Virtual audio system customization using visual matching of ear parameters

Dmitry N. Zotkin, Ramani Duraiswami, Larry S. Davis, Ankur Mohan, Vikas Raykar
Perceptial Interface and Reality Lab, UMIACS
University of Maryland at College Park
College Park, MD 20740 USA
E-mail: {dz,ramani,lsd,ankur,vikas}@umiacs.umd.edu

Abstract

Applications in the creation of virtual auditory spaces (VAS) and sonification require individualized head related transfer functions (HRTFs) for perceptual fidelity. HRTFs exhibit significant variation from person to person due to differences between their pinnae, and their body sizes. In this paper we propose and preliminarily implement a simple HRTF customization based on use of a recently published database of HRTFs [1] that also contains geometrical measurements of subject pinnae. We measure some of these features via simple image processing, and select the HRTF that has features most closely corresponding to the individual's features. This selection procedure is implemented along with the virtual auditory system described in [2], and listener tests conducted comparing the "customized" HRTF and a fixed HRTF. Despite the simplicity of the method, tests reveal average improvement in localization accuracy of about 25 percent, though performance improvement varies with source location and individuals.

1 Head Related Transfer Function

Using just two receivers (ears), humans are able to localize sound with amazing precision [3]. While differences in the time of arrival or level between the signals reaching the two ears (known respectively as interaural time delay, ITD, and interaural level difference, ILD) [4] can partially explain this facility, these 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 distinguish between them. The 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 human body, and for high audible frequencies, the external ear. As a result, the circularly-asymmetric external ear essentially forms a specially-shaped "antenna" that causes a direction-of-arrival (DOA) dependent "filtering" of the sound reaching the eardrums. Thus, scattering of sound by the human body and by the external ears provides additional monaural (and, to a lesser extent, binaural) cues to source position.

The effect of this scattering can be described by a frequency response function called the head-related transfer function (HRTF). For a particular source location, the HRTF is defined as the ratio of the sound pressure level (SPL) at the eardrum to the SPL at the location of center of the head as if the listener is absent. Knowing the HRTF, one can reconstruct the exact pressure waveforms that would reach a listener's ears for any arbitrary source signal arising from the given location, which is a sufficient stimulus for correct perception: if the correct sound signals are delivered to the eardrums, he will perceive a sound source at the correct location in exocentric space.

The HRTF complexity is due to the complex shapes of the pinna. The HRTF varies significantly between individuals as their ear-shapes and to a smaller extent their head and body sizes also vary. While it is possible to obtain HRTFs via direct measurements as it is done for the database [1] acquisition, this is relatively tedious and requires access to specialized experimental facilities. As an alternative, we have recently begun a project which is aimed at the direct computation of the HRTF using three-dimensional ear mesh obtained by computer vision and solving the physical wave propagation equation in the presence of a non-rigid boundary by fast numerical methods [5]. However this work is still under development, and current virtual auditory systems do not have yet any methods for customization of the HRTF. In this paper we seek to customize the HRTF using a database containing the measured HRTFs for 43 subjects along with some anthropometric measurements [1], [6].

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. A sample HRTF showing elevation-dependent changes in sound frequency content is plotted in the Figure 1. The elevation of the sound source in the plot rises from -45° to 225°, and the az-

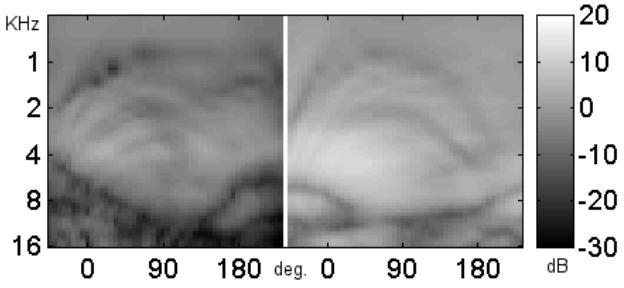

Figure 1. Sample HRTF slice (contralateral and ipsilateral ears, respectively) for the azimuth of 45 degrees and varying elevation for a human subject.

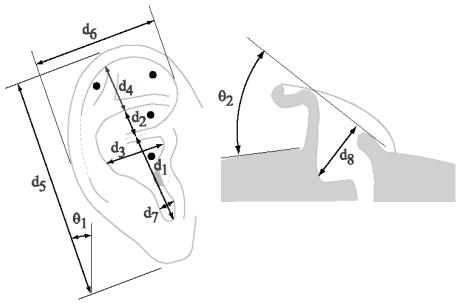

Figure 2. The set of measurements provided with the HRTF database.

imuth is constant. (In the interaural-polar coordinate system used, azimuth $\varphi \in [-90, 90]$ and elevation $\theta \in [-90, 270]$. Points in back of the subject have $\theta = 180°$). The effects of the different body parts show up in different frequency ranges. The head shadow explains the overall level difference in the two pictures; the torso reflections create wide arches in the lower frequency area and the *pinna notches* appear as dark streaks in the high-frequency regions. Feature positions in frequency change with elevation; it is these cues that are used by us to distinguish elevations [7], [8].

