HRTF PERSONALIZATION USING ANTHROPOMETRIC MEASUREMENTS

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

Perceptual Interfaces and Reality Laboratory
Institute for Advanced Computer Studies (UMIACS)
University of Maryland at College Park
College Park, MD 20742
dz@umiacs.umd.edu

ABSTRACT

Individualized head related transfer functions (HRTFs) are needed for accurate rendering of spatial audio, which is important in many applications. Since these are relatively tedious to acquire, they may not be acceptable for some applications. A number of studies have sought to perform simple customization of the HRTF. We propose and test a strategy for HRTF personalization, based on matching certain anthropometric ear parameters with the HRTF database, and incorporation of a low-frequency "head-and-torso" model. We present preliminary tests aimed at evaluation of this customization. Results show that the approach improves both the accuracy of the localization and subjective perception of the virtual auditory scene.

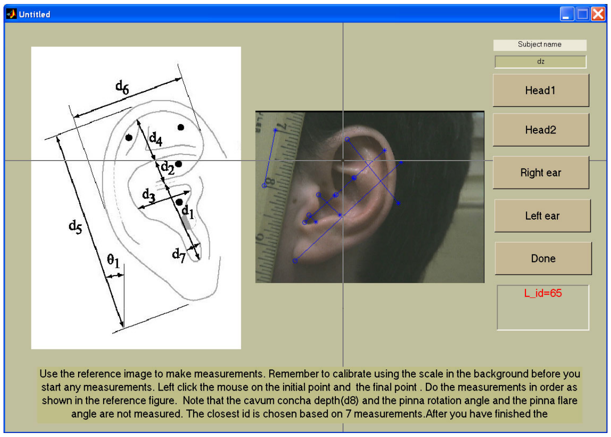
Partial support of NSF award #0086075 and ONR grant N000140110571 is gratefully acknowledged.

1. INTRODUCTION

In many spatial audio applications, it is desirable to have the ability to roughly customize the application to the needs of the user quickly. If a new user of an entertainment system has to wait several minutes or hours before he can use the system, the impression will be that the technology is not yet ready, and/or the user may lose interest. Accordingly, fast methods of spatial audio customization ought to be developed.

Humans are self-trained to localize sounds using their ears starting at birth and localize well even in adverse conditions [1]. It is known that sound scattering by the listener's anatomy plays an important role in sound localization. As the sound scatters off the human torso, head, and outer ear (pinna) characteristic changes in the received sound spectrum of familiar sounds are heard by the listener. These changes depend on the source azimuth φ, elevation θ and range ρ, and encode them. The head-related transfer function (HRTF), which is the ratio of the Fourier transform of the signal at the listener's eardrum to that at the center of the listener's head with the listener absent, characterizes these listener induced changes.

If the HRTF is known, it is simple to synthesize a virtual auditory scene (VAS) that gives the listener the impression of the sound sources being presented in exocentric space. In a static anechoic environment, filtering of the source signal with the HRTF for a given direction delivers to the listener's eardrum exactly the same acoustic pressure waves as the true source in the same environment would produce. By including reverberation and motion cues due to ego-motion of the listener, one can synthesize more realistic environments [2].



Subject name: dz

L_id=65

Use the reference image to make measurements. Remember to calibrate using the scale in the background before you start any measurements. Left click the mouse on the initial point and the final point . Do the measurements in order as shown in the reference figure. Note that the cavum concha depth(d8) and the pinna rotation angle and the pinna flare angle are not measured. The closest id is chosen based on 7 measurements.After you have finished the

Fig. 1. Screenshot of the HRTF customization software.

However, individual differences in the anatomy, especially the shape of the outer ear does not allow use of the same HRTF for all users. The ears, body and head sizes and shapes, vary a lot between people. This means that the pattern of characteristic changes in the spectrum that signifies sound location also vary between individuals, and the HRTF of one individual will create a significantly distorted perception for another one. However, several papers by Middlebrooks et al [3] [4] [5] suggest that one can expect a certain level of correlation between similarity of pinna shapes and similarity of HRTFs between individuals, and it is also reasonable to expect that similarly shaped ears that differ only by a scale factor produce similarly shaped HRTFs shifted in frequency. In the same general shape-to-HRTF correlation framework, Jin et al. [6] explicitly look at individualization of the virtual auditory spaces using morphological measurements. We build upon these approaches by matching the individual against a pre-acquired HRTF database that includes anthropometry. Other personalization methods in addition to ones mentioned above are the direct HRTF measurement using moving speaker and in-the-ear microphone (which is the most accurate but the most time-consuming method), numerical solution of the wave scattering equations [7], [8] and user-aided navigation through large HRTF design choice parameters [9]).

