Rendering Localized Spatial Audio in a Virtual Auditory Space

Dmitry N. Zotkin, Ramani Duraiswami, Larry S. Davis

{dz,ramani,lsd}@umiacs.umd.edu

Perceptual Interfaces and Reality Laboratory

Institute for Advanced Computer Studies

University of Maryland, College Park, MD 20742

Abstract

High-quality virtual audio scene rendering is a must for emerging virtual and augmented reality applications, for perceptual user interfaces, and sonification of data. We describe algorithms for creation of virtual auditory spaces by rendering cues that arise from anatomical scattering, environmental scattering, and dynamical effects. We use a novel way of personalizing the head related transfer functions (HRTFs) from a database, based on anatomical measurements. Details of algorithms for HRTF interpolation, room impulse response creation, HRTF selection from a database, and audio scene presentation are presented. Our system runs in real time on an office PC without specialized DSP hardware.

I. Introduction and Previous Work

Many emerging applications require the ability to render audio scenes that are consistent with reality. In multimodal virtual and augmented reality systems using personal visual and auditory displays the rendered audio and video must be kept consistent with each other and with the user's movements to create a virtual scene [1]. A goal of our work is to create rich auditory environments which can be used as user interfaces for both the visually-impaired and the sighted. These applications require the ability to render acoustical sources at their correct spatial location. Several studies have demonstrated the feasibility of the spatialized audio for data display, e.g., [2]. Real-time spatial displays using specialized hardware have been created [3] and virtual auditory displays have been used as user interfaces for the visually impaired [4], in mobile applications [5], or in the field of sonification ("the use of nonspeech audio to convey information" [6], [7]).

While there have been many successful attempts to render sound to create virtual environments, they have not as yet resulted in a comprehensive theory of rendering audio scenes, and they have not achieved the objective of a set of design rules and guidelines to render consistent personalized design spaces. Often auditory interfaces are based on use of pre-recorded samples created by artists to create auditory objects ("background music"), and on approximate spatial separation of sound. The lack of a theory and methodology has been held to be a chief stumbling block in holding back easy sonification of more complex data [7].

To develop a consistent way to render auditory scenes one must rely on an understanding of how humans segregate the streams of sound they receive into objects and scenes [8], [9], [10]. A key element of this ability, and that which is the main focus of this article, is the human ability to localize sound sources. To successfully render the spatial position of a source we must reintroduce the cues that lead to the perception of that location. This in turn demands an understanding of how the cues are generated and what is the relative importance of different cues [11]. Previous work in the area of localization and spatial sound rendering can be tracked back to the year 1907 [12]. Since then understanding of spatial localization [13], [14], modeling of the involved transfer functions [15], [16], [17], fast synthesis methods [18], environment modeling [19], [20], [21], and implementation of the rendering software [22], [23] have made significant progress. However none of these authors present a comprehensive account of how to render spatially consistent data in real time.

Our goal is to create an auditory display capable of spatial consistency. Achievement of spatial consistency requires rendering static, dynamic and environmental cues in the stream, otherwise the sound is perceived as being inside the head. The static cues are both the binaural difference-based cues, and the monaural and binaural cues that arise from the scattering process from the user's body, head and ears. These localization cues are encoded in the head-related transfer function (HRTF) which varies significantly between people, and it is known [24], [25] that differences in ear shape and geometry strongly distort the perception and that the high-quality synthesis of a virtual audio scene requires the personalization of the HRTF for the particular individual for good virtual source localization. Furthermore, once the HRTF-based cues are added back into the rendered audio stream, the sound is still perceived as non-externalized, since cues that arise from environmental reflections and reverberation are missing. Finally, for proper externalization and localization of the rendered source, dynamic cues must be added back to make the rendering consistent with the user's motion. Thus, dynamic and reverberation cues must be recreated for maximum sense of presence in the virtual audio scene.

In this paper, we present a set of fast algorithms for spatial audio rendering which are able to recreate all the mentioned cues in real time. Our rendering system works on a commercial off-the-shelf PC with no noticeable latency. No additional hardware is used except for the head tracker. This is achieved by using optimized algorithms so that only necessary parts of the spatial audio processing filters are recomputed in each rendering cycle and by highly optimized programming using novel features of the Intel Xeon processors. We also perform address the problem of personalization of the HRTF by selecting the HRTF that corresponds to the closes one from a database of 43 pairs of HRTFs. The selection of the closest HRTF is done by matching of certain anthropometric ear parameters with those in the database. We also present a preliminary investigation of how this personalization can improve the perception of the virtual audio scene.

The rest of the paper is organized as follows. In Section 2, we introduce the head-related transfer function which knowledge is crucial for accurate virtual audio scene rendering. In Section 3, we describe the environmental model which provides important cues for perceptions (in particular, cues that lead to "out-of-the-head" externalization) and in Section 4, the importance of dynamic cues for perception is outlined.. In Section 5, we describe the fast audio rendering algorithms. Section 6 deals with partial HRTF customization using visual matching of ear features. In Section 7, experimental setup and experimental results are presented. Section 8 concludes the paper and outlines directions of future research.

II. Head Related Transfer Function

Using just two receivers (ears), humans are able to localize sound with amazing precision [26]. While differences in the time of arrival or level between the signals reaching the two ears (known respectively as interaural time delay, ITD, and interaural level difference, ILD) [12] can partially explain this facility, interaural differences do not account for the ability to locate a source within the median plane, where both ITD and ILD are essentially zero. In fact, there are many locations in space that give rise to nearly identical interaural differences, yet under most conditions, listeners can determine which of these locations is the “true” source position. The localization is possible because of the other localization cues arising from sound scattering. The wavelength of audible sound $(2cm - 20m)$ is comparable to the dimensions of the environment, the human body, and for high audible frequencies, the dimensions of the external ear (pinna). As a result, the circularly-asymmetric external ear essentially forms a specially-shaped “antenna” that causes a location-dependent and frequency-dependent “filtering” of the sound reaching the eardrums. Thus, scattering of sound by the human body and by the external ears provides additional monaural (and, to a lesser extent, binaural) cues to source position.

The effect of both this scattering and the time and level differences can be described by a frequency response function called the head-related transfer function (HRTF). Knowing the HRTF, one can, in principle, reconstruct the exact pressure waveforms that would reach a listener’s ears for any arbitrary source waveform arising from the particular location. Although the way in which the auditory system extracts information from the stimuli at the ears is only partially understood, the pressure at the eardrums is a sufficient stimulus: if the correct sound pressure signals are presented to the listener’s ears, he will perceive a sound source at the correct location in exocentric space. The process of synthesizing the proper acoustic wave pattern that might have occurred in real environment with real source(s) is referred to as a synthesis of a virtual auditory space (VAS)

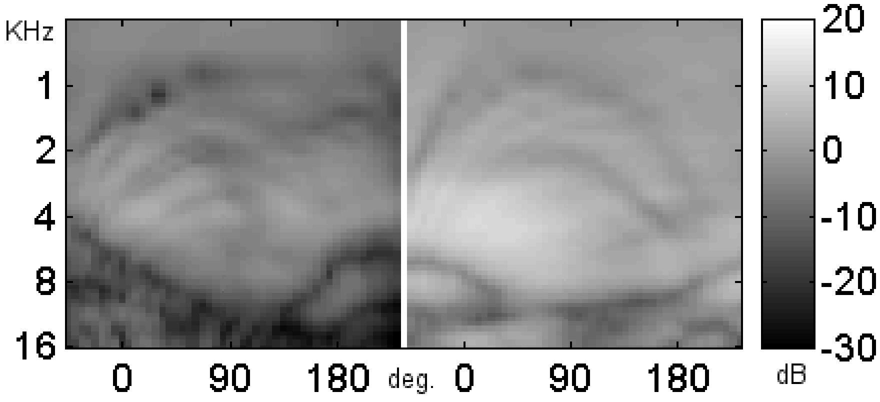

Fig. 1. HRTF slices for contralateral and ipsilateral ears for azimuth of 45 degrees and varying elevation for a human subject.

(for recent review see [27]).

