CREATION OF VIRTUAL AUDITORY SPACES

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

Perceptual Interfaces and Reality Laboratory, UMIACS, University of Maryland, College Park 20742

ABSTRACT

High-quality virtual audio scene rendering is a must for emerging virtual/augmented reality applications and for perceptual user interfaces. We describe algorithms for creation of virtual auditory spaces using measured and non-individualized HRTFs and head tracking. Details of algorithms for HRTF interpolation, room impulse response creation, and audio scene presentation are presented. Tests show that individuals externalize well, and find our interface natural. The system runs in real time with latency of less than 30 ms on an office PC without specialized DSP.

1. INTRODUCTION AND PREVIOUS WORK

Recent advances in our understanding of human ability to perform sound source localization, and the rapid growth in computational power, enable creation of the new generations of multimedia and virtual/augmented reality systems. In such a system, instead of an interface consisting of a screen and speakers, the user would perceive a virtual or augmented multi-modal environment using a personal visual and auditory displays (e.g.,stereo glasses and headphones). The rendered audio and video kept consistent with each other and with the user's movements to create a virtual scene. Alternatively, in user interfaces for the visually impaired or in mobile applications a virtual auditory display alone may be desired. Rendered object characteristics such as pitch, timbre, intensity and spatial location, can be varied to make a UI rich and informative.

Usually, when sound is presented through headphones, it is perceived as being inside the head. Internalized perception occurs because cues that arise from the scattering process from the user's body, head and ears, are not rendered. These localization cues encoded in the head-related transfer function (HRTF) vary significantly between people. If they are added back, the sound is still perceived as non-externalized, since the reverberation and other environmental cues are missing, and, further, any cues rendered may not be consistent with the user's motion. Thus, dynamic cues must be recreated for maximum sense of presence in the virtual audio scene. All these aspects are included in our system.

Previous work in the area can be tracked back to the year 1907 [1]. Since then understanding of spatial localization [2], modeling of the involved transfer functions [3, 4], fast synthesis methods [5], and environment modeling [6, 7] have made significant progress. In addition, several studies have demonstrated the feasibility of the spatialized audio interfaces [8] and real-time spatial displays using specialized hardware have been created [9]. Our work is unique in that it creates a rich audio-visual virtual environment on a commercial off-the-shelf PC. 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 provide details of the algorithms used and of the implementation, along with a report of informal tests of the system on different people using non-individualized HRTFs. The experiments show that all users report consistent externalization and source separation with source motion. In terms of precision, the localization performance varies for different people, with vertical localization being quite accurate for some subjects, suggesting that custom-tailoring of the HRTF [10] can lead to perfect localization.

2. AUDIO SCENE RENDERING

Humans have the remarkable ability to locate a sound source with better than 5° accuracy in both azimuth and elevation, in challenging environments. Multiple cues are involved: interaural time and level differences (ITD, ILD) play a role in azimuth perception but can't explain vertical localization. Additional cues which are used in elevation perception arise from sound scattering off the listener itself (described by the HRTF). It is known that it's 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.

HRTF representation: The HRTF depends on the direction of arrival of the sound, and, for nearby sources, on the source distance, which we neglect. If the sound source is located at polar angles (φ, θ), then the (left and right) HRTFs H_l and H_r are defined as the 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 :

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

In the following we will suppress the dependence on ω. A typical slice of the HRTF for constant azimuth and varying elevation contains several peaks and valleys, which shift as the elevation changes (Fig. 1). These relative amplifications and attenuations create the perception of elevation, and are very individualized, since people's ear shapes and body sizes vary.

To synthesize the audio scene given the source location (φ, θ) one needs to filter the signal with $H(\varphi, \theta)$ and render the result binaurally through headphones. 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.

To compensate for head motion low-latency head tracking is used to stabilize the virtual audio scene. Rendering through loudspeakers [11] 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.

Support of NSF award 0086075 is gratefully acknowledged.