The initial version of our virtual auditory system [2] employs dynamic room models, head tracking, and runs on an off the shelf PC. Most users report very satisfactory experience using the system even with a single measured HRTF. 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 can lead to localization performance comparable to localization with person's own HRTF set.

2 Database matching

Usually, when the waveform is played back for a listener via headphones, the sound appears flat and is perceived as being inside the head. This happens for three reasons. One of them is that the rendered sound lacks those *HRTF-based cues* described above that tell the brain that the sound is "processed" (passively filtered) by ears and therefore has originated from the outside. The second reason is that very important *dynamic cues* are missing (that is, when the listener rotates the head the sound scene does not change, which is not the case for the outside sound since the DOA of the wave changes for the listener). The last important set of cues are the *environmental cues* which are essentially the reflections coming off the room walls and other surfaces, creating reverberation which is important for externalization and distance perception. Provided that the set of HRTF is known, it is easy to create a *virtual audio environment* – a synthesized playback stream which is delivered through headphones but nevertheless appears for the listener to originate from some point or object in 3D-space. In the VAS system, the HRTF-based cues are created by filtering the sound with the appropriate HRTF for the DOA for that sound source, the dynamic cues are implemented by active head tracking using commercially-available tracking systems, and the environmental cues are modeled by a simple multiple reflection model.

The biggest and still-open problem in the synthesis of the virtual auditory spaces is the customization of the HRTF for a particular individual. Each person presumably learns her own HRTF given visual feedback about the source position, but the HRTFs of different people look very different in the plot and are not interchangeable. It is known [9], [10] that differences between individual HRTFs due to differences in ear shape and geometry strongly distort the perception when non-individualized HRTFs are used for rendering, and that the high-quality synthesis of a virtual audio scene requires the personalization of the HRTF for the particular individual for good virtual source localization. The usual customization method is a direct measurement of the HRTF when a tiny microphone is put into the ear canal of the subject and the sound is played through a loudspeaker positioned sequentially over all possible DOA angles in some steps, sampling the whole sphere. This method is accurate but is highly time-consuming. A faster but less accurate approach, which we report on in this paper, is an attempt to select the best-matching HRTF from an existing database of HRTFs and use it for the synthesis of the VAS, thus making the HRTF semi-personalized.

We pose the problem as selecting the most appropriate HRTF from database of HRTFs. The database we used was recently released by the CIPIC lab and contains the measurement of the HRTFs of 43 people, along with the anthropometric information about those subjects. The anthro-

pometric 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 Figure 2 and are as follows: $d_1 ... d_8$ are cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, intertragal incisure width and cavum concha depth, and θ_1 and θ_2 are pinna rotation and flare angles, respectively. For the HRTF matching procedure, we use 7 of these 10 pinna parameters which can be easily measured from the ear picture.

Since the HRTF is the representation of the physical process of the interaction between the oncoming sound wave and the listener pinnae, head and torso, it is natural to assume the hypothesis that the structure of the HRTF is related to body parameters. For example, observe that if the ear is scaled up, the HRTF will maintain its shape but will be shifted toward the lower frequencies on the frequency axis. Since the listener presumably deduces the source elevation from the positions of peaks and notches in the oncoming sound spectrum, usage of the HRTF from the scaled-up ear will result in systematic bias in the elevation estimation. Some studies, such as a structural model for composition and decomposition of HRTF [11] and experiments with HRTF scaling ([12], [13], [14]) already suggested that the hypothesis is somewhat valid, although a perfect match (equivalent to the localization with the person's own HRTF) was not achieved with somebody's else HRTF appropriately scaled up or down. However, the ears of different persons are different in much more ways than just scaling, and seemingly insignificant change in ear shape can cause dramatic changes in HRTF.

We perform an exploratory study on the hypothesis that the HRTF structure is related to the ear parameters. Specifically, given the database of the HRTFs of 43 persons along with their ear measurements we select the closest match to the new person by taking the picture of her ear, measuring the d_i parameters from the image and finding the best match in the database. If the measured value of the parameter is $\hat{d}_i$ and the database value is d_i, then the parameter error $e_i = (\hat{d}_i - d_i)/d_i$, 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 very rare though.

We have developed a graphical interface for fast selection of the best-matching HRTF from the database. The pictures of the left and the right ear of the new virtual audio system user are taken with two cameras (a sample is shown in Figure 3). An operator identifies key points on the ear picture and measures the ear parameters described above. The parameters d_8 and θ_2 are not measured since they can't be reliably estimated from pictures and θ_1 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 using only the ear picture, and no listening tests are necessary.

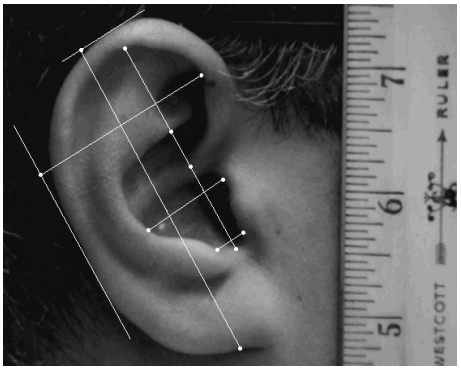

Figure 3. Sample picture of the ear with the measurement control points marked.

