Rendering Localized Spatial Audio in a Virtual Auditory Space

Dmitry N. Zotkin, Ramani Duraiswami, *Member, IEEE*, and Larry S. Davis, *Fellow, IEEE*

***Abstract*—High-quality virtual audio scene rendering is required for emerging virtual and augmented reality applications, 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.**

***Index Terms*—Audio user interfaces, head-related transfer function, spatial audio, 3-D audio processing, user interfaces, virtual auditory spaces, virtual environments, virtual reality.**

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 that 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 (e.g., [2]) have demonstrated the feasibility of spatial audio for data display. 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]).

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]–[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 their relative importance [11]. Previous work in the area of localization and spatial sound rendering can be tracked back to 1907 [12]. Since then, understanding of spatial localization [13], [14], modeling of the involved transfer functions [15]–[17], fast synthesis methods [18], environment modeling [19]–[21], and implementation of the rendering software [22], [23] have made significant progress.

Our goal is the creation of 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) that varies significantly between people. It is known [24]–[26] that differences in ear shape and geometry strongly distort perception and that the high-quality synthesis of a virtual audio scene requires 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 nonexternalized, because reverberation cues that arise from environmental reflections 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 fidelity of the virtual audio scene.

In this paper, we present a set of fast algorithms for headphones-based spatial audio rendering that are able to recreate all these mentioned cues in near real time. Our rendering system has a rendering latency that is within the acceptable limits reported in [27]. It is implemented on a commercial off-the-shelf PC, and needs no additional hardware other than a 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 utilizing optimizations available on Intel Xeon processors. We also partially address the problem of personalization of the HRTF by selecting the HRTF that corresponds to the closest one from a database of 43 pairs of HRTFs. This selection is performed by matching a person's 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 II, we consider the scattering related cues arising from interaction of the sound wave with the anatomy, and the environment. We

Manuscript received March 20, 2002; revised October 15, 2002. This work was supported in part by the National Science Foundation under Award ITR-0086075 and in part by the Office of Naval Research under Grant N000140110571. The associate editor coordinating the review of this paper and approving it for publication was Prof. Yoshinori Kuno.

The authors are with the Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA (e-mail: dz@umiacs.umd.edu; ramani@umiacs.umd.edu; lsd@umiacs.umd.edu).

Digital Object Identifier 10.1109/TMM.2004.827516

1520-9210/04$20.00 © 2004 IEEE

Meta Platforms, Inc.
Exhibit 1111
Page 001

introduce the head-related transfer function, the knowledge of which is crucial for accurate spatial audio. We also describe the environmental model that provides important cues for perception (in particular, cues that lead to "out-of-the-head" externalization) and its characterization via the room impulse response. In Section III, the importance of dynamic cues for perception is outlined. In Section IV, we describe the fast audio-rendering algorithms. Section V deals with partial HRTF customization using visual matching of ear features. In Section VI, our experimental setup and experimental results are presented. Section VII concludes the paper.

II. Scattering Based Cues

Using just two receivers (ears), humans are able to localize sound with amazing accuracy [28]. Although 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 is the "true" source position. This localization is possible because of the other localization cues arising from sound scattering.

The wavelength of audible sound (2 cm–20 m) is comparable to the dimensions of the environment, the dimensions of the human body, and for high frequencies, the dimensions of the external ear (pinna). As a result, the circularly-asymmetric pinna forms a specially-shaped "antenna" that causes a location-dependent and frequency-dependent "filtering" of the sound reaching the eardrums, especially at higher frequencies. 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. Scattering off the environment (room walls, etc.) provides additional cues for the source position.

The effect of both the anatomical scattering and the time and level differences can be described by a head-related impulse response (HRIR), or alternatively its Fourier transform, which is called the head-related transfer function (HRTF). Similarly, environmental scattering can be characterized by a room impulse response (RIR). Knowing the HRIR and the RIR, 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 sound pressure signals that are identical to the stimulus in the real scene are presented at the listener's ears, and they change the same way with his motion, he will get the same perception as he would have had in the real scene, including the perception of the presence of a sound source at the correct location in exocentric space, the environmental characteristics, and other scene aspects. Thus knowledge of the RIR and the HRTF is the key to rendering virtual spatial audio.

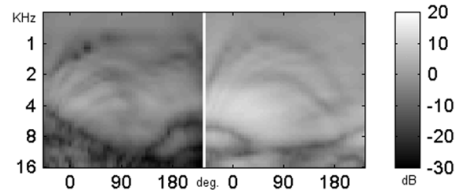


Fig. 1. HRTF magnitude slices for the contralateral and ipsilateral ears for a fixed azimuth of 45° and varying elevation for a subject from the CIPIC HRTF database [48].

A. Head Related Transfer Function

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 [30]–[32], 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 sound pressure level (SPL) at the corresponding eardrum $\Phi_{l,r}$ to the free-field SPL at the center of the head as if the listener were absent Φ_f:

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

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 $225°$ for an azimuth of 45°. The plot contains several peaks and valleys, which shift as the elevation changes. The effects of the different body parts are visible 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 [33]–[35].

Typically the HRTF is measured by presenting sound stimuli from different locations to a listener whose ears are instrumented with microphones, and then using (1). The measurements can only be made at a finite set of locations, and when a sound source at an intermediate location must be rendered, the HRTF must be interpolated. If a nearest neighbor approach is used instead of interpolation, the user hears audible sudden changes in the sound spectrum when the source position changes. The spectrum changes manifest themselves as clicks and noise, and perception of position can be adversely affected.

Some of the features of the HRTF arise due to coherent addition or cancellation of waves after reflection or diffraction. Simple interpolation does not preserve these features, but would rather result in a broad peaks of nulls. Other HRTF features arise due to resonance effects, which are poorly understood. A simple additive interpolation scheme may thus have difficulty in producing perceptually valid interpolations. The phases of the transfer function can be defined only within a multiple of 2π, which introduces further phase unwrapping errors on the interpolated value of the phase. The phases of the measured HRTFs

Meta Platforms, Inc.
Exhibit 1111
Page 002

encode the time delays, and often are in error. Finally, to capture fine details of the HRTF the sampling must be fine enough, i.e., satisfy a Nyquist criterion.