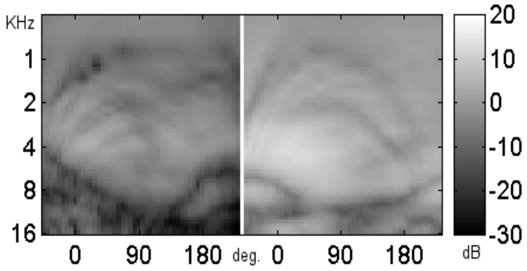

Fig. 1. HRTF slices for contralateral and ipsilateral ears for azimuth of 45 degrees and varying elevation for a human subject.

Head tracking: In addition to the static localization cues (ITD, ILD and HRTF-based), 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. 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. However, if the sound scene is presented through headphones without compensation for head 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. 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.

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.

HRTF interpolation: Currently, we use pre-measured sets of HRTFs. (Another related project deals with numerical synthesis of HRTF from ear meshes and, once completed, will eliminate the need for tedious HRTF measurements). 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 HRTF $H'(\omega)$ 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 $H'(\omega)$ is required. Simple separate interpolation of the amplitude and the phase is flawed; the reason is phase uncertainty. It is interesting to note that the phase uncertainty does not arise if the spatial sampling frequency for the HRTF is above a certain threshold determined by a Nyquist criterion. It is also noticeable that for a fixed sampling grid, HRTF interpolation over lower frequencies (longer wavelengths) will not be disturbed 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 $H'(\omega)$ can be constructed uniquely by interpolating amplitudes of $H_i(\omega)$, but the phase reconstruction is problematic. However, it is not really necessary to preserve phase information in the interpolated HRTF, as the measured phase is likely to be contaminated. The phase is of course necessary to reconstruct the ITD, but given the correct ITD, only the frequency content 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 Woodward's formula

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

The only unknown value here is the head radius r which we take to be approximately 8.9 cm in most of our simulations [12]. (We are in the process of implementing on the fly customization using video). We interpolate the amplitudes of $H_i(\omega)$ and add or subtract time delays of $\tau/2$ by setting the phase to $e^{-i\omega\tau/2}$ or $e^{i\omega\tau/2}$, respectively. (Time shift is done in frequency domain since humans are sensitive to ITD variations as small as 7 μs [13] which is 1/3 of a sampling period at the rendering rate of 44.1 kHz). The IFFT of $H_i(\omega)$ provides the desired interpolated head-related impulse response (HRIR).

The database of HRTFs used in our work includes several measurements for different people on a lattice with 10 degree step in azimuth and elevation. To compute interpolated HRIRs for a source at (φ, θ), we find the three closest lattice points $(\varphi_i, \theta_i), i = 1...3$. (The distance between lattice points is defined as a Euclidean distance on the unit sphere). Interpolation with inverse distance weighting is employed using the values at these three points. The phase is set as described earlier. The perceived sound motion is quite smooth, and no jumps or clicks are noticeable.

Room model: Using the HRTF alone to render the sound scene results in a "flat" environment where the sounds are not well externalized. Users report correct perception of azimuth and elevation, but the sound is felt to be excessively close to the head surface. If the sound source is placed close to the ear, the sound is perceived correctly. However, when the source is moved away, it feels as if the sound is still in the ear but with decreased volume. To achieve good externalization, environmental scattering cues must be incorporated. We use either a simple image model for rectangular rooms [14] or a more complicated model for arbitrary piecewise-planar room [15]. 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 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 IR. It depends upon the relative locations of the source and receiver in the room. (For computing the IR at multiple room points we presented a fast algorithm based on the multipole method in [16] and are now working on expanding it to the case of non-rectangular rooms).

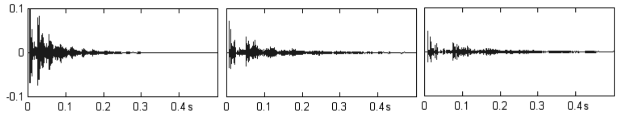

Fig. 2. A room impulse response for the audio scene scaled up by a factor 1, 2 and 3, respectively.