In this paper, we expand and further evaluate the HRTF personalization method based on anthropometric features of the outer ear suggested in [10]. The method is based on finding the best match to the outer ear shape of an individual using a set of seven distances between easily identified anatomic ear features. The CIPIC database [11] contains HRTF measurements of 45 individuals along with their anthropometry, including the ear parameters mentioned. We take a digital image of the ear of an individual, identify the anatomic features and find the best match in the CIPIC database. We perform experimental evaluation of the method, asking the subjects to localize a virtual sound source using: a) generic HRTF (which we choose to be the KEMAR HRTF); b) personalized HRTF (HRTF of the best-matched CIPIC database subject); and c) personalized HRTF amended accordingly to the "snowman" (head-and-torso) model [12]. Experiments show the effectiveness of both personalization and HAT model incorporation.

2. HRTF PERSONALIZATION

We perform HRTF customization based on a digital image of the ear taken by a video camera. A screenshot of our customization software is shown in Figure 1. On the left a reference image is shown, with the seven measurements identified as $d_1, \ldots, d_7$ (they are, in order, cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width and intertragal incisure width). On the right, the operator acquires images of the left ear, right ear, and frontal and side views of the body, and marks feature points on the images and a one-inch interval on a ruler in the image. If $\hat{d}_i$, $i = 1...7$, is the value of the i^{th} parameter in the image, and d_i^k is the value of the same parameter for the k^{th} subject of the database, then the matching is performed by minimizing the measure over all k subjects

$$E^k = \sum_{i=1}^{7} \frac{(\hat{d}_i - d_i^k)^2}{\sigma_i^2}.$$

Here σ_i^2 is the variance of the i^{th} parameter across all subjects in the database. Subject k, $k = \arg\min_k E^k$, is chosen to be the best match. In the case shown, the best matching subject for the left ear is subject ID 45. We had verified accuracy of the measurement software by comparing the obtained values of ear parameters with the physical ear measurements done with a ruler. Note that the matching is done separately for the two ears, which sometimes results in different matching subjects for the left and right ears because of individual anatomical asymmetries.

3. HAT MODEL

Direct HRTF measurements have always been problematic at low frequencies (see [12]). A simple head-and-torso model was proposed in [12] to compensate for these deficiencies. In the model, the human body is approximated by a "snowman" consisting of a spherical head and a spherical torso, separated by a certain distance (neck). Analytical HRTF computation is possible with this model. The model of course lacks all the high-frequency features introduced by pinna, but the effects of pinna, head and torso are separable on frequency axis (an object contributes to scattering only when the object size and the wave length are comparable) as experimentally shown in [13].

The HAT model uses two different algorithms depending on whether the source is located inside or outside of the torso shadow cone (Figure 2). The source outside the shadow has both a direct path from the source to the ear, and an indirect "shoulder bounce" path. The source inside the torso shadow must diffract around the body to reach the ear. All propagation paths, to the contralateral ear are also potentially shadowed by the head. We compute HAT model HRTF $H_h(\omega)$ following the algorithm in [12]. We use the phase of $H_h(\omega)$ for all ω and gradually blend the log-magnitudes of $H_h(\omega)$ and the HRTF $H_c(\omega)$ selected from the CIPIC database to get the combined HRTF $H_s(\omega)$:

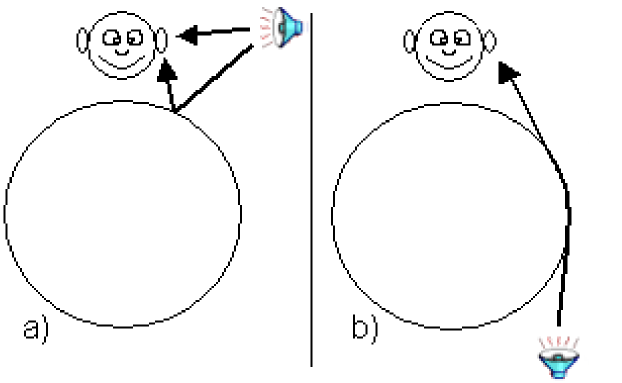


Fig. 2. a) Sound propagation path simulated by the HAT model in case of the source being outside of the torso shadow; b) source is now inside the torso shadow.

$$\begin{aligned}
A_s(\omega) &= \begin{cases} A_h(\omega) & \omega < \omega_l \\ A_h(\omega) + \frac{A_c(\omega) - A_h(\omega)}{\omega_h - \omega_l}(\omega - \omega_l) & \omega_l < \omega < \omega_h \\ A_c(\omega) & \omega > \omega_h \end{cases} \\
A_s(\omega) &= \log|H_s(\omega)|, \quad A_h(\omega) = \log|H_h(\omega)|, \\
A_c(\omega) &= \log|H_c(\omega)|, \quad \omega_l = 500Hz, \quad \omega_h = 3000Hz.
\end{aligned}$$

Thus, the HAT model HRTF magnitude is used below 500 Hz. From 500 Hz to 3 kHz the database HRTF is progressively blended in and above 3 kHz the HAT model magnitude is not used.