For relatively distant sources the HRTF is a function of source direction and frequency, with a weaker dependence on the distance to the sound source [28], which we neglect. If the sound source is located at azimuth φ and elevation θ in a spherical coordinate system, then the (left and right) HRTFs H_l and H_r are defined as the frequency-dependent ratio of the SPL at the corresponding eardrum $\Phi_{l,r}$ to the free-field SPL at the center of the head as if the listener is absent Φ_f:

$$H_l(\omega, \varphi, \theta) = \frac{\Phi_l(\omega, \varphi, \theta)}{\Phi_f(\omega)}, \quad H_r(\omega, \varphi, \theta) = \frac{\Phi_r(\omega, \varphi, \theta)}{\Phi_f(\omega)}. \tag{1}$$

In the following we will suppress the dependence on the frequency ω. A typical slice of an HRTF is shown in Fig. 1. In the plot, the elevation rises from -45 to 225 degrees along the cone of confusion for the azimuth of 45 degrees. The plot contains several peaks and valleys, which shift as the elevation change. The effects of the different body parts show up in different frequency ranges. Shadowing by the head explains the overall level difference in the two pictures; torso reflections create wide arches in the low frequency area of the plot, and *pinna notches* appear as dark streaks in the high-frequency regions. The locations of these features change with frequency and with elevation. These cues are thought to be very important to our ability to distinguish elevations [29], [30], [31].

III. Environmental modeling

Using the HRTF alone to render the sound scene results in perception of a "flat" environment where the sounds are not well externalized. Users usually report correct perception of azimuth and elevation, but the sound is felt to be excessively close to the head surface. As a source is moved away from the ear, users report that the sound is still in the ear but with decreased volume. To achieve good externalization, we have observed that environmental scattering cues must be incorporated in the simulation of the auditory space. We use either a simple image model for rectangular rooms [32] or a more complicated model for arbitrary piecewise-planar rooms [33]. Multiple reflections create an infinite lattice of virtual sources, whose positions can be found by simple geometric computations and visibility testing. Absorption is accounted for by multiplying virtual source strengths by a heuristic coefficient β for every reflection occurred. (We use $\beta = 0.9$ for walls and 0.7 for the carpeted floor and ceiling). Summing the peaks at time instants $\tau = d/c$, where d is the distance from the ith virtual source, with amplitudes determined by the distance and the source strength, we can compute the room impulse response (IR). It depends upon the relative locations of the source and the receiver in the room. For computing the IR at multiple room points in a rectangular room we presented a fast algorithm based on the multipole method in [34].

IV. Dynamics

In addition to the static localization cues (ITD, ILD and anatomical scattering), humans use dynamic cues to reinforce localization. These arise from active, sometimes unconscious, motions of the listener, which change the relative position of the source [35]. It is reported that front/back confusions which are common in static listening tests disappear when listeners are allowed to slightly turn their heads to help them in localization.

When the sound scene is presented through headphones without compensation for head and body motion, the scene does not change with the user's motion, and dynamic cues are absent. The virtual scene essentially rotates with the user, creating discomfort and preventing externalization. The effect of the source staying at the same place irrespective of the listener's motion is for it to be placed at the one location that stays fixed in the moving coordinate system — the origin of that coordinate system, inside the head. A low latency head position and orientation tracking is necessary so that dynamic cues are recreated, and delay between head motion and resulting changes in audio stream is not distracting.

V. Audio scene rendering algorithms

As described above, it is sufficient to recreate the sound pressure at the eardrums to make a synthetic audio scene indistinguishable from the real one, and the synthesis of the virtual audio scene must include both HRTF-based and environmental cues to achieve accurate simulation. We use a set of real-time sound rendering algorithms described below. The level of detail in the simulation (interpolation quality and number of room reflections traced) is automatically adjusted to match the processing power available.

To synthesize the audio scene given the source location(s) (φ, θ) one needs to filter the signal with the appropriate HRTF(s) $H(\varphi, \theta)$ and render the result binaurally through headphones. To compensate for head motion low-latency head tracking is employed to stabilize the virtual audio scene.[1] Additionally, the HRTF must be interpolated between discrete measurement positions to avoid audible jumps in sound, and appropriate reverberation must be mixed into the rendered signal to create good externalization.

A. Head tracking

We use a Polhemus tracker for head tracking. The tracker provides the position (Cartesian coordinates) and the orientation (Euler angles) of up to 4 receivers with respect to a transmitter. A receiver is mounted on the headphones. The transmitter might be fixed, creating a reference frame, or moved by the user, creating a perception of a moving sound source. Then, positions of virtual sources in the listener's frame of reference are computed by simple geometric inversion, and sources are rendered at their appropriate locations. Tracking latency is approximately 40 ms. Multiple receivers are used to enable multiple people participation; our Polhemus transmitter though has an error-free operation range of only about 1.5 m, limiting the system's spatial extent.

Since the tracker provides the position and orientation data of the receiver with respect to the

[1] Rendering through loudspeakers [36] does not require precise head tracking, but does need additional processing to cancel the crosstalk. Further, the "sweet spot" where correct perception is achieved is quite small ($\sim$ 20 cm), and it is harder to render multiple sources.

transmitter, simple geometric inversion of coordinates must be performed for virtual scene stabilization if the scene is to stay stable with respect to a fixed transmitter. The coordinate system used in this work is a vertical-polar system with X-axis horizontal pointing from the center of the head to the right ear, Z-axis horizontal pointing from the center of the head to the nose, and Y-axis vertical pointing up. In this system, Y is the polar axis, the azimuth $\varphi \in [-\pi, \pi]$, the elevation $\theta \in [-\pi/2, \pi/2]$ and a point position on a unit sphere is given by

$$x = \sin\varphi\cos\theta,\ y = \sin\theta,\ z = \cos\varphi\cos\theta.$$

For rendering, the coordinates (X', Y', Z') of a virtual sound source in a coordinate system bound to a receiver (which is mounted on the headphones) is necessary. The information obtained from the head tracker includes the coordinates (X_r, Y_r, Z_r) and the orientation (φ, θ, ψ) (azimuth, elevation and roll, respectively) of the receiver with respect to transmitter. The rotation matrix R then can be written as

$$R = \begin{bmatrix} \cos\varphi\cos\theta & \cos\varphi\sin\theta\sin\psi - \sin\varphi\cos\psi & \cos\varphi\sin\theta\cos\psi + \sin\varphi\sin\psi \\ \sin\varphi\cos\theta & \sin\varphi\sin\theta\sin\psi + \cos\varphi\cos\psi & \sin\varphi\sin\theta\cos\psi - \cos\varphi\sin\psi \\ -\sin\theta & \cos\theta\sin\psi & \cos\theta\cos\psi \end{bmatrix}.$$

Given the position of the virtual sound source (X, Y, Z) in the transmitter coordinate system, the vector of (X', Y', Z') can be simply found as

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = R^{-1} \begin{bmatrix} X - X_r \\ Y - Y_r \\ Z - Z_r \end{bmatrix},$$

and the source view angles (φ', θ') (azimuth and elevation) in the receiver-bound coordinate system are simply

$$\theta' = \arcsin(\frac{Y'}{R'}), \varphi' = \arcsin(\frac{X'}{R'\cos\theta'}), R'^2 = X'^2 + Y'^2 + Z'^2.$$

Once the direction of arrival is computed, the corresponding HRTF is retrieved or interpolation between closest measured HRTFs is performed.

B. HRTF interpolation

Currently, we use pre-measured sets of HRTFs.[2] As measured, the HRTF corresponds to a sounds generated 1 meter away and sampled at a sphere at a fixed angular resolution. The measurements are performed at discrete points and have to be interpolated to avoid audible sudden changes in sound spectrum when the source position changes. The spectrum changes are very noticeable if white noise is used and the HRTF for the closest measured point is used instead of interpolation.

Several papers report on different possible interpolation methods. Assume that the source is located at a azimuth and elevation of (φ, θ) and the N closest available measurements of HRTF are at (φ_i, θ_i). The resulting interpolated HRTF $\tilde{H}(\omega)$ should be computed as a weighted average of those N HRTFs $H_i(\omega)$ with weights w_i which sum up to one, and ultimately the impulse response (IR) corresponding to $\tilde{H}(\omega)$ is required. Note that the HRTF at a given frequency is simply a complex number. Simple interpolation of a real and an imaginary parts separately is flawed.[3] The paper [38] suggests the geometric interpolation as a way to properly interpolate complex valued frequency response functions. We use instead arithmetic interpolation of the amplitude and the phase separately, which gives the same result for the interpolated phase as the geometric interpolation. However, the problem of phase uncertainty still exists since the phase of a transfer function can be defined only within a multiple of 2π, which introduces phase unwrapping errors on the interpolated value of the phase. The phase uncertainty will not arise if the spatial sampling frequency for the HRTF is fine enough as determined by a Nyquist criterion. Thus for a fixed sampling grid, HRTF interpolation at lower frequencies (longer wavelengths) will not be affected by this problem. This is confirmed in our experiments; major lower-frequency content of the acoustic signal is perceived at the correct place, but phantom sources containing mostly high-frequency components appear in various places, often inside the head.