The image model results in a small number of relatively strong image sources from the early reflections, and very large numbers of later weaker sources (see Fig. 2). The earlier reflections will in turn be scattered by the listener's body. Thus they must be convolved with the appropriate HRIR (For example, the first reflection is usually the one from the floor, and should be perceived as such). In addition, it is believed [17] that at least the first few reflections provide additional information that help in sound localization. To accomplish this we use the HRIR for the image source direction with the correct attenuation and delay.

This approach results in a much greater computational load. Even the pulse-constituted room IR requires recomputation if the source or receiver position changes. Now, the IR composed of pasted-on HRIRs has to be recomputed even if the receiver orientation changes! Full recomputation of IR is not feasible in real-time. We adopt a combined approach where the direct signal arrival and the room reflections are put together in one finite impulse response (FIR) filter of length of approximately 100 ms. The direct path arrival and the first few reflection components of FIR are recomputed in real time and the rest of the filter is computed once for a given room geometry and materials.

FIR computation: The substantial length of the FIR filter results in delays due to convolution. To avoid delays, the convolution can be carried out in the time domain but this is inefficient, and only a short FIR can be used. In the frequency-domain the blocky nature of the convolution causes latency of at least one block. A nonuniform block partitioned convolution algorithm was proposed in [18], but this algorithm is proprietary and somewhat inefficient and difficult to optimize on regular hardware. We instead use frequency-domain convolution with short blocks ($N = 2048$ or 4096 samples) which results in tolerable delays of 50 to 100 milliseconds (at a sampling rate of 44.1 KHz). We compute in real-time the FIR filter of the same length N. However, processing only with this filter will limit the reverberation time to the filter length, while reverberation for real rooms is from 400ms (office environment) to 2 s and more (concert halls). Thus, we use an additional longer (65536 samples) FIR for simulation of reverberation. This filter contains only the reverberant tail of the room response, and the part from 0 to N in it is zeroed out.

The algorithm proceeds as follows. The FIR is 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 order L_2 . Higher order reflections are those that are too small, or are outside the chosen buffer length. The second part is precomputed at the start of the program for a given room geometry, and a set of 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 FIR and sticks it on top of the second part. As can be seen in Fig. 3, the reverberant tail of room response function stays relatively the same for different source and receiver locations in the room, which justifies our approach.

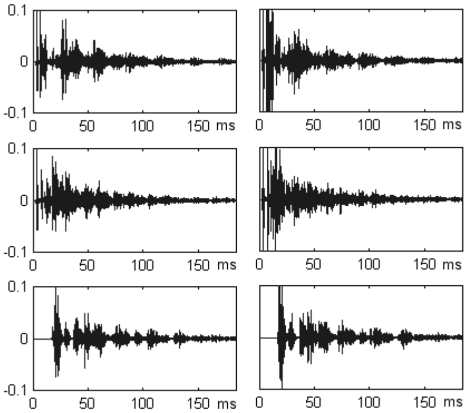

Fig. 3. Left and right IR for three different source and listener positions in the same room.

Rendering: 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 FIRs. 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.

3. SYSTEM SETUP

The current setup uses a high-end office computer which is dual Xeon P4-1.7 GHz Dell 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. 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.

A simple game with 3D-audio cues for localization was developed where spatialized sound, personalized stereo display, head tracking and multiplayer capabilities all come 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 is tracked, and appropriate panning of the video scene takes place. The rendered world stays stable in both video and audio modalities. (Video is rendered using OpenGL). The participant learns an intuitive set of commands that are conveyed by head motion. During the course of game, players navigate a virtual world and hit targets. Some targets appear for short periods of time and manifest themselves with different sounds. In this way, the audio significantly extends the user's field of regard, since, often, a new target is initially located by sound. For multiplayer capability, several PCs linked together via Ethernet render the audio and video streams for participants.

4. EXPERIMENTAL RESULTS

A number of volunteers were subjects of listening experiments. Generally, people achieve very good externalization. Reported experience varies from "I can truly believe that this box [transmitter] is making sound" to "Sound is definitely outside of my head, but elevation perception is distorted" (probably due to non-personalized HRTFs). Thus, the system was capable of making the people think that the sound is coming from the external source. Presumably, reverberation cues and highly natural changes of the audio scene with head motion and rotation create this realistic perception. Even better results should be achievable with personalized HRTFs. We will soon have a mechanism to compute personalized HRTFs using video and numerical analysis [10].

