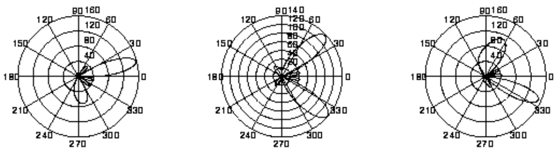

Figure 4: Directivity Pattern of Broadside Array with Cardioid Elements steered at 15, 45, and 75 degrees. Note that the reference point of the broadside array geometry (Figure 3) should be aligned with the center of each polar plot

A detailed examination of the response patterns with the different array geometries and element responses is developed in [4]. Through this study, it was found that four microphones in endfire arrange would provide a very directional beam, but would produce a symmetric lobe at $-\theta$. 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 detailed account of this proposed setup is in [4].

4 Producing Audio Information

We have only presented half of the story so far; we have yet to show how we return audio information to the user. To truly create a 3-D feel in the virtual environment, sound sources in different locations in the virtual environment must sound as though they were physically in those locations. In other words, it is not sufficient to simply send all of the sound through a single loudspeaker.

The naive solution to this problem is a balance control scheme, i.e., setting up four or more speakers surrounding the user and then adjusting the level of a given sound on each speaker. For example, a sound source to the front and left of a user would be simulated by increasing the level of the sound on the front left speaker and reducing the level (or cutting it off) on the other speakers. A sound source in between two speakers would be simulated by mediating the levels between the two closest speakers.

This solution doesn't work for relatively subtle reasons that have their basis in the human auditory system. We perceive the location of a sound not only on the basis of the magnitude difference between the two ears (i.e., balance), but also on the basis of the phase and timing difference between the ears (see p.99 of [7]). Though this latter difference may seem to be small, human listeners can detect interaural time differences as short as 0.01 msec, which corresponds to a difference in sound source orientations of roughly one degree [7]. It has been shown that we use both magnitude and phase information to perform the subtle discrimination tasks we are capable of, such as being able to discern the words of one person from those of an adjacent person (the canonical "cocktail party" problem). Thus, in order to exploit this perceptual capability and create the illusion of a 3D auditory scene, it is necessary to accurately reproduce both the phase and magnitude of the virtual sound source.

4.1 The Phase-Magnitude Solution

Indeed, the correct phase and magnitude for a given pair of sound source position and user position can be found and constructed at each ear. We solve the problem in two parts: a technique known as binaural spatialization can be used to find the sound that each ear should receive. A second stage can then do "transaural rendering" to produce these sounds for a given user location from two statically positioned frontal speakers. There are some obvious difficulties with this approach - the signal that supplies the correct signal to one ear will travel through the transfer function of the head and reach the other ear, and thus must be cancelled by the negative of the resultant signal at this ear. This cancellation signal must then be cancelled at the first ear, and so on. Though complex, this does not render the solution impractical. The cancellation described can be achieved quite effectively, and the computation necessary to do both the binaural spatialization and the transaural rendering can be performed on a single Silicon Graphics Indigo workstation.

The basics of the theory behind these techniques is presented below. We first demonstrate the spatialization process with headphones and then extend this to the free-field situation with transaural rendering. For a more detailed discussion and a description of the system used in our virtual environment, please refer to [4].

4.2 Audio Synthesis Principles

As described above, a binaural spatializer simulates the auditory experience of one or more sound sources arbitrarily located around a listener [3]. 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 to ear frequency response [19, 13]. A schematic diagram of a single source system is shown in Figure 4.2. The direction of the source (θ = azimuth, ϕ = elevation) determines which pair of HRTFs to use, and the distance (r) determines the gain. A multiple source spatializer then adds a constant level of reverberation to enhance distance perception (see [4]).

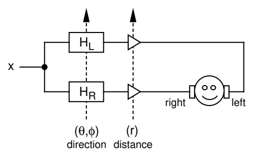

Figure 5: Single source binaural spatializer.

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 [8], 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 [12].

4.3 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 [16, 15] and later refined by Cooper and Bauck [5], who referred to it as "transaural audio". The stereo listening situation is shown in Figure 6, 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:

$$\mathbf{y} = \mathbf{H}\hat{\mathbf{x}} \tag{4}$$

where:

$$\mathbf{y} = \left[\begin{array}{c} y_L \\ y_R \end{array}\right], \mathbf{H} = \left[\begin{array}{cc} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{array}\right], \hat{\mathbf{x}} = \left[\begin{array}{c} \hat{x}_L \\ \hat{x}_R \end{array}\right] \tag{5}$$

[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.

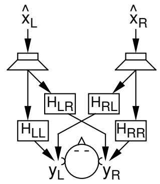

Figure 6: Transfer functions from speakers to ears in stereo arrangement.