Thus, the magnitude part of $\tilde{H}(\omega)$ can be constructed uniquely by interpolating amplitudes of

[2] Another related project deals with numerical synthesis of HRTF from ear meshes and, once completed, will eliminate the need for tedious HRTF measurements [37].

[3] Consider interpolation of two complex numbers with the same amplitude and different phases. If the real and the imaginary part are interpolated separately, the result of interpolation will have smaller amplitude than the original vectors – while it is obvious that the result of the correct interpolation should have the same amplitude and intermediate phase.

$H_i(\omega)$, but the phase reconstruction is problematic. However, in practice, it is not really necessary to preserve phase information in the interpolated HRTF, as humans are sensitive mostly to the magnitude spectrum for the localization purposes [39] and the measured phase is likely to be contaminated anyway due to difficulties of measuring it accurately because of sampling and other problems. The phase is necessary to reconstruct the ITD, but given the correct ITD, only the frequency dependence of the magnitude matters for perception. If the phase information is lost, the resulting response is minimum-phase (the highest peak occurs at zero time, and the response decays quite rapidly, which means that the ITD can be accounted for by a simple time shift). The ITD can be approximated using the Woodworth's formula [40]

$$\hat{\tau} = r(\varphi + \cos\varphi)\cos\theta/c, \tag{2}$$

where c is the sound speed. The only unknown value here is the head radius r which can be customized for the particular user using video, as is described below.

The database of HRTFs used in our work measures the directional transfer functions on a lattice with 10 degree step in azimuth and elevation for different people. To compute interpolated HRIRs for a source at $P = (\varphi, \theta)$, we find the three closest lattice points $P_i = (\varphi_i, \theta_i), i = 1...3$, with corresponding distances d_i between P and P_i.[4] Then, if the HRTF at point P_i is represented by $H_i = A_i(\omega)e^{-i\varphi_i(\omega)}$, the interpolated HRTF magnitude is

$$\tilde{A}(\omega) = \frac{\sum w_i A_i(\omega)}{\sum w_i},$$

with weights $w_i = 1/d_i$ (w_i is bounded from above by some constant $C = 100$ to prevent numerical instabilities). Then, the phase $\hat{\varphi}(\omega)$ of the interpolated HRTF corresponding to the leading ear is set to the $\hat{\tau}/2$:

$$\tilde{\varphi}(\omega) = \omega\hat{\tau}/2,$$

and the phase for the lagging ear is set to the $-\hat{\tau}/2$ in the similar way. (Time shifts are performed in frequency domain since humans are sensitive to ITD variations as small as 7 μs [41] which is $1/3$ of a sampling period at the rendering rate of 44.1 kHz). The resulting HRTF $\tilde{H}(\omega) = \tilde{A}(\omega)e^{-i\tilde{\varphi}(\omega)}$

[4] The distance between lattice points is defined as a Euclidean distance between the points with corresponding azimuth and elevation placed on the unit sphere.

is the desired interpolation. The inverse Fourier transform of $\tilde{H}(\omega)$ provides the desired interpolated head-related impulse response (HRIR) which can be directly used for convolution with the sound source waveform. The perceived sound motion is quite smooth, and no jumps or clicks are noticeable.

It is also desirable to find the closest lattice points quickly (as opposed to finding the distances from P to all lattice points). A fast search for the three nearest points P_i in a lattice is performed using a lookup table. The lookup table is a 360-by-180 table covering all possible integer values of azimuth and elevation. The cell (i, j) in the table stores n identifiers of lattice points that are closest to the point with azimuth $i + 0.5$ and elevation $j + 0.5$. To find the closest points to the point P, only the n points referred to by a cell corresponding to the integer parts of P's azimuth and elevation are checked. It is clear that for a regular lattice some small value of n is sufficient to always obtain the correct closest points. We use $n = 5$ which is practically errorless (in over 99.95% cases the closest three points are found correctly in random trials) which significantly improves the performance of the on-line renderer compared to a brute-force search.

C. Incorporation of the room model

The room impulse response (RIR) can be written for rectangular rooms using a simple image model [32]. A more complex image model with visibility criteria [33], that is somewhat more heuristic, can be applied for the case of more general rooms. The RIR is a function of both the source and receiver locations, and as the listener's position relative to the source changes, so does the RIR.

The RIR from the image model has a small number of relatively strong image sources from the early reflections, and very large numbers (tens of thousands) of later weaker sources. The earlier reflections will in turn be scattered by the listener's anatomy. Thus they must be convolved with the appropriate HRIR for the direction of an image source. (For example, the first reflection is usually the one from the floor, and should be perceived as such). The large number of image sources presents first a problem of evaluating the RIR, and the length of the RIR presents a problem for low-latency rendering, since convolution with long filters may introduce latencies. We present below a solution to this problem, based on a decomposition of the RIR.

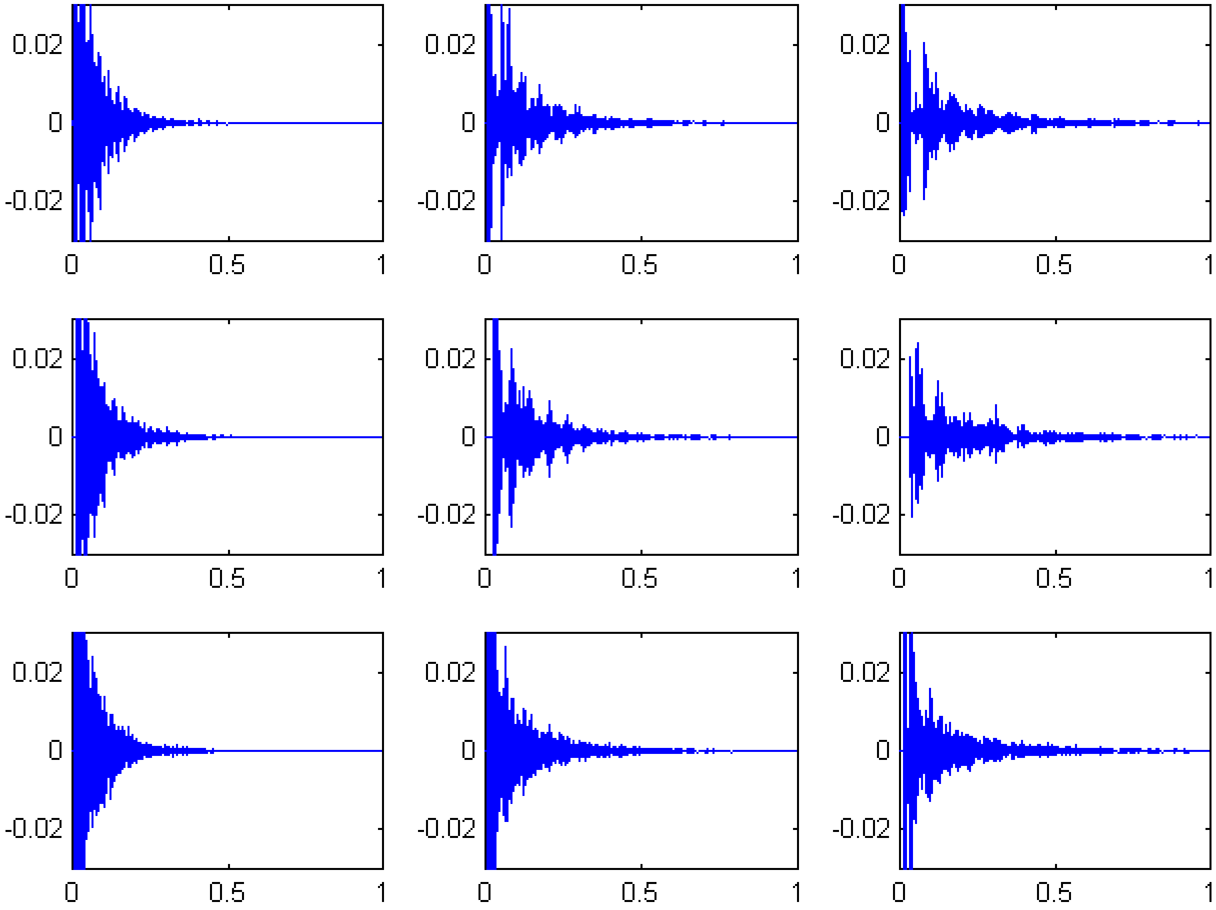

Fig. 2. Sample room impulse response functions. Column 1: Three different arrangements of source and receiver within a room of fixed size. Column 2: The source and the receiver are placed in the same positions as for column 1, but the scene is scaled up by a factor of 2. Column 3: same as column 2, with the scaling factor of 3.