We performed informal 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 75 ms bursts of white noise with 75 ms pauses between them, repeated every second. 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 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 below.

	s1	s2	s3	s4	s5	s6
avg $\lvert\varphi\rvert$	1.7	6.3	5.1	4.3	6.4	8.0
avg $\lvert\theta\rvert$	3.1	9.0	9.5	5.5	16.7	14.4
avg φ	0.3	-5.3	4.8	2.7	3.3	4.2
avg θ	1.2	-4.0	-4.5	5.0	-9.0	-8.3

(3)

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. Subject 1 is the best localizer in both azimuth and elevation. Performance of subject 4 is also quite good. Subjects 5 and 6 perform poorly, but azimuth localization is still better than 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 mostly 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.

5. CONCLUSIONS

We have created a prototype virtual audio space rendering system which runs in real-time on a typical office PC. Static, dynamic and environmental sound localization cues are accurately reconstructed, creating highly convincing experience for participants. The system is in use and will be enhanced with several user interface projects for sighted and low-vision users in the works.

6. REFERENCES

[1] J.W.Strutt (Lord Rayleigh). "On our perception of sound direction", Phil.Mag., **13**, 1907.

[2] D. Wright, J. Hebrank, B. Wilson. "Pinna reflections as cues for localization", J. Acoustic Soc. Am., **56**, 1974.

[3] R. Duda. "Modeling head related transfer function", Proc. Asilomar conf. on Signal, Systems and Computers, 1993.

[4] E. Lopez-Poveda, R. Meddis. "A physical model of sound diffraction and reflections in the human concha", J. Acoust. Soc. Am., **100**, 1996.

[5] C. Brown, R. Duda. "A structural model for binaural sound synthesis", IEEE Trans. Speech Aud. Proc., **6**, Sep. 1998.

[6] J.-M. Jot. "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces," Multimedia Systems, vol. 7(1), 1999.

[7] B. Shinn-Cunningham. "Distance cues for virtual auditory space", IEEE-PCM2000.

[8] R. McKinley, M. Ericson, "Flight demonstration of a 3-D auditory display", in Binaural and Spatial Hearing in Real and Virtual Env., ed. by R.Gilkey and T.Anderson, 1997.

[9] M Casey, W. Gardner, S. Basu. "Vision steered beamforming and transaural rendering for the artificial life interactive video environment", Proc. Audio Eng. Soc. Conv., 1995.

[10] R. Duraiswami et al. "Creating virtual spatial audio via scientific computing and computer vision", Proc. of 140th meeting of the ASA, Newport Beach, CA, December 2000.

[11] C. Kyriakakis, P. Tsakalides, T. Holman. "Surrounded by sound: Immersive audio acquisition and rendering methods", IEEE Signal Processing Magazine, **16**, Jan 1999.

[12] R. Algazi, C. Avendano, R. Duda. "Estimation of a Spherical-Head Model from Anthropometry", J. Audio Eng. Soc., **49**, 2001.

[13] C. Kyriakakis. "Fundamental and technological limitations of immersive audio systems", Proc. IEEE, **86**, 1998.

[14] J. Allen, D. Berkeley. "Image method for efficiently simulating small-room acoustics", J. Acoust. Soc. Am., **65**, 1979.

[15] J. Borish. "Extension of the image model to arbitrary polyhedra", J. Acoust. Soc. Am., **75**, 1984.

[16] R. Duraiswami, N. Gumerov, D. Zotkin, L. Davis. "Efficient Evaluation of Reverberant Sound Fields", Proc. WASPAA01, New Paltz, NY, October 2001.

[17] B. Rakerd, W. Hartmann. "Localization of sound in rooms, II: The effects of a single reflecting surface", J. Acoust. Soc. Am., **78**, Aug. 1985.

[18] W. Gardner, "Efficient Convolution without Input-Output Delay". J. Audio Eng. Soc., vol. 43(3), March 1995.