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:

$$\mathbf{H}^{-1} = \frac{1}{H_{LL}H_{RR} - H_{LR}H_{RL}} \left[\begin{array}{cc} H_{RR} & -H_{RL} \\ -H_{LR} & H_{LL} \end{array}\right] \tag{6}$$

This leads to the general transaural filter shown in Figure 7. 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:

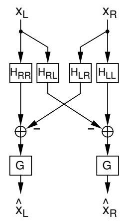

Figure 7: General transaural filter, where $G = 1/(H_{LL}H_{RR} - H_{LR}H_{RL})$.

$$\mathbf{H}^{-1} = \frac{1}{H_i^2 - H_c^2} \left[\begin{array}{cc} H_i & -H_c \\ -H_c & H_i \end{array}\right] \tag{7}$$

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 [15].
- 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 [5].
- 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 [14]. Consequently, there is little point in modeling pinna response when constructing a transaural filter.

[2]Suggested by David Griesinger in personal communication

4.4 Performance of combined system

The binaural spatializer and transaural filter were combined into a single program which runs in real time 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 and 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.

4.5 Current Work

We now discuss efforts underway to extend this technology by adding 6 DOF head tracking capability. The head tracker should provide the location and orientation of the head. The current system can provide an accuracy of 10cm with a single camera and 1.5cm with a stereo pair in real time (10 Hz) but no orientation information. While this is more than accurate enough for the adaptive beamforming algorithm, it is not sufficient for high-quality transaural rendering: the detailed orientation of the head is also necessary.

To attain this additional information, we can use the 6 DOF rigid motion head-tracking algorithm described in [2]. This method models the head as a rigid ellipsoid and projects the frame to frame motion onto the possible rigid motions of the model. Plots of the orientation tracking are shown for a calibrated sequence in Figure 8. The orientation is correct within .2 radians (12 degrees) over a large range of motions. This method has been found to be robust over many frames and a variety of heads. We are currently working to make this tracking system run in real time.

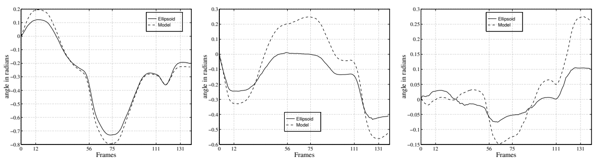

Figure 8: Head-tracking results for calibrated sequence: plots shown are for the alpha, beta, and gamma parameters (rotations around the z,y, and x axes, respectively).

4.6 Preliminary results

In order to simulate the head tracking while a real-time implementation of this method is developed, we are currently using a Polhemus sensor. This sensor returns the position and orientation of a sensor with respect to a transmitter (6 degrees of freedom). The head position and orientation can be used to update the parameters of the 3-D spatializer and transaural audio system.

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 suggested by Dave Griesinger, in which the ipsilateral response is taken to be unity and the contralateral response is modelled as a delay, attenuation, and a lowpass filter:

$$\begin{aligned} H_i(z) &= 1 \\ H_c(z) &= g z^{-m} H_{LP}(z) \\ H_{LP}(z) &= \frac{1-a}{1-az^{-1}} \end{aligned} \tag{8}$$

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. 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 6) 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 [15].

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. As a result, the perception of the rendered sound source locations is *stronger* if small head rotations are ignored.

5 Conclusions

We have presented techniques for the localized sensing and production of sound in an unencumbered environment. The key idea to absorb from this work is that we have used vision information to accomplish both of these tasks. It is the interaction of the two modalities that is truly interesting here: the fact that difficult or impossible problems in one domain can be solved with high level information from another. In addition, we have presented a general framework for audio interaction in virtual environments. It is not possible to fully develop the idea of a virtual environment without the inclusion of sound. In addition, if we want users to be able to interact freely with the environment, it does not seem reasonable to ask them to strap on microphones, headphones, or other sensors every time they use it. The methods we have presented are free from such constraints, and have been shown in preliminary tests to perform effectively in an interactive environment.

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] Sumit Basu, Irfan Essa, and Alex Pentland. "Motion Regularization for Model-Based Head Tracking". M.I.T. Media Laboratory Perceptual Computing Technical Report No. 362.

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

[4] Michael A. Casey, William G. Gardner, and Sumit Basu. "Vision Steered Beam-forming and Transaural Rendering for the Artificial Life Interactive Virtual Environment (ALIVE)". In *Proc. Audio Eng. Soc. Conv.*, 1995.

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

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

[7] Stephen Handel. *Listening: An Introduction to the Perception of Auditory Events.* MIT Press, Cambridge, MA, 1989.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

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

[20] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. "Pfinder: Real-Time Tracking of the Human Body". *SPIE Photonics East*, 2615:89–98, 1995.