The paper [36] suggests geometric interpolation as a way to properly interpolate complex valued frequency response functions. Separate arithmetic interpolation of the amplitude and the phase gives the same result for the interpolated phase as the geometric interpolation. More complex interpolation methods aimed specifically at HRTF interpolation are interpolation using pole-zero approximation models [37], [38] and spherical spline-based methods [39]. It is also known that 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 [40] and the measured phase is likely to be contaminated anyway due to difficulties of measuring it accurately because of sampling and other problems. It is safe to say that the subject of HRTF interpolation is an area likely to see further research.

B. Environmental Modeling

Using the HRTF alone to render the sound scene results in perception of a "flat" or "dry" auditory space 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. To achieve good externalization and distance perception, environmental scattering cues must be incorporated in to the simulation of auditory space [21]. Environmental scattering is characterized by a room transfer function, or alternatively a room impulse response (RIR). The RIR includes effects due to reflection at the boundaries, sound absorption, diffraction around edges and obstacles, and low frequency resonance effects. The RIR depends on the locations of both the source and the receiver. For computing the RIR, a simple image model has been proposed for rectangular rooms [41] and has been extended to the case of piecewise-planar rooms [42]. For computing the IR at multiple room points in a rectangular room we presented a fast algorithm based on the multipole method in [43]. These models capture some of the features of the RIR and are adopted in our system.

Multiple reflections create an infinite lattice of image sources in these image models. The positions of these image sources can be found by simple geometric computations and visibility testing. Absorption is accounted for by multiplying image source strengths by a heuristic coefficient β for every reflection occurred. (We use $\beta = 0.9$ for walls and 0.7 for carpeted floors and false ceilings). Summing the peaks at time instants $\tau = d/c$, where d is the distance from the ith image source, with amplitudes determined by the distance and the source strength, we can compute the room impulse response.

III. Dynamics

In addition to static localization cues (ITD, ILD, and anatomical scattering) and environmental scattering, humans use dynamic cues to reinforce localization. Studies on the importance of these cues date back to 1940 [44]. They arise from active, sometimes unconscious, motions of the listener, which change the relative position of the source [45]. It is reported that front/back confusions that are common in static listening tests disappear when listeners are allowed to slightly turn their heads to help them in localizing sound [46].

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 causes it to be perceived at the one location that stays fixed in the moving coordinate system—the origin of that coordinate system, inside the head. Low latency head position and orientation tracking is necessary so that dynamic cues are recreated, and delay between head motion and the resulting changes in audio stream are not distracting.

IV. Audio Scene Rendering Algorithms

As described above, 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 head tracking is employed to stabilize the virtual audio scene. 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.

When rendering the environmental model, one is faced with competing requirements of low latency rendering and the necessity for convolution with long filters. Convolution in the Fourier domain is efficient, but requires delays of at least one frame. Convolution in the time-domain is inefficient, but capable of low latency. In our system we take the approach that some latency is unavoidable in the rendering, and use this latency, decomposition of the filtering, and the linearity of the convolution operation to achieve better system performance. The output audio stream is synthesized block-by-block by performing frequency domain convolution of the input stream with a short rendering filter. The length of this filter is a parameter, and we typically set it at 100 ms.

In addition to the virtual source, image sources created by reflections off the room walls that simulate reverberation must be rendered. In several existing systems (for example, in [20]), the input data stream is convolved separately in the time domain with the HRIR of each image source, often using specialized hardware, and the results are summed up. The length of each HRIR is a few milliseconds (we use a 256-tap HRIR corresponding to 5.8 ms at a sampling rate of 44.1 kHz). As the number of image sources increases with the number of reflections simulated, this approach becomes infeasible. In our system, we first pack all HRIRs into a rendering filter that

Meta Platforms, Inc.
Exhibit 1111
Page 003

consists of the sum of the appropriately delayed head-related impulse responses for the virtual source and for the image sources. Then, the frequency-domain convolution with the rendering filter is performed using the extremely efficient fast Fourier transform software library, FFTW, which is freely available on the Web [47].

Frequency-domain convolution introduces latency due to the blocky nature of convolution. This latency is inevitable, along with the latency of the tracking subsystem, and the challenge is to use it wisely. Essentially, during the time of playback of one data block the next block should be computed (which includes updating of the rendering filter to accommodate new positions of the source and the receiver and convolving the next data block with the freshly computed filter), and maximum fidelity should be achieved in this time. Increasing block size will obviously increase the available processing time and the amount of processing that can be done, but the system latency will also increase. We use a block size of 4096 samples and use the resulting latency as a time frame to update the rendering filter as much as possible. It turns out that on our desktop system the processing power available is sufficient to simulate as many as five orders of reflections in real time. We report our estimations of the latency of the system and compare them to the published studies of acceptable latency later in the section devoted to experimental results.

A. Head Tracking

We use a Polhemus system for head tracking. The tracker provides the position (Cartesian coordinates) and the orientation (Euler angles) of up to four receivers with respect to a transmitter. In our system a receiver is mounted on headphones. The transmitter might be fixed, creating a reference frame, or be used to simulate a virtual sound source that can be moved by the user. The positions of the virtual sources in the listener's frame of reference are computed by simple geometry, and these virtual sources are rendered at their appropriate locations. The tracking latency is limited by the operating system and the serial port operation speed, and is approximately 40 ms (we found that the data becomes corrupted if smaller delays are used). Multiple receivers are used to enable multiple participants. The Polhemus transmitter has a tracking range of only about 1.5 m, limiting the system's spatial extent. Because the tracker provides the position and orientation data of the receiver with respect to the 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. 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 premeasured sets of HRTFs from the HRTF database recently released by the CIPIC laboratory at UC Davis. The database and measurement procedure are described in detail in [48]. As measured, the HRTF corresponds to sounds generated 1 m away and sampled on a sphere at a given angular resolution. For rendering the virtual sound source at an arbitrary spatial location, the HRTF for the corresponding direction is required. Since HRTFs are measured only at a number of fixed directions, given an arbitrary source position, the HRTF must be interpolated.