Room impulse responses for three rooms of different sizes and different positions of the source and the receiver are shown in Figure 2 (the plots are intentionally clipped along the Y-axis so that details of the reverberation tail are visible). The room size stays constant along a column, while the relative positions of the source and the receiver stays the same along the rows. Three random positions of the source and the receiver were selected for plots in the first column. The second column is the same three source-receiver arrangements within a room which is twice as large as for the first column (i.e., the whole scene is scaled up by a factor of two), and the third column is for the room which is three times larger.

These plots illustrate that the reverberation time of the room (formally defined as the time it takes for the sound level to drop by 60 dB) and the decay rate of the reverberation tail changes with room geometry – the reverberation decays slower in bigger room. Obviously, the decay rate depends also on the reflective properties of the room walls. [5] However, it can be seen that this rate doesn't depend on the position of the source and the receiver within a room, which can be expected since the reverberation tail consists of a mixture of weak late reflections and is essentially directionless. We performed Monte-Carlo experiments with random positions of the source and the receiver and found that for a given room size, the variance in the reverberation time is less that 20 percent. This observation is useful because humans perceive the size and the properties of the room using primarily the reverberation time for different frequencies. Independence of the tail decay rate of the positions of the source and the receiver enables us to avoid expensive recomputation of the tail with every source and receiver motion.

However, the positions of early reflections do change significantly when the source or the receiver is moved. It is believed ([43], [44]) that at least the first few reflections provide additional information that help in sound localization. Full recomputation of the IR is not feasible in real-time; still, some initial part of the room response must be reconstructed on the fly to accommodate changes in the positions of the early reflections. We adopt an approach where the direct path arrival and the first few reflection components of IR are recomputed in real time and the rest of the filter is computed once for a given room geometry and materials. As can be seen from Figure 2, the reverberant tail of room response function stays relatively the same for different source and receiver locations in the same room, which justifies our approach. (If a virtual world consists of several rooms, the reverberation tail for each of those can be pre-computed and appropriate switch be made when participant moves from one room to another).

D. Rendering filter computation

As described before, we construct in real-time the finite-impulse-response (FIR) filter H that consists of a mix of appropriately delayed individual impulse responses corresponding to the signal arrivals from the virtual source and its images created by reflections. The substantial length of

[5] The Sabine equation relates the reverberation time T of a room to be proportional to the room volume V and inversely proportional to its total sound absorption A, $T \sim V/A$. [42]

the filter H (which contains the direct arrival and room reverberation tail) results in delays due to convolution. For accurate simulation of the room response, the length of H must be not less than the room reverberation time, which ranges for real rooms from 400 ms (typical office environment) to 2s and more (concert halls). If the convolution is carried out in the time domain, the processing lag is essentially zero, but due to high computational complexity of time-domain convolution only a very short filter can be used if processing is to be done in real-time, hindering reverberation rendering abilities. The frequency-domain processing using fast Fourier transforms are much faster, but the blocky nature of the convolution causes latency of at least one block. A nonuniform block partitioned convolution algorithm was proposed in [45], but this algorithm is claimed to be proprietary, and is somewhat inefficient and difficult to optimize on regular hardware. We instead use frequency-domain convolution with short data blocks ($N_1 = 2048$ or 4096 samples) which results in tolerable delays of 50 to 100 milliseconds (at a sampling rate of 44.1 kHz). We split the filter H into two pieces H_1 and H_2; H_1 has length N_1 (same as data block length) and is recomputed in real-time. However, processing only with this filter will limit the reverberation time to the filter length. The second part of the filter, H_2, is much longer ($N_2 = 65536$ samples) and is used for the simulation of reverberation. This filter contains only the constant reverberant tail of the room response, and the part from 0 to N_1 in it is zeroed out.

By splitting the convolution in two parts and exploiting the fact that the filter H_2 is constant in our approximation, we are able to convolve the incoming data block X of the length N_1 with the combined filter H of length $N_2 \gg N_1$ with delays only of order N_1 (as opposed to unacceptable delay of order N_2 if a non-partitioned convolution is used). This is due to the linearity of convolution with allows us to split the filter impulse response into blocks of different sizes, compute the convolution of each block with the input signal, and sum appropriately delayed results to obtain the output signal. In our example, the combined filter $H = H_1 + H_2$ (since the samples from 0 to N_1 in H_2 is zeroed out) and no delays are necessary.

Mathematically, the (continuous) input data stream $X = \{x(1), x(2), ..., x(n), ...)$ is convolved with the filter $H = \{h(1), , h(2), ...h(N_2))$ to produce the output data stream $Y = \{y(1), y(2),$

$..., y(n), ...)$. The convolution is defined as

$$y(n) = \sum_{k=1}^{N_2} x(n-k)h(k),$$

and we break the sum into two parts of lengths $N_1$and $N_2 - N_1$ as

$$y(n) = \sum_{k=1}^{N_1} x(n-k)h_1(k) + \sum_{k=N_1+1}^{N_2} x(n-k)h_2(k).$$

The second sum can be also taken from 0 to N_2 with $h_2(1), h_2(2), ..., h_2(N_1)$ set to zero. The filter H_1 is resynthesized in real-time to account for the source and receiver relative motion. The filter H_2 contains the fixed reverberant tail of the room response. The first part of the sum is recomputed in real time using fast Fourier transform routines with appropriate padding of the input block and the FIR filter to prevent wrap-around and ensure continuity of the output data. The delay introduced by the inherent buffering of the frequency-domain convolution is limited to $2N_1$ at worst, which is acceptable. The second part of the sum (which is essentially the late reverberation part of a given signal) operates with a fixed filter H_2 and for a given source signal is simply precomputed off-line. If the source signal is also obtained on-line, it must be delayed sufficiently to allow reverb precomputation before the actual output starts, but once the reverb is precomputed the reaction of the system to the user's head motion is fast because only the frequency-domain convolution with short H_1 (which changes on-the-fly to accommodate changes in user position) is done on-line. In this way, both the low-latency real-time execution constraint and the long reverberation time constraint are met without resorting to the slow time-domain convolution.

The algorithm for real-time recomputation of H_1 proceeds as follows. The filter H_1 is again separated into two parts. The first part contains the direct path arrival and first reflections (up to reflections of order L_1 – where L_1 is chosen by the constraint of real time execution). This part is recomputed in real time to respond to the user or source motion. The second part consists of all the reflections from order L_1 to the end of the filter. This second part is precomputed at the start of the program for a given room geometry, and some fixed location of source and receiver. Once the new coordinates of the source and the receiver are known, the algorithm recomputes the first part of FIR filter and sticks it on top of the second part. Figure 3 shows the process of composition for two different arrangements of the source and the receiver and $L_1 = 4$. The composition starts