4. EXPERIMENT DESCRIPTION

We performed a series of experiments to verify whether the HRTF customization method and incorporation of the HAT model improves the localization ability of the individual and subjective quality of the audio scene. Nine subjects were selected to participate, all reporting no problems with their hearing. (One of the subject's data is not presented here because of high localization bias and inconsistent results, which suggests some error in the system setup for that experiment). Each subject had their pictures (the front view of the upper body, the side view of the upper body, and a detailed close up on the right and left ears) taken. All pictures included a scale to relate the known length in world to image measurements. From these images, seven ear measurements for the HRTF database matching and three body measurements for the HAT model computations are obtained. The torso radius, head radius, as well as the neck length were measured. The torso radius is measured as half the length between the left and right shoulders. The head radius is half the length between the two ears. The neck length was taken as the length between the chin and the collar bone. The best-matching database subject's HRTF(s) for the left and for the right ears is selected and will be referred to as "personalized" (PRS) HRTF. It is also modified at low frequencies in accordance with the HAT model using the body measurements to create a "personalized-plus-snowman" (PPS) HRTF.

	M_φ	M_θ	E_φ	E_θ	b_φ	b_θ
s1 GEN	2.81	8.50	3.61	11.24	1.81	2.75
s1 PRS	3.69	6.56	4.41	8.81	3.69	4.31
s1 PPS	7.31	8.31	7.58	9.81	7.31	1.06
s2 GEN	3.06	10.88	4.07	14.38	-1.44	-10.13
s2 PRS	4.69	10.31	6.56	13.69	-1.94	-0.19
s2 PPS	6.00	13.38	7.39	16.56	-3.50	-7.50
s3 GEN	5.63	10.31	6.51	12.54	-5.63	4.06
s3 PRS	4.31	8.44	4.89	10.26	-2.69	4.94
s3 PPS	3.50	7.81	4.21	11.04	2.75	-5.19
s4 GEN	3.13	5.94	3.54	6.69	-2.75	-5.94
s4 PRS	3.13	3.13	3.79	3.94	2.75	-2.25
s4 PPS	2.69	3.88	3.46	4.72	1.81	-2.88
s5 GEN	1.63	14.94	2.06	15.81	-0.63	-14.94
s5 PRS	1.38	5.81	1.77	7.15	-1.13	-4.69
s5 PPS	3.69	7.25	4.38	8.60	-3.69	-4.88
s6 GEN	2.00	11.75	2.29	13.66	0.88	-9.50
s6 PRS	5.75	14.50	6.33	18.06	5.38	-8.00
s6 PPS	5.81	8.13	6.75	9.31	-5.81	-7.38
s7 GEN	4.81	6.50	5.25	7.61	-4.31	-3.38
s7 PRS	2.69	10.94	3.44	13.12	0.81	6.06
s7 PPS	4.38	14.00	4.86	20.17	1.88	11.38
s8 GEN	5.25	8.38	5.80	8.59	-4.88	8.38
s8 PRS	11.81	5.44	12.60	6.85	-11.81	3.94
s8 PPS	9.75	6.88	10.67	7.61	-9.75	4.13

Table 1. Absolute error (M), r.m.s. error (E) and bias (b) for continuous stimulus

We use KEMAR-with-small-pinna HRTF (also from CIPIC database) as a "generic" (GEN) HRTF, and we apply the HAT model to it with KEMAR head radius, torso radius and neck height (8.7 cm, 16.9 cm and 5.3 cm, respectively) creating "generic-plus-snowman" (GPS) HRTF.