As discussed previously, the phase of the measured HRTFs is prone to noise and other errors and is difficult to interpolate. We replace the phase of the HRTFs with a value that gives the correct interaural time difference for a sphere (Woodworth's formula [50])

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

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

As far as the magnitude is concerned, we interpolate the values from the three closest available measurements of the HRTF. The database we use has the directional transfer functions on a lattice with 5°step in azimuth and elevation for different people. We interpolate the associated HRIRs by finding the three closest lattice points $P_i = (\varphi_i, \theta_i), i = 1 \ldots 3$, with corresponding distances d_i between P and P_i.[1] 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 taken as

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

with weights $w_i = 1/d_i$. To prevent numerical instability, w_i is bounded from above by some constant $C = 100$. Using the value $\hat{\tau}$ from (2), the phase $\hat{\varphi}(\omega)$ of the interpolated HRTF corresponding to the leading and lagging ears are respectively set to

$$\tilde{\varphi}_{\text{leading}}(\omega) = \omega\hat{\tau}/2, \qquad \tilde{\varphi}_{\text{lagging}}(\omega) = -\omega\hat{\tau}/2.$$

Time shifts are performed in the frequency domain because humans are sensitive to ITD variations as small as 7 μs [51], which is $1/3$ of a sampling period at a rendering rate of 44.1 kHz. The resulting HRTF $\tilde{H}(\omega) = \tilde{A}(\omega)e^{-i\tilde{\varphi}(\omega)}$ 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.

It is also desirable to find the closest interpolation 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 360×180 table covering all possible integer values of azimuth and elevation. The cell (i, j) in the table stores the n identifiers of the lattice points that are closest to the point with azimuth $i + 0.5$ and elevation $j + 0.5$. To find the closest points to 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). This significantly improves the performance of the online renderer compared to a brute-force search.

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

Meta Platforms, Inc.
Exhibit 1111
Page 004

C. Incorporation of the Room Model

The room impulse response (RIR) can be analytically approximated for rectangular rooms using a simple image model [41]. A more complex image model with visibility criteria [42] can be applied for the case of more general rooms. The RIR is a function of both the source and receiver locations, and as their positions change 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. These 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 a problem for evaluating the RIR, and the length of the RIR presents a problem for low-latency rendering. Time-domain convolution with long filters is computationally inefficient and frequency-domain convolution introduces significant processing delays due to block segmentation. We present below a solution to this problem, based on a decomposition of the RIR.

The reverberation time of a 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. However, the behavior of the tail does not appear to depend on the position of the source and the receiver within a room, which can be expected because 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%. This observation suggests that the tail may be approximated by a generic tail model for a room of similar size, thereby avoiding expensive recomputation of the tail for every source and receiver position.

However, the positions of early reflections do change significantly when the source or the receiver is moved. It is believed ([53], [54]) 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. Similarly to numerous existing solutions [55], [56], we break the impulse response into first few spatialized echoes and a decorrelated reverberant tail. 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 boundary. However, due to the fact that the early reflection filter is performed in the frequency domain, we are able to include many more reflection components.

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 image source and its images created by reflections. The substantial length of 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 from 400 ms (typical office environment) to 2 s 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 the processing is to be done in real-time. Frequency-domain processing using fast Fourier transform is 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 [57], 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 ms (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 ($N_2 = 65\,536$ 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 zero to N_1 in it is zeroed out.

By splitting the convolution into 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 having an unacceptable delay of order N_2 if a full 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 is $H = H_1 + H_2$ (because the samples from zero to N_1 in H_2 is zeroed out) and no delays are necessary.

Mathematically, the (continuous) input data stream $X = \{x(1), x(2), \ldots, x(n), \ldots)$ is convolved with the filter $H = \{h(1), h(2), \ldots, h(N_2))$ to produce the output data stream $Y = \{y(1), y(2), \ldots, y(n), \ldots)$. 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 zero to N_2 with $h_2(1), h_2(2), \ldots, h_2(N_1)$ 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 computed in real time