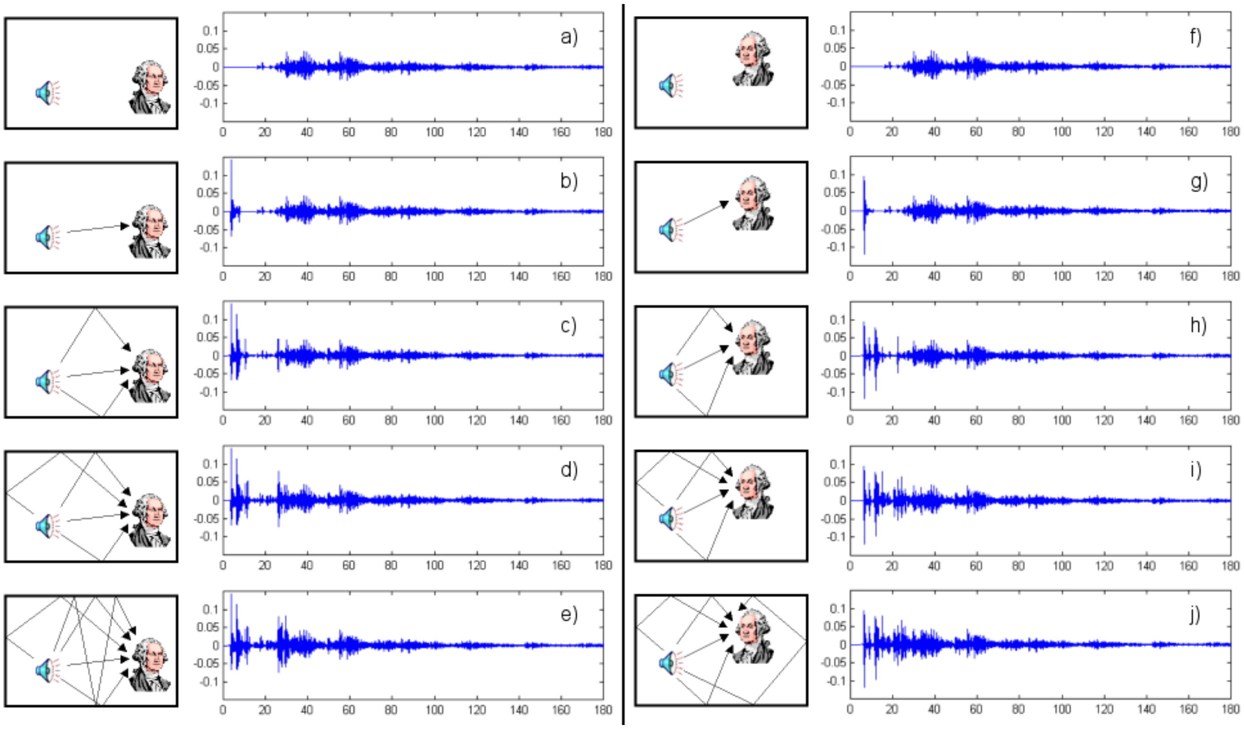

Fig. 3. Synthesis of a rendering FIR filter in real time. a) Precomputed tail of the filter (reflections of order 4 and above). b)-e) Addition of reflections of order 0, 1, 2 and 3, respectively. f) Same as a) but for different position and orientation of the receiver. g)-j) Same as b)-e) above.

with the precomputed tail of IR which stays constant independent of the source and the receiver positions, and in four shown steps adds the direct arrival component and reflections of order 1, 2 and 3 to the IR. It is interesting to note that some reflections of small order may come later than some reflections of larger order because of the room geometry, so the fixed part does overlap with the part that is recomputed on the fly. When new H_1 is available, it is used to filter the new incoming blocks of input data, and the precomputed result of convolution with H_2 is added to the result of convolution with H_1 to form the playback stream.

E. Playback synthesis

The computations described above can be performed in parallel for multiple virtual sound sources at different positions. In a rendering cycle, the source signals are convolved with their appropriate FIR filters. The convolution is done in the frequency domain. The convolved streams are mixed together for playback. A separate thread computes the reverberation tail, which is easier, since all streams share the same precomputed reverberation FIR filter. The streams are first mixed together and then the reverberation filter is applied, also in the frequency domain. The result of this convolution is mixed into the playback. The playback is performed through standard operating system calls, which is the source of small additional system latency.

VI. Customizing the HRTF

The biggest and still-open problem in the synthesis of the virtual auditory spaces is the customization of the HRTF for a particular individual. The HRTF complexity is due to the complex shapes of the pinna, which lead to several resonances and antiresonances. Each person presumably learns her own HRTF given auditory or visual feedback about the source position, but the HRTFs of different people look very different and, not surprisingly, are not interchangeable. In order to accurately simulate the pressure waveforms that a listener would hear in the real world, HRTFs must be separately determined for each individual (e.g., see [25], [46]). The problem of HRTF customization is currently a subject of a open research. The usual customization method is a direct measurement of HRTF when a tiny microphone is placed in the ear canal of the subject and a sound is played through a loudspeaker positioned sequentially over all possible direction of arrival (DOA) angles in some given steps. In the database used by us for matching [47], [48] a resolution of 5 degrees is used over the whole sphere (except for the lower part, which cannot be measured in the experiments usually). This method is accurate but highly time-consuming, and there are different measurement issues complicating the procedure [49]. There also exist alternative approaches such as allowing participant to manipulate different characteristics of HRTF set used for rendering until she achieves satisfactory experience (see, e.g., [50]), though it is not clear if the correct HRTF is achieved. A novel and promising approach is the direct computation of the HRTF using three-dimensional ear mesh obtained via computer vision and solving the physical wave propagation equation in the presence of a non-rigid boundary by fast numerical methods [37]. However this work is still under development, and current virtual auditory systems do not have yet any methods for customization of the HRTF. In this paper we seek to customize the HRTF using a database containing the measured HRTFs for 43 subjects along with some anthropometric measurements [47], [48].

A. Approaches to Customization

Since the HRTF is the representation of the physical process of the interaction between the oncoming sound wave and the listener's pinnae, head and torso, it is natural to make the hypothesis that the structure of the HRTF is related to body scattering part dimensions and orientation. For example, observe that if the ear is scaled up the HRTF will maintain the shape but will be shifted toward the lower frequencies on the frequency axis. Since the listener presumably deduces the source elevation from the positions of peaks and notches in the oncoming sound spectrum, usage of the HRTF from the scaled-up ear will result in systematic bias in the elevation estimation. Some studies, such as functional representation of HRTFs using spatial feature extraction and regularization model [51], a structural model for composition and decomposition of HRTF [52], and especially experiments with HRTF scaling ([53], [54], [55]) already suggested that the hypothesis is somewhat valid, although a perfect localization (equivalent to the localization with the person's own HRTF) was not achieved with other people's HRTFs appropriately scaled up or down. However, the ears of different persons are different in much more ways than just a simple scaling, and a seemingly insignificant small change in ear structure can cause dramatic changes in HRTF.

B. Database Matching

An intermediate approach which we use in our system is an attempt to select the best-matching HRTF from an existing database of HRTFs and use it for the synthesis of the virtual audio scene, thus making the HRTF semi-personalized.

Thus, the problem is to select the most appropriate HRTF from a database of HRTFs indexed in some way. The database we used was recently released by the UC Davis CIPIC laboratory and contains the measurement of the HRTFs of 43 people, along with some anthropometric information about the subjects. The HRTFs are measured on a spherical lattice using a speaker positioned 1 meter away from the subject. HRTF measurements below -45 degrees of elevation

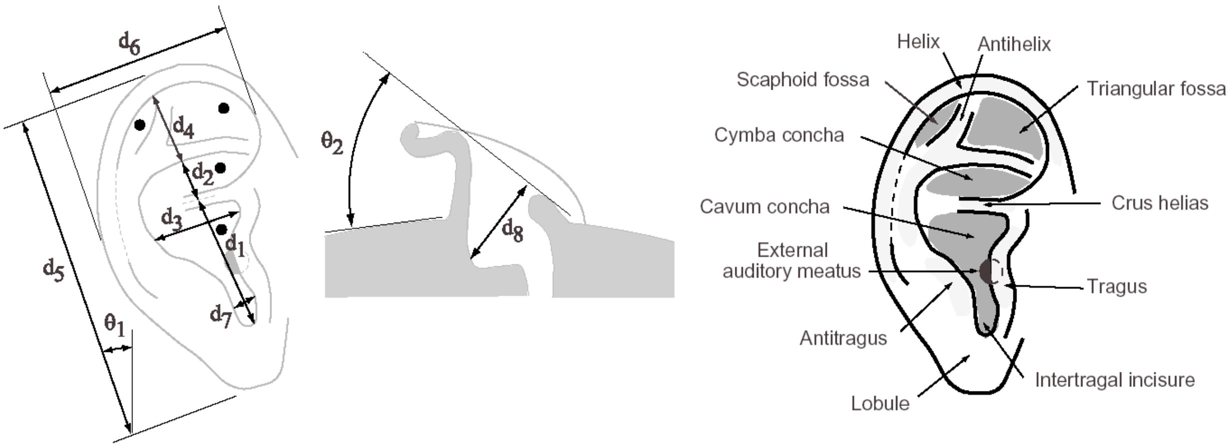

Fig. 4. The set of measurements provided with the HRTF database.

are not available (the speaker can't be placed that low since it would hit the person's legs). [6] The anthropometric information in the database consists of 27 measurements per subject – 17 for the head and the torso and 10 for the pinna. Pinna parameters are summarized in the Figure 4 and are as follows: $d_1 ... d_8$ are cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, intertragal incisure width and cavum concha depth, and the θ_1 and θ_2 are pinna rotation and flare angles, respectively. For the HRTF matching procedure, we use 7 of these 10 pinna parameters which can be easily measured from the ear picture.