3 Results

We performed a series 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. The audio rendering system is described in detail in [2]. The system is based on a fast dual-processor PC with no specialized hardware, except for the head tracker. The test sounds are presented through headphones, and the head tracker measures the head position when the subject "points" to the virtual sound source with her nose. The test sound was three 75ms bursts of white noise with 75 ms pauses between them, repeated every second.

We performed two series of the experiments with the same 6 subjects. In the first series, the "generic" HRTF was used for the VAS rendering, which was taken to be the HRTF of a real person measured in an anechoic chamber. That person was not among the test subjects. In the second series, the best-matching HRTF was selected from the database and used for VAS rendering. The test sessions themselves consisted of a short training period and the real test, when the test sound described above is played in some location in the space with the azimuth $\varphi \in [-90, 90]$ and the elevation $\theta \in [-45, 45]$, and the subject is asked to point at the location of the source with her nose. The perfect localization corresponds to $\varphi = 0, \theta = 0$ in the coordinate system linked to the subject's head. On localization, the subjects presses the button on the keyboard and is presented with the next source. The series consists of 20 randomly selected positions. The results for the generic HRTF are presented in the Table 1, and the Table 2 contains the results when the HRTF that best matches that subject's ear

parameters are used. For both tables, the average value of the modulus of error in both azimuth and elevation and the average value of the azimuth and elevation error themselves (the bias) are reported.

Table 1

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

Table 1 corresponds to the results obtained with one generic HRTF for all six subjects. Some subjects perform better than the others, and localization in azimuth is generally better than in elevation. Considering elevational localization (which is believed to be hampered most by using of non-individualized HRTF), subject 3 performs quite good; performance of subjects 1, 2 and 6 is close to the average and subjects 4 and 5 perform poorly.

Table 2

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

Table 2 results are for the case of the best-matching HRTF from the HRTF database. Azimuthal localization was not the priority task for the subjects for this test. It is clear that the elevation localization performance is improved consistently by 20-30% for 4 out of 6 subjects. Improvement for the subject 5 is marginal and subject 6 performs actually worse with the customized HRTF.

The objective performance criteria agrees with the subjective performance estimated by subjects themselves. Subjects 1 through 4 reported that they are able to better feel the sound source motion in the median plane and the VAS synthesized with the personalized HRTF sounds better (better externalization and better perception of DOA and source distance). Subject 5 reported that motion can't be perceived reliably both with generic and customized HRTF, which agrees with experimental data. Subject 6 also reports that the generic HRTF just "sounds better".

4 Conclusions and future work

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 "fits" better or worse. These two measures correlate well, and if the customized HRTF does not "sound" good for a user a switchback to the generic HRTF can be made easily.

The performed customization is a coarse "nearest-neighbor" approach, and the HRTF certainly depends on much more than the 7 parameters measured. Still, even with such a limited parameter space the approach is shown to achieve good performance gain.

For performing statistically valid customization we must develop a predictive theory that can link anthropometric measurements to specific features. Such an approach can perhaps be developed using accurate modeling of the ear shape and numerical solution of the wave propagation equation. This approach is also a subject of active research work both in our lab and at other institutions worldwide.

References

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

[2] D. N. Zotkin, R. Duraiswami, and L. S. Davis. "Creation of Virtual Auditory Spaces", Accepted, IEEE ICASSP 2002, Orlando, FL, May 2002.

[3] W. M. Hartmann. "How We Localize Sound", Physics Today, November 1999, pp. 24-29.

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

[5] R. Duraiswami et al. "Creating virtual spatial audio via scientific computing and computer vision", Proc. 140th ASA conf., Newport Beach, CA, 2000, p. 2597, available at http://www.acoustics.org/press/140th/duraiswami.htm.

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

[7] H. L. Han. "Measuring a dummy head in search of pinnae cues", J. Audio Eng. Society, 42(1):15-37, 1994.

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

[9] M. B. Gardner and R. S. Gardner. "Problem of localization in the median plane: effect of pinna cavity occlusion", J. Acoust. Soc. Am., 53(2):400-408, 1973.

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

[11] C. P. Brown and R. O. Duda. "A structural model for binaural sound synthesis", IEEE Trans. on Speech and Audio Processing, 6(5):476-488, 1998.

[12] J. C. Middlebrooks. "Individual differences in external-ear transfer functions reduced by scaling in frequency", J. Acoust. Soc. Am., 106(3):1480-1492, 1999.

[13] J. C. Middlebrooks. "Virtual localization improved by scaling non-individualized external-ear transfer functions in frequency", J. Acoust. Soc. Am., 106(3):1493-1510, 1999.

[14] J. C. Middlebrooks, E. A. Macpherson, Z. A. Onsan. "Psychophysical customization of directional transfer functions for virtual sound localization", J. Acoust. Soc. Am., 108(6):3088-3091, 2000.