Meta Platforms, Inc.
Exhibit 1111
Page 005

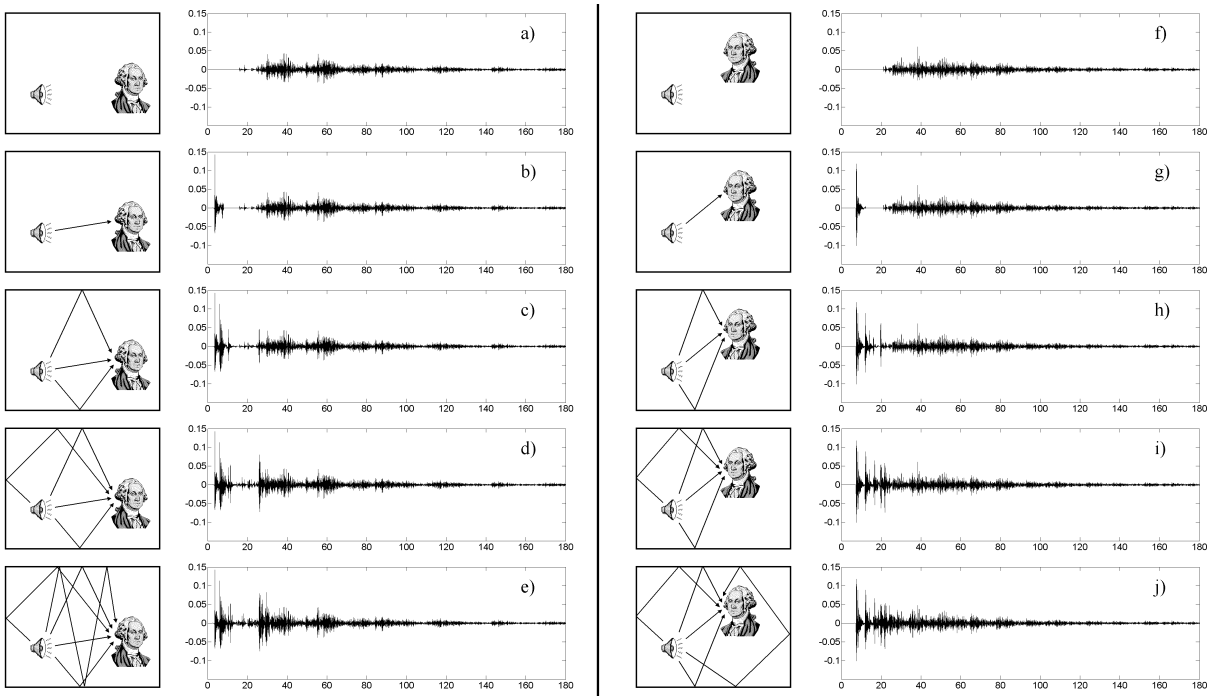


Fig. 2. Synthesis of a rendering FIR filter in real time. (a) Precomputed tail of the filter (reflections of order 4 and above). (b)–(e) Addition of reflections of order 0, 1, 2, and 3, respectively. (f) Same as (a) but for different position and orientation of the receiver. (g)–(j) Same as (b)–(e).

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 offline.

In case the source signal is only available online, it must be delayed sufficiently to allow the reverberant precomputation to be done before the actual output starts, but once this is done, the reaction of the system to the user's head motion is fast because only the frequency-domain convolution with a short filter H_1 (which changes on-the-fly to accommodate changes in user position) is done online. 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 locations of source and receiver. Once the new coordinates of the source and the receiver are known, the algorithm recomputes the first part of the FIR filter and places it on top of the second part. Fig. 2 shows the process of composition for two different arrangements of the source and the receiver and $L_1 = 4$. The composition starts with the precomputed tail of IR that stays constant independent of the source and receiver positions. In the four steps shown, it 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 smaller order may come later than some reflections of larger order because of the room geometry, so the fixed part overlaps with the part that is recomputed on the fly. When a 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 because 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 via the PC soundcard using standard operating system calls. Due to the internal buffering of the operating system, it is necessary to have at least one full buffer in the output queue of the PC sound interface. Therefore, the sound output subroutine initially outputs two buffers of data and upon receiving a buffer completion event for the first of these two computes the next buffer using the currently available

Meta Platforms, Inc.
Exhibit 1111
Page 006

source positions to synthesize the system IR and performing frequency-domain convolution of the computed IR with the data. These computations take place within the allowed time frame, which is determined by the time of playback of the single data buffer that is currently playing. The freshly computed buffer is then placed in the queue, and the process repeats. Thus, the maximum latency of the playback subsystem from the availability of new source position to the corresponding change in the rendered sound is limited by the length of the playback buffer.

F. Headphone Equalization

Equalizing the sound to account for the headphones is relatively simple to do, and is well described in [71]. While significant effects are seen, they do not change with the location of the rendered source, and it is still an open issue whether headphone compensation and missing ear-canal response reintroduction [71] are necessary for proper perception of the rendered source in exocentric space. A recent study [71] suggests that only the variations of sound spectrum across source locations provide the localization cues that the listener uses to determine source position, and static features (even of comparable magnitude) do not influence localization. In our system we do not perform headphone or ear-canal compensation. Preliminary testing with addition of such equalization suggests that while the perceived sound is different, perception of externalization and localization is not affected.

V. 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 his/her own HRTF using feedback about the source position through lifelong experience, 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 [26], [58]). The problem of HRTF customization is currently an open question that is the subject of much research. The usual customization method is by direct measurement of the HRTF for the particular user. This method is accurate but highly time-consuming, and there are different measurement issues complicating the procedure [60]. Alternative approaches such as allowing the participant to manipulate different characteristics of the HRTF set used for rendering until she achieves satisfactory experience have been proposed (see, e.g., [61]), although it is not clear if the correct HRTF is achieved. A novel and promising approach is the direct computation of the HRTF using a three-dimensional ear mesh obtained via computer vision and solving the physical wave propagation equation in the presence of the boundary by fast numerical methods [49]. However, this work is still under development, and current virtual auditory systems do not yet have 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 [48], [59].

A. Approaches to Customization

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; therefore, it is natural to make the hypothesis that the structure of the HRTF is related to scattering body part dimensions and orientation. Some studies, such as HRTF clustering and selection of a few most representative ones [62], functional representation of HRTFs using spatial feature extraction and regularization model [63], a structural model for composition and decomposition of HRTF [64], and especially experiments with development of a functional model relating morphological measurements to HRTFs [65] and with HRTF scaling [66]–[68] 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 modified accordingly. For example, the work of Middlebrooks [66], [67] is based on the idea of frequency scaling: 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. Because 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. However, the ears of different persons are different in more ways than can be captured by just a simple scaling, and a seemingly insignificant small change in ear structure can cause dramatic changes in the HRTF.

B. Database Matching

An intermediate approach that 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 [48] contains the measurement of the HRTFs of 43 persons, along with some anthropometric information about the subjects. The HRTFs are measured on a spherical lattice using a speaker positioned 1 m away from the subject. The subject's ear canals were blocked, and the measurement results were free-field compensated. HRTF measurements below -45° of elevation are not available (the speaker cannot be placed below the person). 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 Fig. 3 and are as follows: $d_1 \cdots d_8$ 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 seven of these ten pinna parameters that can be easily measured from a frontal ear picture.