We perform an exploratory study on the hypothesis that the HRTF structure is related to the ear parameters. Specifically, given the database of the HRTFs of 43 persons along with their ear measurements we select the closest match to the new person by taking the picture of her ear, measuring the d_i parameters from the image and finding the best match in the database. If the measured value of the parameter is $\hat{d}_i$, the database value is d_i and the variance of the parameter in the database is $Var(d_i)$, then the error for this parameter $e_i = (\hat{d}_i - d_i)/Var(d_i)$, the total error $E = \sum_i e_i^2$ and the subject that minimizes the total error E is selected as the closest match. Matching is performed separately for the left and the right ears, which sometimes leads to the selection of left and right HRTFs belonging to two different database subjects; these cases are rare

[6] It is possible that absence of low-elevation measurements has negative impact on the VAS synthesis because the first reflection usually comes from the floor (ie. low elevation) and might be not rendered correctly (a closest available HRTF will be used for rendering instead).

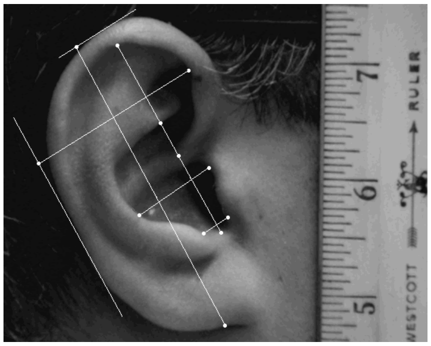

Fig. 5. Sample picture of the ear with the measurement control points marked.

though.

We have developed a simple interface which allows to perform almost on-the-fly selection of the best-matching HRTF from the database. The picture of the left and the right ear of the new virtual audio system user is taken with two cameras, with the user holding the ruler in the frame to provide scale of reference. Sample picture used in one of the sessions of HRTF matching is shown in the Figure 5. Once the picture is taken, an operator identifies key points on the ear and measures the ear parameters described above. The user interface enforces certain constraints on the measurements (for example, d_1, d_2 and d_4 should lie on the same straight line which is the ear axis, d_3 and d_7 should be perpendicular to the ear axis and the bounding rectangle formed by d_5 and d_6 is axis-aligned). The parameters d_8 and θ_2 are not measured since they can't be reliably estimated from pictures and θ_1 is not used for the matching, but is used to compensate for the difference between pinna rotation angles of the system user and the selected best-matching subject. The matching is done in less than a minute, and no extended listening tests have to be performed for customization – only the ear picture is required, which is a significant advantage of our method.

VII. Experimental Results

A number of volunteers (~50) were subjects of some informal listening experiments, in which a virtual source was generated at the location of the transmitter of the Polhemus tracker (a small cube of side 4 cm). Generally, people reported achieving very good externalization. Reported experience varies from "I can truly believe that this box is making sound" to "Sound is definitely outside my head, but my elevation perception is distorted" (probably due to non-personalized

HRTFs). Thus, the system was capable of making people think that the sound was coming from the external source, even though it was being rendered at the headphones. Presumably, correct ITD cues, reverberation cues and highly natural changes of the audio scene with head motion and rotation create this realistic perception, along with the non-personal HRTF cues are responsible for these reports. The stability of the synthesized virtual audio scene is also remarkable and latency is noticeable only if user rotates her head or moves the source in a very fast, jerky motion. Even better results should be achievable with personalized HRTFs. We will soon have a mechanism to compute personalized HRTFs using video and numerical analysis [37].

A. System Setup and Performance

The current setup used for experiments is based on a high-end office computer which is dual Xeon P4-1.7 GHz Dell Precision 530 PC with Windows 2000, with the tracker connected to the serial port. One receiver is fixed providing a reference frame, and another is mounted on the headphones. The setup also includes stereo head-mounted display Sony LDI-D100B which is used for immersive virtual environment in the developed software. The programming is done in Microsoft Visual C++ 6.0, using OpenGL for video. Computations are parallelized for multiple sources and for left and right playback channels, which results in good efficiency. The number of recomputed reflections is adjusted on the fly to be completed within the head-tracker latency period. For one source, up to five levels of reflection can be recomputed in real time. The algorithm can easily handle up to 16 sources with two levels of reflections, doing video rendering in parallel.

B. Non-personalized HRTF set

We performed small-scale formal tests of the system on six people. The test sounds are presented through headphones, and the head tracker measures the head position when the subject “points” to the virtual sound source. The sounds used for the tests were three 75ms bursts of white noise with 75ms pauses between them, repeated every second. As a “generic” HRTF set, we used HRTFs that were measured from a real person in an anechoic chamber. This person was not a test subject.

The test sessions were fairly short and involved calibration, training and measurement. For calibration, subjects were asked to look at the source placed at a known spatial location (coinciding with the tracker transmitter) and the position of the sensor on the subject’s head was adjusted to read 0 degrees of azimuth and elevation. Then, the sound was presented at random position, with $\varphi \in [-90°, 90°], \theta \in [-45°, 45°]$. Subjects were asked to "look" at the virtual source in the same way that they looked at the source during calibration (e.g., point with their forehead). For training feedback, the program constantly outputs the current bearing of the virtual source; perfect pointing would correspond to $\varphi = 0, \theta = 0$. During test sessions, 20 random positions are presented. The subject points at the perceived sound location and on localization hits the button. The localization error is recorded and the next source is presented. Results are summarized in the Table 1 below.

Table 1

	s1	s2	s3	s4	s5	s6
avg $\lvert\varphi\rvert$	6.3	5.1	4.3	6.4	8.0	8.4
avg $\lvert\theta\rvert$	9.0	9.5	5.5	16.7	14.4	7.2
avg φ	-5.3	4.8	2.7	3.3	4.2	-5.7
avg θ	-4.0	-4.5	5.0	-9.0	-8.3	3.8

The results for the "generic" HRTF set are interesting. Some subjects perform better than the others, and localization in azimuth is generally better than in elevation. In addition, for all subjects bias accounts for at least half of the error, and may be removed with a better pointing mechanism, For subjects 2, 3 and 4 bias accounts for almost all of the localization error in azimuth. Considering elevational localization (which is believed to be hampered most by using of non-individualized HRTF), subject 3 performs quite good; performance of subjects 1, 2 and 6 is close to the average and subjects 4 and 5 perform poorly, but azimuthal localization is still better that elevation. Errors are probably due to non-individualized HRTFs.

The results show that the localization with non-individualized HRTF tends to introduce significant errors in elevation, either by "shifting" the perceptual source position up or down or by disrupting the vertical spatialization more dramatically. Still, the elevation perception is consistent and the source can be perceived as being "above" or "below". The azimuth perception remains reasonable since it depends primarily on the ITD/ILD cues. Overall, the system is shown to be able to create convincing and highly accurate virtual auditory displays. With the personalized HRTFs, the same degree of accuracy is expected for every user.

C. Personalized HRTF set

We performed a second set of tests to verify whether the customization has a significant effect on the localization performance and the subjective experience of the virtual audio system user. For this set, the best-matching HRTF was selected from the database and used for virtual audio scene rendering. The test sessions themselves was conducted in the same manner as in the first set, with the same test sound used.

Table 2

	s1	s2	s3	s4	s5	s6
avg $\|\varphi\|$	13.5	5.9	7.5	13.4	10.2	7.6
avg $\|\theta\|$	7.6	7.2	4.4	12.9	13.6	12.5
avg φ	-9.4	-3.3	-4.8	-3.1	-1.5	0.7
avg θ	-1.4	-7.0	-2.0	4.8	4.8	-6.3

Table 2 results are for the case of the best-matching HRTF from the HRTF database. Azimuthal localization was not the priority task for the subjects for this test. It is clear that the elevation localization performance is improved consistently by 20-30% for 4 out of 6 subjects, although it would take a larger number of trials to be sure that a reduction in elevation error is statistically significant. We are currently working on full-scale set of experiments to confirm the statistical significance of these results. Improvement for the subject 5 is marginal and subject 6 performs actually worse with the customized HRTF.

The objective performance criteria agrees with the subjective performance estimated by subjects themselves. Subjects 1 through 4 reported that they are able to better feel the sound source motion in the median plane and the virtual auditory scene synthesized with personalized HRTF sounds better (better externalization and better perception of DOA and source distance is achieved). Subject 5 reported that motion can not be perceived reliably both with generic and customized HRTF, which agrees with experimental data (It was later discovered that the subject 5 has tinnitus – “ringing” in the ears). Subject 6 also reports that the generic HRTF just “sounds better”.