For the listening test, the subjects sits down in the chair and puts on the headphones with the head tracker attached. We used our real-time spatial audio synthesis software described in [2] for synthesis and deliver the stimulus through Sennheiser HD-470 headphones. The system latency was under 100 ms and no headphone compensation was done. Before the test is started, instructions are given to the subjects. They are told that the sound is emitted from a virtual sound source located at a randomly chosen point in space in front (azimuth $\varphi \in [-90^\circ, 90^\circ]$, elevation $\theta \in [-45^\circ, 65^\circ]$), at a distance of 1 meter. The sound is heard through the headphones and they are to turn their head in order to "look" in the direction of the sound (point at it with their nose). Then, the headphone position on subject's head is calibrated by placing a visual marker directly in front of the subject, producing the virtual sound from the marker's location and asking the subject to point his nose at this marker. The headphone position is adjusted then to read zero in both azimuth and elevation. After this calibration, the actual test starts. The sound is played through headphones, and the subject points to the sound source and hits the space bar to record the perceived position of the sound. The head tracker records the position of the head. The error in azimuth and in elevation is found as a difference between true virtual source location and the pointed-to location. The next sound is then emitted at a different random position.

	M_φ	M_θ	E_φ	E_θ	b_φ	b_θ
s1 GEN	17.88	9.25	21.45	10.99	3.63	-0.50
s1 GPS	N/A	N/A	N/A	N/A	N/A	N/A
s1 PRS	21.06	12.94	24.00	14.53	4.06	-1.44
s1 PPS	19.88	8.88	22.00	11.61	1.00	-4.00
s2 GEN	27.13	12.75	33.83	15.72	7.13	-4.00
s2 GPS	N/A	N/A	N/A	N/A	N/A	N/A
s2 PRS	28.63	11.56	36.64	14.55	6.75	-5.94
s2 PPS	19.44	12.50	24.96	18.83	-5.94	-7.88
s3 GEN	9.94	12.69	12.35	15.86	7.06	-11.19
s3 GPS	7.06	11.31	9.08	13.77	3.56	-9.81
s3 PRS	10.88	12.50	12.74	14.19	6.88	-6.88
s3 PPS	5.25	10.13	6.35	13.62	4.13	-8.00
s4 GEN	13.25	14.13	17.13	16.47	-4.00	-5.38
s4 GPS	7.50	9.38	9.14	10.94	-5.63	-3.63
s4 PRS	9.38	11.94	11.22	15.45	3.38	1.06
s4 PPS	10.38	9.94	11.66	12.72	-2.50	-5.06
s5 GEN	20.94	13.94	23.00	17.68	3.44	-9.69
s5 GPS	5.75	10.13	6.75	12.45	1.75	-6.88
s5 PRS	8.88	9.75	9.76	12.23	-3.50	-7.50
s5 PPS	6.19	6.31	7.43	8.13	1.06	-5.56
s6 GEN	12.94	17.88	15.99	23.04	-7.81	-16.38
s6 GPS	N/A	N/A	N/A	N/A	N/A	N/A
s6 PRS	10.44	15.75	12.13	19.12	-8.69	-12.50
s6 PPS	13.19	13.63	16.78	16.53	-7.81	-5.88
s7 GEN	18.81	9.00	23.95	11.68	1.94	-0.88
s7 GPS	N/A	N/A	N/A	N/A	N/A	N/A
s7 PRS	22.50	17.00	28.18	19.75	2.13	0.00
s7 PPS	18.88	12.38	23.55	15.92	-9.50	-4.38
s8 GEN	7.63	8.25	9.29	10.51	-1.50	-4.75
s8 GPS	7.13	7.44	9.47	9.86	3.38	-2.31
s8 PRS	5.81	8.25	8.63	10.52	-1.31	2.00
s8 PPS	3.38	9.13	4.87	10.34	-1.75	3.75

Table 2. Performance data, as in table 1, for short bursts

The subject repeatedly listens to and locates the sound for 3 series of about 30 localization attempts for each tested HRTF. We used two types of signals: a continuous sound that stays on during the whole task, and single sound bursts. The first task is essentially more like "centering" on the sound rather than localization, and the second task is the true localization task. A single sound burst consists of 93 ms of white noise repeated three times with 93 ms pauses; in continuous mode, these single bursts repeat every second. The experimental setup and the response procedure of the subjects are described above and are identical for both tasks. Three different tests are performed for each type of the sound using GEN, PRS and PPS HRTFs. For some of the subjects, we also tested GPS HRTF for the single burst sounds.

5. RESULTS

We show results by the subject and by the HRTF tested in the Table 1 for continuous and in the Table 2 for short burst stimulus.

In the tables, s1 – s8 are the subject numbers, and columns are as follows: M_φ and M_θ are the average absolute values of the localization errors in azimuth and elevation, E_φ and E_θ are the root mean square errors in azimuth and elevation, and b_φ and b_θ are the localization biases in azimuth and elevation, respectively.