We performed 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

Meta Platforms, Inc.
Exhibit 1111
Page 007

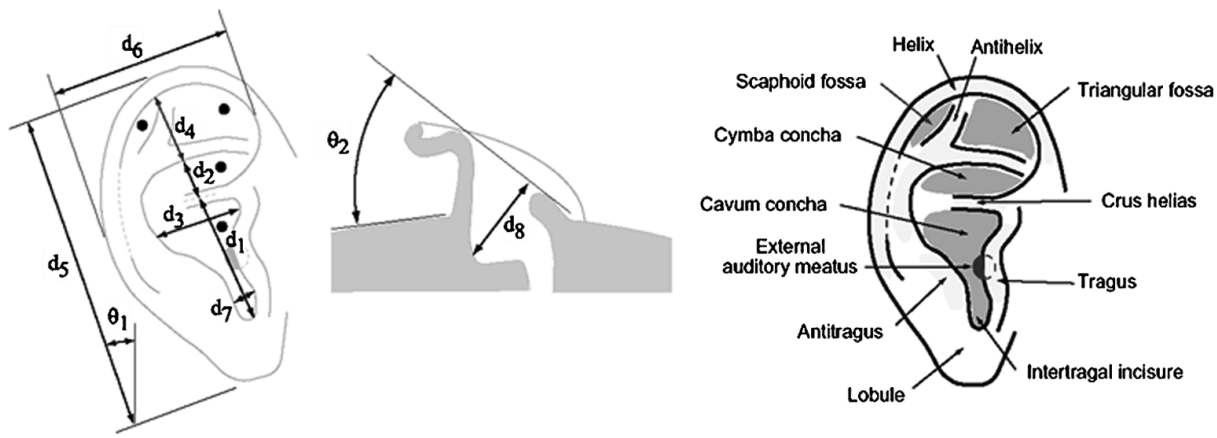


Fig. 3. Set of measurements provided with the HRTF database.

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 $\mathrm{Var}(d_i)$, then the error for this parameter $e_i = (\hat{d}_i - d_i)/\mathrm{Var}(d_i)$, 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 though.

We have developed a simple interface that allows us to perform quick selection of the best-matching HRTF from the database. The pictures of the left and the right ears of the new virtual audio system user are taken with two cameras, with the user holding a ruler in the frame to provide a scale of reference. A sample picture used in one of the sessions of HRTF matching is shown in Fig. 4. An operator identifies key points on the ear picture and measures the 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 that 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 because they cannot 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.

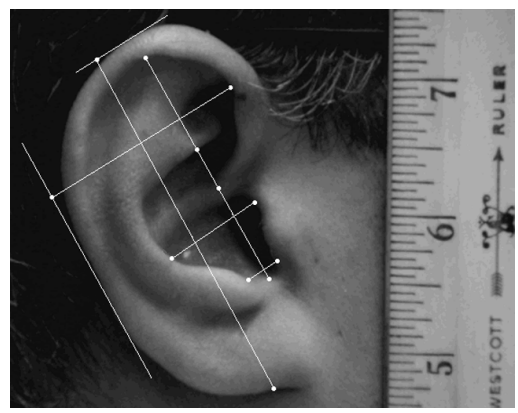


Fig. 4. Sample picture of the ear with the measurement control points marked.

VI. Experimental Results

A number of volunteers ($\sim$150) were subjects of some informal listening experiments, in which a image 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 nonpersonalized 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, along with the nonpersonal HRTF cues, are responsible for these reports. The perceived sound motion is quite smooth, and no jumps or clicks are noticeable. The stability of the synthesized virtual audio scene is also remarkable and latency is noticeable only if the user rotates her head or moves the source in a very fast, jerky motion. Even better results should be achievable with personalized HRTFs.

A. System Setup and Performance

The current setup used for experiments is based on a 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 a Sony LDI-D100B stereo head-mounted display, which is used for creating an immersive virtual environment. The programming is done in Microsoft Visual C++ 6.0, using OpenGL for video. Computations are

Meta Platforms, Inc.
Exhibit 1111
Page 008

TABLE I
LOCALIZATION ERROR FOR GENERIC AND PERSONALIZED HRTF SETS

	s1	s2	s3	s4	s5	s6
Average elevation error, generic HRTF, degrees	9.0	9.5	5.5	16.7	14.4	7.2
Average elevation bias, generic HRTF, degrees	-4.0	-4.5	5.0	-9.0	-8.3	3.8
Average elevation error, personalized HRTF, degrees	7.6	7.2	4.4	12.9	13.6	12.5
Average elevation bias, personalized HRTF, degrees	-1.4	-7.0	-2.0	4.8	4.8	-6.3

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 system 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.

We estimate total system latency (TSL) similar to [69] by adding the individual latencies for the different processing steps. The Polhemus tracker is operating over the serial link with a baud rate of 57 600. There is an artificial delay of 40 ms between sending a command to the tracker and reading back the response. This delay is introduced into the tracking thread to avoid data corruption. The length of the tracker response message is 47 bytes in ASCII format and it takes approximately 9 ms to transmit it over the serial link. As described in the Section IV-E, the latency of the playback synthesis is limited by the playback buffer length which is 4096 samples, corresponding to a time of 93 ms. Then the TSL is bounded from above by the sum of these numbers, which is 142 ms. It was reported in [70] that the minimum detectable delay in case of audio–video asynchrony is 187.5 ms, and in [27] the latency of the dynamic virtual audio system was not obvious to the subjects until it reached 250 ms; and even with a latency of 500 ms, the localization performance was comparable to the no latency case, suggesting that the listeners are able to ignore latency for localization purposes. We conclude that the latency of our system falls within the limits of perceptually unnoticeable latency.

B. Nonpersonalized HRTF Set

While most people reported very consistent sound externalization and localized a source when given a visual cue to its location, we wished to test the ability of the noncustomized system to render virtual sources accurately. We performed small-scale formal tests of the system on six people. The test sounds were presented through headphones. The head tracker was used to obtain the head position when the subject "points" to a rendered virtual sound source. The pointing mechanism is calibrated by having subjects look at a source placed at a known spatial location. The sounds used for the tests were three 75 ms bursts of white noise with 75 ms pauses between them, repeated every second. The sound stayed on for 3 s. 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° of azimuth and elevation. Then, the sound was presented at a random position, with $\varphi \in [-90°, 90°], \theta \in [-45°, 45°]$. Subjects were asked to "look" at the image source in the same way that they looked at the source during calibration. 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 a button. The localization error is recorded and the next source is presented. Results are summarized in the first two lines of Table I. For clarity, we present the average localization error and average localization bias only for elevation measurement of the virtual source, the perception of which is believed to be hampered most by use of a nonindividualized HRTF. The errors in azimuth are much lesser.