Overall, it can be said that the customization based on visual matching of ear parameters can provide significant enhancement for the users of the virtual auditory space. This is confirmed

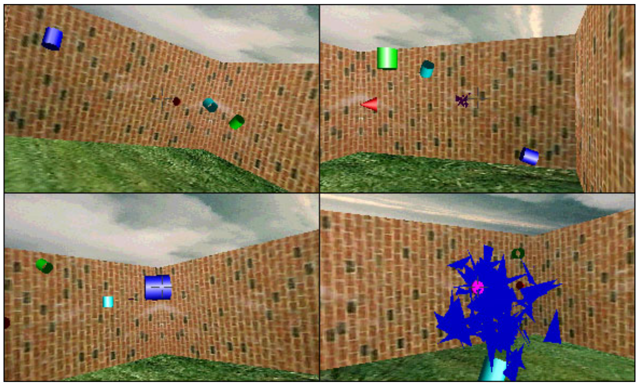

Fig. 6. Sample screenshots from the developed audio-video game with spatialized audio user interaface.

both by objective measures, where the localization performance increases by 30% for some of the subjects (the average gain is about 15%), and by subjective reports, where the listener is able to distinguish between HRTFs that "fits" better or worse. These two measures correlate well, and if the customized HRTF does not "sound" good for a user a switch back to the generic HRTF can be made easily. The performed customization is a coarse "nearest-neighbor" approach, and the HRTF certainly depends on much more than the 7 parameters measured. Still, even with such a limited parameter space the approach is shown to achieve good performance gain, and combined with the audio algorithms presented, should allow for creation of realistic virtual auditory spaces.

D. Application example

An alternative way to evaluate the benefits of the auditory display is by looking at user's informal reports of their experience with an application. We did this by developing a simple game with spatialized sound, personalized stereoscopic visual display, head tracking and multi-player capabilities all combined together. In the game, the participant wears stereo glasses, headphones and a Polhemus tracker. Participants are immersed in the virtual world and are free to move. The head position and orientation of the players are tracked, and appropriate panning of the video scene takes place. The rendered world stays stable in both video and audio modalities. The video stream is rendered using standard OpenGL.

In the game, the participant is piloting a small ship and can fly in a simulated room. The participant learns an intuitive set of commands that are controlled by his head motion like in an airplane simulator game. (For example, head bowing forward is interpreted as a command to accelerate forward, while the opposite motion slows down the ship; rolling the head to the left or right makes the ship turn in the corresponding direction). Multi-player capability is implemented using a client-server model, when the state of the game is maintained on one computer in a game server program which keeps and updates the game state (object positions, ship positions, collision detection etc.) periodically. Information required for game scene rendering (positions and video/audio attributes of objects) is sent by the server after each update to the video and audio client programs which do corresponding rendering. Clients in turn send back to the server any input received from the keyboard or the tracking unit so that the server can process the input (e.g. spawn a missile object in response to a fire key pressed on the client). Several PCs linked together via Ethernet participate in the rendering of the audio and video streams for the players.

During the course of game, players navigate a virtual world and hit targets, either cooperatively or competitively. In cooperative mode, target changes color and sound when hit by one player, and breaks when another one hits it within 20 seconds of the first hit, so active cooperation of participants is necessary for successful scoring. In competitive mode, the target just breaks when it's hit. Four sample screenshots from the game are shown in Figure 6. Three colored cylindrical objects that can be seen in the field of view are the game targets; they are playing different sounds – music, speech and noise bursts, respectively, and their intensities and spatial positions agree with current position of the player in the world. On the fourth screenshot, one of them gets destroyed and the corresponding sound ceases. The colored cone in one of the screenshots corresponds to the ship of the second participant.

An alternative implementation of the game is an interactive news reader installation when three cubes that simulate the TV screens are floating around, and each cube is broadcasting some randomly selected audio stream from some news site on the World Wide Web. The listener can listen to some or all of them, and select their favorite one by get into its closer proximity for selective listening, or shoot and break some cubes if they doesn't like the news being broadcasted by them, in which case new cubes emerge later on connected to new live audio streams.

The audio modality supports the video modality in these applications since some targets appear only for short periods of time and manifest their presence by playing different sound content. In this way, the audio significantly extends the user's field of regard because, often, a new target is initially localized by the sound it makes. It is also possible to play the game using the audio modality alone, by aiming at the target exclusively by the sound it makes, although it's much harder to hit targets in this way. Slowing down the targets significantly and making the projectile bigger is necessary to get satisfactory performance and user satisfaction in this case, compared to the case when video is available. An interesting audio-only strategy invented by one participant is firing a test missiles and listening to the sound of missile splashing against the wall – and determining whether it's to the left or to the right, above or below the sound made by target.

We performed several informal experiments with both generic HRTFs and personalized HRTFs using our video feature matching personalization algorithm. Participants report that the acoustical presence of several externalized objects is very convincing and it is possible to match the acoustical stream with the video image of the target. The general drawback of the interface noted by several participants is that spatial matching of video and audio is not quite easy because the visual objects are physically on the plane in front of a user (screen or personal LCD glasses) while their acoustical counterparts are much more externalized and are floating in the space surrounding the user, resulting in misalignment between line of sight and direction of sound arrival. This problem can be solved by using a panoramic video display. However, participants usually are able to associate sounds with targets and keep following the sound they are most interested in. The game experience (externalization, localization, navigation, targeting and sound following ability) are generally improved by using semi-personalized HRTFs instead of a generic HRTF set. Customization is especially helpful in achieving good externalization in the front, when the sound often tends to jump inside the head or to the back hemisphere of the listener when a generic HRTF set is used. With a personalized HRTF set, participants report that the sounds of missile splash and target breaking really happen in front and seem to come from the speakers installed next to the PC display, while the true playback is of course happening through the headphones.

VIII. Conclusions and future work

We have presented a set of algorithms for creating a virtual auditory space rendering systems. These algorithms were used to create a prototype system that runs in real-time on a typical office

PC. Static, dynamic and environmental sound localization cues are accurately reconstructed with no noticeable latency, creating highly convincing experience for participants. The system is in use and will be the basis for several user interface projects for sighted and low-vision users.

Acknowledgments

We would like to gratefully acknowledge the support of NSF award 0086075. The authors would like to thank Prof. Richard O. Duda (San Jose State University), Prof. V. Ralph Algazi (University of California, Davis) and Prof. Shihab A. Shamma (University of Maryland) for many illuminating discussions on spatial audio. We would also like to thank Mr. Vikas Raykar and Mr. Ankur Mohan, of our laboratory, for assistance with some of the experiments and software. Finally, thanks are due to the volunteers who participated in the tests of the developed auditory display and game.

References

[1] Y. Bellik (1997). "Media integration in multimodal interfaces", Proc. IEEE First Workshop on Multimedia Signal Processing, Princeton, NJ, pp. 31-36.

[2] R. L. McKinley and M. A. Ericson (1997). "Flight demonstration of a 3-D auditory display", in Binaural and Spatial Hearing in Real and Virtual Environments, ed. by R. H. Gilkey and T. R. Anderson, Lawrence Earlbaum Associates, Mahwah, NJ, pp. 683-699.

[3] M Casey, W. G. Gardner, and S. Basu (1995). "Vision steered beamforming and transaural rendering for the artificial life interactive video environment", Proc. 99th AES Convention, New York, NY, pp. 1-23.

[4] S. A. Brewster (1998). "Using nonspeech sounds to provide navigation cues", ACM Transactions on Computer-Human Interaction (TOCHI), vol. 5, no. 3, pp. 224-259.

[5] J. M. Loomis, R. G. Golledge, and R. L. Klatzky (1998). "Navigation system for the blind: Auditory display modes and guidance", Presence, vol. 7, no. 2, pp. 193-203.

[6] G. Kramer et al. (1997). "Sonification report: Status of the field and research agenda", Prepared for the NSF by members of the ICAD. (Available on the World Wide Web at http://www.icad.org/websiteV2.0/References/nsf.html).

[7] S. Bly (1994). "Multivariate data mappings". In Auditory display: Sonification, audification and auditory interfaces, G. Kramer, ed. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XVIII, pp. 405-416. Reading, MA, Addison Wesley.

[8] A. S. Bregman (1994). "Auditory scene analysis: The perceptual organization of sound", MIT Press, Cambridge, MA.

[9] J. P. Blauert (1997). "Spatial hearing" (revised edition), MIT Press, Cambridge, MA.

[10] M. Slaney (1998). "A critique of pure audition", in Computational Auditory Scene Analysis, ed. by D. F. Rosenthal, H. G. Okuno, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 27-42.

[11] C. Jin, A. Corderoy, S. Carlile, and A. van Schaik (2000). "Spectal cues in human sound localization", Advances in Neural Information Processing Systems 12, edited by S.A. Solla, et al., MIT Press, Cambridge, MA, pp. 768-774.