Table 1 shows that, generally, for continuous stimulus personalization somewhat helps to improve localization in elevation (which is believed to be hampered most by using non-individualized HRTF), but not for everybody. For 5 subjects out of 8, the PRS results are better than the GEN. Subject 2 shows no improvement, and subject 7 performance worsens when personalization is performed. The performance decrease can be attributed to the imperfections of matching or to the tiredness of the subject. Subject 6 showed a decrease in performance when going from GEN to PRS, but an increase again when the HAT model was added; it is possible that low-frequency cues played more prominent role for that person than for everybody else. Localization error in azimuth does not seem to exhibit any sort of regular pattern.

In Table 2, the results are presented for the short stimulus (which is the "true" localization task). We will consider first localization in elevation. Now, for almost all of the subjects, incorporation of the HAT model improves performance more than personalization. For subjects 1 and 8, the localization worsens when going from GEN to PRS but improves again in PPS case (with the HAT model). Personalization still seems to help for about half of the subjects, and for the other ones there is either no change, or a degradation of performance. For some of the subjects, we tested GPS (generic-with-snowman) HRTF there, which also seems to improve performance for all subjects tested (unlike personalization, which does not always work well). Error in azimuth also decreases as personalization and HAT model incorporation are done.

The elevation localization error M_θ averaged over 8 subjects is, per HRTF in degrees, GEN 9.65, PRS 8.14, PPS 8.70 for table 1 and GEN 12.24, PRS 12.46, PPS 10.36 for table 2. Indeed, it can be expected that the HAT model will be important for the short bursts but will not matter much for continuous sound. In the continuous sound localization, the source is essentially staying on-axis in the final stages of localization. When short bursts are used, the source could be located anywhere on the sphere, and it is known that the low-frequency localization cues introduced by the HAT model help localization only for off-axis sources [14]. Thus, we expect that HAT model would not increase performance for on-axis sources but might help for short off-axis stimulus presentation. The experimental data support these expectations. It is also possible that even if the personalization is misfit, the HAT model still provides correct localization cues (it was reported by one of the subjects that the low-frequency components of the GPS and PPS HRTFs have "depth", sound more "focused" and the bass part of the noise provide some sort of "stability" to the sound and removes ambiguities in the source position). HAT model also seems to improve localization in azimuth in the short stimulus case, which is probably due to the incorporation of the correctly personalized ITD cues computed by HAT model using measured head radius.

Usually in audio user interfaces the localization task involves responding quickly to a short sound stimulus outside the field of view. Therefore, short burst localization (table 2 data) would probably be more important, compared to the results of the continuous sound localization. We thus conclude that incorporation of the low-frequency localization cues provided by the head-and-torso model is desirable for rendering with non-personalized HRTFs.

6. CONCLUSIONS

We performed the experiments to compare virtual sound source localization performance using the generic (KEMAR) HRTF, the HRTF personalized based on pinna anthropometry, and the HRTF modified in accordance with the "snowman" (head-and-torso) model. The generic trend observed is that incorporation of the HAT model almost always improves localization performance, whereas the ear parameters based personalization method does not always perform well. In ongoing work we are performing a comparison of the different methods of quick customization proposed in the literature. We suggest that researchers and software developers use the HAT model to both improve subjective quality of the audio scene and help correct localization of the acoustic signals in virtual audio environments. We plan to further pursue the development of fast HRTF measurement and customization methods in the near future.

7. REFERENCES

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

[2] D. N. Zotkin, R. Duraiswami, and L. S. Davis. "Rendering localized spatial audio in a virtual auditory space", IEEE Trans. on Multimedia, in press.

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

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

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

[6] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile (2000). "Enabling individualized virtual auditory space using morphological measurements", Proceedings of the First IEEE Pacific-Rim Conf. on Multimedia, pp. 235-238.

[7] Y. Kahana (2000). "Numerical modelling of the head-related transfer function", Ph.D. Thesis, ISVR, University of Southampton, UK.

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

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

[10] D. N. Zotkin, R. Duraiswami, L. S. Davis, A. Mohan, V. Raykar (2002). "Virtual audio system customization using visual matching of ear parameters", Proc. IEEE ICPR 2002, Quebec City, Canada, pp. 1003-1006.

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

[12] V. R. Algazi, R. O. Duda, D. M. Thompson (2002). "The use of head-and-torso models for improved spatial sound synthesis", Proc. AES 113th Convention, Los Angeles, CA, preprint 5712.

[13] V. R. Algazi, R. O. Duda, R. P. Morrison, D. M. Thompson (2001). "Structural composition and decomposition of HRTF's", IEEE WASPAA01, New Paltz, NY, pp. 103-106.

[14] V. R. Algazi, C. Avendano, and R. O. Duda. "Elevation localization and head-related transfer function analysis at low frequencies", J. Acoust. Soc. Am., vol. 109(3), pp. 1110-1122.