The results for the "generic" HRTF set are interesting. Some subjects perform better than the others in the elevational localization; subject 3 performs quite well, while the performance of subjects 1, 2, and 6 is close to the average and subjects 4 and 5 perform poorly. Errors are probably due to nonindividualized HRTFs. The results show that, as can be reasonably expected, the localization with nonindividualized HRTFs 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". Overall, the system is shown to be able to create convincing and accurate virtual auditory displays, and the accuracy can be improved significantly by personalization as follows.

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 tests were conducted in the same manner as above. The two last lines of Table I are the results for the case where the sound is rendered with the best-matching HRTF from the HRTF database. It is clear that the elevation localization performance is improved consistently by 20–30% for four out of the six 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 fullscale set of experiments to confirm the statistical significance of these results. Improvement for the subject 5 is marginal and subject 6 performs worse with the "customized" HRTF.

Meta Platforms, Inc.
Exhibit 1111
Page 009

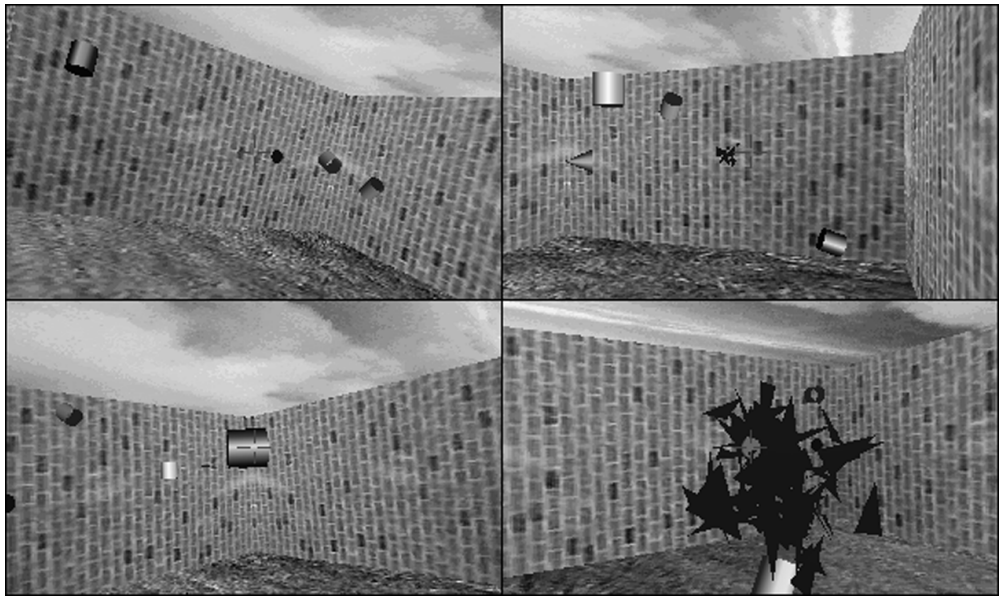


Fig. 5. Sample screenshots from the developed audio–video game with spatialized audio user interaface.

To test if the personalization resulted in a better subjective experience for the participants, we asked them which rendering they liked better. Subjects 1–4 reported that they are able to better feel the sound source motion in the median plane and that the virtual auditory scene synthesized with personalized HRTF sounds better (the externalization and the perception of DOA and source distance is better, and front-back confusions occur much less often). 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 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 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 "fit" 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 seven 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 informal reports of users' experience with an application. To do this we developed a simple game with spatialized sound, personalized stereoscopic visual display, head tracking and multiplayer capabilities all combined together. In the game, the participant wears stereo glasses and headphones with an attached head-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, which are given by his head motion like in an airplane simulator game. Multiplayer capability is implemented using a client-server model, in which the state of the game is maintained on one computer in a game server program that 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 that 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.

Four sample screenshots from the game are shown in Fig. 5. Three 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 cone in one of the screenshots corresponds to the ship of the second participant.

Meta Platforms, Inc.
Exhibit 1111
Page 010

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 various news sites on the World Wide Web. The listener can listen to some or all of them, and select their favorite one by getting into its proximity for selective listening, or shoot and break some cubes if they do not like the news being broadcast, in which case new cubes emerge later on connected to new live audio streams.

VII. Conclusions and Future Work

We have presented a set of algorithms for creating 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 perceptually unnoticeable latency, creating a highly convincing experience for participants.

Acknowledgment

The authors would like to thank Prof. R. O. Duda (San Jose State University), Prof. V. R. Algazi (University of California, Davis) and Prof. S. A. Shamma (University of Maryland) for many illuminating discussions on spatial audio. They would also like to thank V. Raykar and A. Mohan, of the Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies, University of Maryland, for assistance with some of the experiments and software. Thanks are due to the volunteers who participated in the tests of the developed auditory display and game. Finally, they would like to thank the reviewers for their insightful comments and suggestions that helped us to improve the manuscript.

References

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

[2] R. L. McKinley and M. A. Ericson, "Flight demonstration of a 3-D auditory display," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 683–699.

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

[4] S. A. Brewster, "Using nonspeech sounds to provide navigation cues," *ACM Trans. Comput.-Human Interact.*, vol. 5, no. 3, pp. 224–259, 1998.

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

[6] G. Kramer *et al.*. (1997) Sonification Report: Status of the Field and Research Agenda. *Prepared for the NSF by Members of the ICAD* [Online]. Available: http://www.icad.org/websiteV2.0/References/nsf.html

[7] S. Bly, "Multivariate data mappings," in *, Proc. Auditory Display: Sonification, Audification, and Auditory Interfaces, Santa Fe Institute Studies in the Sciences of Complexity*, G. Kramer, Ed. Reading, MA: Addison Wesley, 1994, vol. XVIII, pp. 405–416.

[8] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1994.

[9] J. P. Blauert, *Spatial Hearing*, revised ed. Cambridge, MA: MIT Press, 1997.

[10] M. Slaney, "A critique of pure audition," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno, Eds. Mahwah, NJ: Lawrence Erlbaum, 1998, pp. 27–42.