[12] J.W.Strutt (Lord Rayleigh) (1907). "On our perception of sound direction", Phil.Mag., vol. 13, pp. 214-232.

[13] D. W. Batteau (1967). ”The role of the pinna in human localization”, Proc. Royal Society London, vol. 168 (series B), pp. 158-180.

[14] D. Wright, J. H. Hebrank, and B. Wilson (1974). “Pinna reflections as cues for localization”, J. Acoustic Soc. Am., vol. 56, no. 3, pp. 957-962.

[15] R. O. Duda (1993). “Modeling head related transfer functions”, Proc. 27th Asilomar conf. on Signal, Systems and Computers, Asilomar, CA, pp. 457-461.

[16] E. A. Lopez-Poveda and R. Meddis (1996). “A physical model of sound diffraction and reflections in the human concha”, J. Acoust. Soc. Am., vol. 100, no. 5, pp.3248-3259.

[17] E. A. Durant and G. H. Wakefield (2002). “Efficient model fitting using a genetic algorithm: pole-zero approximations of HRTFs”, IEEE Trans. on Speech and Audio Processing, vol. 10, no. 1, pp. 18-27.

[18] C. P. Brown and R. O. Duda (1998). “A structural model for binaural sound synthesis”, IEEE Trans. Speech and Audio Processing, vol. 6, no. 5, pp. 476-488.

[19] T. Funkhouser, P. Min, and I. Carlbom (1999). “Real-time acoustic modeling for distributed virtual environments”, Proc. SIGGRAPH 1999, Los Angeles, CA, pp. 365-374.

[20] J.-M. Jot (1999). “Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces”, Multimedia Systems, vol. 7, no. 1, pp. 55-69.

[21] B. G. Shinn-Cunningham (2000). “Distance cues for virtual auditory space”, Proc. IEEE PCM2000, Sydney, Australia, pp. 227-230.

[22] E. M. Wenzel, J. D. Miller, and J. S. Abel (2000). “A software-based system for interactive spatial sound synthesis”, Proc. ICAD 2000, Atlanta, GA, pp. 151-156.

[23] N. Tsingos (2001). “A versatile software architecture for virtual audio simulations”, Proc. ICAD 2001, Espoo, Finland, pp. 38-43.

[24] M. B. Gardner and R. S. Gardner (1973). ”Problem of localization in the median plane: effect of pinna cavity occlusion”, J. Acoust. Soc. Am., vol. 53, no. 2, pp. 400-408.

[25] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman (1993). “Localization using non-individualized head-related transfer functions”, J. Acoust. Soc. Am., vol, 94, no. 1, pp. 111-123.

[26] W. M. Hartmann (1999). “How we localize sound”, Physics Today, November 1999, pp. 24-29.

[27] S. Carlile, ed. (1996). “Virtual auditory space: Generation and applications”, R. G. Landes Company, Austin, TX.

[28] D. S. Brungart and W. R. Rabinowitz (1996). “Auditory localization in the near field”, Proc. ICAD 1996, Palo Alto, CA (http://www.santafe.edu/~icad/ICAD96/proc96/INDEX.HTM)

[29] R. A. Butler (1975). ”The influence of the external and middle ear on auditory discriminations”, in Handbook of Sensory Physiology, edited by W. D. Keidel and W. D. Neff (Springer Verlag, New York), pp. 247-260.

[30] H. L. Han (1994). “Measuring a dummy head in search of pinnae cues”, J. Audio Eng. Society, vol. 42, no. 1, pp. 15-37.

[31] E. A. G. Shaw (1997). “Acoustical features of the human external ear”, in Binaural and Spatial Hearing in Real and Virtual Environments, ed. by R. H. Gilkey and T. R. Anderson, Lawrence Earlbaum Associates, Mahwah, NJ, pp. 25-48.

[32] J. B. Allen and D. A. Berkeley (1979). “Image method for efficiently simulating small-room acoustics”, J. Acoust. Soc. Am., vol. 65, no.5, pp. 943-950.

[33] J. Borish (1984). “Extension of the image model to arbitrary polyhedra”, J. Acoust. Soc. Am., vol. 75, no. 6, pp. 1827-1836.

[34] R. Duraiswami, N. A. Gumerov, D. N. Zotkin, and L. S. Davis (2001). “Efficient evaluation of reverberant sound fields”, Proc. IEEE WASPAA01, New Paltz, NY, pp. 203-206.

[35] S. Perret and W. Noble (1997). “The effect of head rotations on vertical plane sound localization”, J. Acoust. Soc. Am., vol. 102, no. 4, pp. 2325-2332.

[36] C. Kyriakakis, P. Tsakalides, and T. Holman (1999). “Surrounded by sound: Immersive audio acquisition and rendering methods”, IEEE Signal Processing Magazine, vol. 16, no. 1, January 1999, pp. 55-66.

[37] R. Duraiswami et al. (2000). “Creating virtual spatial audio via scientific computing and computer vision”, Proc. of 140th meeting of the ASA, Newport Beach, CA, December 2000, p. 2597. Available on the world wide web at http://www.acoustics.org/press/140th/duraiswami.htm

[38] J. Schoukens and R. Pintelon (1990). “Measurement of frequency response functions in noisy environments”, IEEE Trans. on Instrumentation and Measurement, vol. 39, no. 6, pp. 905-909.

[39] A. Kulkami, S. K. Isabelle, and H. S. Colburn (1999). “Sensitivity of human subjects to head-related transfer-function phase spectra”, J. Acoust. Soc. Am., vol. 105, no. 5, pp. 2821-2840.

[40] R. S. Woodworth and G. Schlosberg (1962). “Experimental psychology”, Holt, Rinehard and Winston, NY, pp.349-361.

[41] C. Kyriakakis (1998). “Fundamental and technological limitations of immersive audio systems”, Proc. IEEE, vol. 86, no. 5, pp. 941-951.

[42] L. E. Kinsler (editor), A. R. Frey, A. B. Coppens, and J. V. Sanders (1982). “Fundamentals of acoustics” (third edition), John Wiley & Sons, pp. 313-321.

[43] B. Rakerd and W. M. Hartmann (1985). “Localization of sound in rooms, II: The effects of a single reflecting surface”, J. Acoust. Soc. Am., vol. 78, no.2, pp. 524-533.

[44] B. G. Shinn-Cunningham (2001). “Localizing sound in rooms”, proc. of the ACM SIGGRAPH and Eurographics Campfire: Acoustic Rendering for Virtual Environments, Snowbird, Utah.

[45] W. G. Gardner (1995). “Efficient convolution without onput-output delay”. J. Audio Eng. Soc., vol. 43, no.3, pp. 127-136.

[46] D. R. Begault, E. M. Wenzel, and M. R. Anderson (2001). “Direct comparison of the impact of head-tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source”, J. Audio Eng. Soc., vol. 49, no. 10, pp. 904-916.

[47] V. R. Algazi, R. O. Duda, D. P. Thompson, and C. Avendano (2001). “The CIPIC HRTF database”, Proc. IEEE WASPAA01, New Paltz, NY, pp. 99-102.

[48] CIPIC HRTF Database Files, Release 1.0, August 15, 2001, available from http://interface.cipic.ucdavis.edu/

[49] S. Carlile, C. Jin, V. Harvey (1998). “The generation and validation of high fidelity virtual auditory space”, Proc. 20th Annual Intl. Conf. of the IEEE Engineering in Medicine and Biology Society, vol. 20, no. 3, pp. 1090-1095.

[50] P. Runkle, A. Yendiki, and G. Wakefield (2000). “Active sensory tuning for immersive spatialized audio”, Proc. ICAD 2000, Atlanta, GA.

[51] J. Chen, B. D. van Veen, K. E. Hecox (1993). “Synthesis of 3D virtual auditory space via a spatial feature extraction and regularization model”, Proc. IEEE Virtual Reality Annual Int. Symp., pp. 188-193.

[52] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson (2001). “Structural composition and decomposition of HRTFs”, Proc. IEEE WASPAA01, New Paltz, NY, pp. 103-106.

[53] J. C. Middlebrooks (1999). “Individual differences in external-ear transfer functions reduced by scaling in frequency”, J. Acoust. Soc. Am., vol. 106, no.3, pp. 1480-1492.

[54] J. C. Middlebrooks (1999). “Virtual localization improved by scaling non-individualized external-ear transfer functions in frequency”, J. Acoust. Soc. Am., vol. 106, no.3, pp. 1493-1510..

[55] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan (2000). “Psychophysical customization of directional transfer functions for virtual sound localization”, J. Acoust. Soc. Am., vol. 108, no. 6, pp. 3088-3091.