[11] C. Jin, A. Corderoy, S. Carlile, and A. van Schaik *et al.*, "Spectral cues in human sound localization," in *Advances in Neural Information Processing Systems 12*, S. A. Solla *et al.*, Eds. Cambridge, MA: MIT Press, 2000, pp. 768–774.

[12] J. W. Strutt, "On our perception of sound direction," *Philos. Mag.*, vol. 13, pp. 214–232, 1907. Lord Rayleigh.

[13] D. W. Batteau, "The role of the pinna in human localization," in *Proc. Royal Soc. London B*, vol. 168, 1967, pp. 158–180.

[14] D. Wright, J. H. Hebrank, and B. Wilson, "Pinna reflections as cues for localization," *J. Acoustic Soc. Amer.*, vol. 56, no. 3, pp. 957–962, 1974.

[15] R. O. Duda, "Modeling head related transfer functions," in *Proc. 27th Asilomar Conf. on Signal, Systems and Computers*, Asilomar, CA, 1993, pp. 457–461.

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

[17] E. A. Durant and G. H. Wakefield, "Efficient model fitting using a genetic algorithm: Pole-zero approximations of HRTFs," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 18–27, Jan. 2002.

[18] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 476–488, Sept. 1998.

[19] T. Funkhouser, P. Min, and I. Carlbom, "Real-time acoustic modeling for distributed virtual environments," in *Proc. SIGGRAPH 1999*, Los Angeles, CA, 1999, pp. 365–374.

[20] J.-M. Jot, "Real-time spatial processing of sounds for music, multimedia, and interactive human-computer interfaces," *Multimedia Syst.*, vol. 7, no. 1, pp. 55–69, 1999.

[21] B. G. Shinn-Cunningham, "Distance cues for virtual auditory space," in *Proc. IEEE PCM2000*, Sydney, Australia, 2000, pp. 227–230.

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

[23] N. Tsingos, "A versatile software architecture for virtual audio simulations," in *Proc. ICAD 2001*, Espoo, Finland, 2001, pp. 38–43.

[24] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: Effect of pinna cavity occlusion," *J. Acoust. Soc. Amer.*, vol. 53, no. 2, pp. 400–408, 1973.

[25] M. Morimoto and Y. Ando, "On the simulation of sound localization," *J. Acoust. Soc. Jpn. E*, vol. 1, no. 2, pp. 167–174, 1980.

[26] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 111–123, 1993.

[27] E. M. Wenzel, "Effect of increasing system latency on localization of virtual sounds," in *Proc. 16th AES Conf. Spatial Sound Reproduction*, Rovaniemi, Finland, 1999, pp. 42–50.

[28] W. M. Hartmann, "How we localize sound," *Phys. Today*, pp. 24–29, Nov. 1999.

[29] S. Carlile, Ed., *Virtual Auditory Space: Generation and Applications*. Austin, TX: R. G. Landes Company, 1996.

[30] D. S. Brungart and W. R. Rabinowitz. Auditory localization in the near field. presented at *Proc. ICAD 1996*. [Online]. Available: http://www.santafe.edu/~icad/ICAD96/proc96/INDEX.HTM

[31] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Amer.*, vol. 104, no. 5, pp. 3048–3058, 1998.

[32] B. G. Shinn-Cunningham, S. G. Santarelli, and N. Kopco, "Tori of confusion: Binaural cues for sources within reach of a listener," *J. Acoust. Soc. Amer.*, vol. 107, no. 3, pp. 1627–1636, 2000.

[33] *Handbook of Sensory Physiology*, W. D. Keidel and W. D. Neff, Eds., Springer–Verlag, New York, 1975, pp. 247–260. R. A. Butler, "The influence of the external and middle ear on auditory discriminations".

[34] H. L. Han, "Measuring a dummy head in search of pinnae cues," *J. Audio Eng. Soc.*, vol. 42, no. 1, pp. 15–37, 1994.

[35] E. A. G. Shaw, "Acoustical features of the human external ear," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 25–48.

[36] J. Schoukens and R. Pintelon, "Measurement of frequency response functions in noisy environments," *IEEE Trans. Instrum. Meas.*, vol. 39, pp. 905–909, Dec. 1990.

[37] A. Kulkarni and H. S. Colburn, "Efficient finite-impulse-response models of the head-related transfer function," *J. Acoust. Soc. Amer.*, vol. 97, p. 3278, 1995.

[38] M. A. Blommer and G. H. Wakefield, "Pole-zero approximations for head-related transfer functions using a logarithmic error criterion," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 278–287, May 1997.

Meta Platforms, Inc.
Exhibit 1111
Page 011

[39] S. Carlile, C. Jin, and V. van Raad, "Continuous virtual auditory space using HRTF interpolation: Acoustic and psychophysical errors," in *Proc. Int. Symp. Multimedia Information Processing*, Sydney, NSW, Australia, 2000, pp. 220–223.

[40] A. Kulkami, S. K. Isabelle, and H. S. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2821–2840, 1999.

[41] J. B. Allen and D. A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 5, pp. 943–950, 1979.

[42] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. Amer.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[43] R. Duraiswami, N. A. Gumerov, D. N. Zotkin, and L. S. Davis, "Efficient evaluation of reverberant sound fields," in *Proc. IEEE WASPAA01*, New Paltz, NY, 2001, pp. 203–206.

[44] H. Wallach, "The role of head movement and vestibular and visual cues in sound localization," *J. Exper. Psychol.*, vol. 27, pp. 339–368, 1940.

[45] S. Perret and W. Noble, "The effect of head rotations on vertical plane sound localization," *J. Acoust. Soc. Amer.*, vol. 102, no. 4, pp. 2325–2332, 1997.

[46] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Amer.*, vol. 105, pp. 2841–2853, 1999.

[47] [Online]. Available: http://www.fftw.org/

[48] V. R. Algazi, R. O. Duda, D. P. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE WASPAA01*, New Paltz, NY, 2001, pp. 99–102.

[49] R. Duraiswami *et al.*. Creating virtual spatial audio via scientific computing and computer vision. presented at *Proc. 140th Meeting of the ASA*. [Online]. Available: http://www.acoustics.org/press/140th/duraiswami.htm

[50] R. S. Woodworth and G. Schlosberg, *Experimental Psychology*. New York: Holt, Rinehard and Winston, 1962, pp. 349–361.

[51] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," *Proc. IEEE*, vol. 86, no. 5, pp. 941–951, 1998.

[52] A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 3rd ed, L. E. Kinsler, Ed: Wiley, 1982, pp. 313–321.

[53] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms, II: The effects of a single reflecting surface," *J. Acoust. Soc. Amer.*, vol. 78, no. 2, pp. 524–533, 1985.

[54] B. G. Shinn-Cunningham, "Localizing sound in rooms," in *Proc. ACM SIGGRAPH and Eurographics Campfire: Acoustic Rendering for Virtual Environments*, Snowbird, UT, 2001.

[55] M. R. Schroeder, "Natural-sounding artificial reverberation," *J. Audio Eng. Soc.*, vol. 10, no. 3, pp. 219–223, 1962.

[56] J.-M. Jot, "Efficient models for reverberation and distance rendering in computer music and virtual audio reality," in *Proc. 1997 Int. Computer Music Conf.*, 1997, pp. 236–243.

[57] W. G. Gardner, "Efficient convolution without input-output delay," *J. Audio Eng. Soc.*, vol. 43, no. 3, pp. 127–136, 1995.

[58] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "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, 2001.

[59] CIPIC HRTF Database Files, Release 1.0 (2001, Aug. 15). [Online]. Available: http://interface.cipic.ucdavis.edu/

[60] S. Carlile, C. Jin, and V. Harvey, "The generation and validation of high fidelity virtual auditory space," in *Proc. 20th Annu. Int. Conf. IEEE EMBS*, vol. 20, 1998, pp. 1090–1095.

[61] P. Runkle, A. Yendiki, and G. Wakefield, "Active sensory tuning for immersive spatialized audio," in *Proc. ICAD 2000*, Atlanta, GA, 2000.

[62] S. Shimada, M. Hayashi, and S. Hayashi, "A clustering method for sound localization transfer functions," *J. Audio Eng. Soc.*, vol. 42, no. 7/8, pp. 577–584, 1994.

[63] J. Chen, B. D. van Veen, and K. E. Hecox, "Synthesis of 3D virtual auditory space via a spatial feature extraction and regularization model," in *Proc. IEEE Virtual Reality Annual Int. Symp.*, 1993, pp. 188–193.

[64] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural composition and decomposition of HRTFs," in *Proc. IEEE WASPAA01*, New Paltz, NY, 2001, pp. 103–106.

[65] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, "Enabling individualized virtual auditory space using morphological measurements," in *Proc. First IEEE Pacific-Rim Conf. Multimedia (2000 Int. Symp. Multimedia Information Processing)*, 2000, pp. 235–238.

[66] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1480–1492, 1999.

[67] ——, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1493–1510, 1999.

[68] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan, "Psychophysical customization of directional transfer functions for virtual sound localization," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3088–3091, 2000.

[69] E. M. Wenzel, "The impact of system latency on dynamic performance in virtual acoustic environments," in *Proc. 16th Int. Congr. Acoustics and 135th Meeting Acoust. Soc. Amer.*, Seattle, WA, 1998, pp. 2405–2406.

[70] N. F. Dixon and L. Spitz, "The detection of auditory visual desynchrony," *Perception*, vol. 9, pp. 719–721, 1980.

[71] H. Moller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, pp. 171–218, 1992.

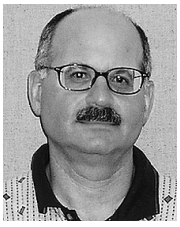
[72] K. McAnally and R. Martin, "Variability in the headphone-to-ear-canal transfer function," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 263–267, 2001.

Dmitry N. Zotkin was born in Moscow, Russia, in 1973. He received combined B.S./M.S. degrees in applied mathematics and physics from the Moscow Institute of Physics and Technology, Dolgoprudny, in 1996, and the M.S. and Ph.D. degrees in computer science from University of Maryland at College Park, in 1999 and 2002, respectively.

He is currently a Research Associate at the University of Maryland. His current research interests are in multimodal signal processing for multimedia and virtual reality, particularly in algorithms and methods for information presentation using customizable virtual auditory displays and perceptual processing interfaces.

Ramani Duraiswami (M'99) was born in Madras, India, in 1963. He received the B.Tech. degree from the Indian Institute of Technology, Bombay, in 1985 and the Ph.D. degree from The Johns Hopkins University, Baltimore, MD, in 1991.

He is a Research Scientist at the Institute for Advanced Computer Studies, University of Maryland. He is also Director of the Perceptual Interfaces and Reality Laboratory at the Institute, where multidisciplinary research in perceptual interfaces and virtual reality is conducted. He has broad research interests in the areas of virtual reality, computer vision, scientific computing, modeling human audition, computational acoustics, applied mathematics, and fluid dynamics.



Larry S. Davis (F'97) received the B.A. degree from Colgate University, Hamilton, NY, in 1970 and the M.S. and Ph.D. degrees in computer science from the University of Maryland, College Park, in 1974 and 1976, respectively.

From 1977 to 1981, he was an Assistant Professor, Department of Computer Science, University of Texas, Austin. He returned to the University of Maryland as an Associate Professor in 1981. From 1985 to 1994, he was the Director of the University of Maryland Institute for Advanced Computer Studies. He is currently a Professor in the Institute and the Computer Science Department, as well as Chair of the Computer Science Department. He is known for his research in computer vision and high performance computing. He has published over 75 papers in journals and has supervised over 12 Ph.D. students. He is an Associate Editor of the *International Journal of Computer Vision* and an area editor for *Computer Models for Image Processor: Image Understanding*.

Dr. Davis has served as program or general chair for most of the field's major conferences and workshops, including the 5th International Conference on Computer Vision, the field's leading international conference.

Meta Platforms, Inc.
Exhibit 1111
Page 012