PROJECT REPORT ON STUDY OF CIPIC HRTF DATABASE

BY

MANU MITRA 0795410

UNIVERSITY OF BRIDGEPORT

2008-2009

Name : - *Manu Mitra*

Student Id : - *0795410*

Major : - *Electrical Engineering*

Class : - *Masters Project*

Subject : - *Project Report on Study of CIPIC HRTF Database*

Submission : - *Oct 13th 2009*

Comments:-

Marks Obtained *Professor's Signature.*

Contents

Study of CIPC HRTF Database

1. Abstract……..1

2. Introduction……...1

3. Properties of HRTF……...2

4. Anthropometry……...3

5. HRTF Personalization……...6

6. HAT Model……7

7. HRTF Variation……...13

8. Composition of the HRTF……...15

9. Decomposition……17

10. Feature Extraction……..21

11. Result……23

12. Reference……...26

Study of CIPC HRTF Database

1. Abstract:

This project describes a public-domain database of high-spatial-resolution head-related transfer functions measured. Head Related Transfer Function (HRTF) characterizes the auditory cues created by scattering of sound off a person's anatomy. While it is known that features in the HRTF can be associated with various phenomena, such as head diffraction, head and torso reflection, knee reflection and pinna resonances and anti-resonances, identification of these phenomena are usually qualitative and/or heuristic. The objective of this project is to attempt to decompose the HRTF and extract significant features that are perceptually important for source localization. Some of the significant features that have been identified are the pinna resonances and the notches in the spectrum caused by various parts of the body.

2. Introduction:

Humans are self trained to localize sounds using their ears starting at birth and localize well even in adverse conditions. 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, elevation, range 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.

Head-related transfer functions (HRTFs) capture the sound localization cues created by the scattering of incident sound waves by the body, and play a central role in spatial audio systems. Most HRTF-based commercial system convolves the input signal with a single, "standard" head-related impulse response (HRIR), and several studies have employed the public-domain dataset for the KEMAR mannequin. However, it is well known that HRTFs vary significantly from person

to person and that serious perceptual distortion (particularly front/back confusion and elevation error) can occur when one listens to sounds spatialized with a non individualized HRTF.

Although the determination of individual HRTFs can addressed in a number of ways, most recently by numerical computations based on a detailed geometric mesh of the human body, the study of individual variations requires a database of uniformly measured HRTFs. Several laboratories have developed HRTF databases to support their own research. However, the only publicly available database is the AUDIS catalog, which is limited to 12 subjects measured at approximately 120 positions in space, and cannot be used for commercial purposes.

The CPIC Interface Laboratory at U.C. Davis has measured HRTFs at high spatial resolution for more than 90 subjects. Release 1.0 – a public –domain subset for 45 subjects (including KEMAR with large and with small pinnae) – is available by downloading from the website (http://interface.cipic.ucdavis.edu). In addition to including impulse responses for 1250 directions for each ear of each subject, the database includes a set of anthropometric measurements that can be used for scaling studies.

3. Properties of HRTF

HRTF impulse responses are the output of a linear and time-invariant system, that is, the diffraction and reflections around the human head, the outer ear, and the torso. Thus the impulse response responses can directly be represented as finite-impulse response (FIR) filters. There are often computational constraints that lead to the need of HRTF impulse response approximation. This can be carried out using conventional digital filter design techniques. It is, however, necessary to note that the filter design problem is not a straightforward one.

An attractive property of HRTFs is that they may be modeled as minimum-phase structures. The excess phase that is the result of subtracting the original phase response from its minimum-phase counterpart has been found to be approximately linear. This suggests that the excess phase can be separately implemented as an all pass filter or a simple delay line. In the case of binaural synthesis, the ITD part of the two HRTFs may be modeled as a separate delay line, and minimum-phase HRTFs may be used for synthesis (see Fig.1). According to

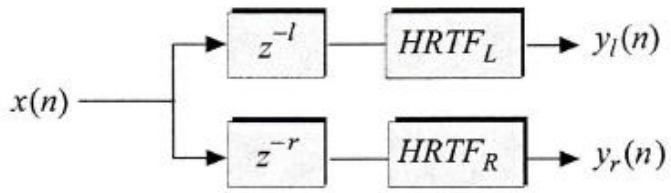

Figure 1. Implementation of HRTFs using minimum-phase HRTF approximation and pure delays for realization of ITD.

Kistler and Wightman, minimum-phase reconstruction does not have any perceptual consequences. Furthermore, with this method the designed FIR filters are of the shortest possible length for that magnitude response (property of minimum-phase filters).

4. Anthropometry

Although the exact HRTF is complicated, its general behavior can be estimated from fairly simple geometric models of the torso, head and pinnae. These models can be individualized to particular listeners if appropriate anthropometric measurements are available. However, specifying a general set of well defined and relevant measurements is problematic. The problem is particularly difficult for the pinna, where small variations can produce large changes in the HRTF. Anthropometric measurements, even if imperfect, enable the investigation of possible correspondences or correlations between physical dimensions are HRTF features.

The choice of anthropometry relevant to understanding or estimating HRTFs lead us to follow an approach proposed by Genuit and to define a set of 27 anthropometric measurements – 17 for the head and torso and 10 for the pinna.

The range of variation for the individuals in the CPIC database can be measured by some statistics for the anthropometric measurements. In general, histograms of the individual measurements indicate a basically normal distribution of values. The means and standard deviations for the anthropometric parameters are listed in Table 1. Here distances are measured in cm and angles in degrees, and the percentage variation is $2s/\mu$ in percent. For example, the mean head width was

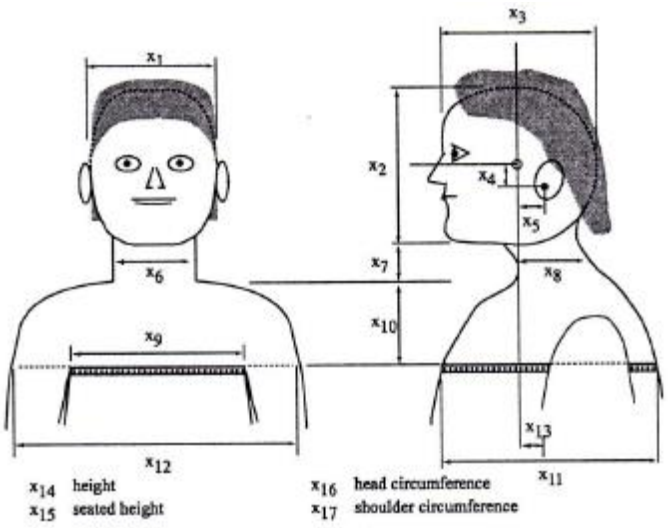

Figure 2: *Head and torso measurements*

14.49 cm and, assuming a normal distribution, 95% of the cases were within +/- 13% of the mean. Excluding the offset measurements x_4, x_5 and x_{13}, for which percentage deviation is not meaningful, we see that the average percentage deviation is +/- 26%. Thus, there is considerable variation in the sizes and shapes of the subjects in the database.

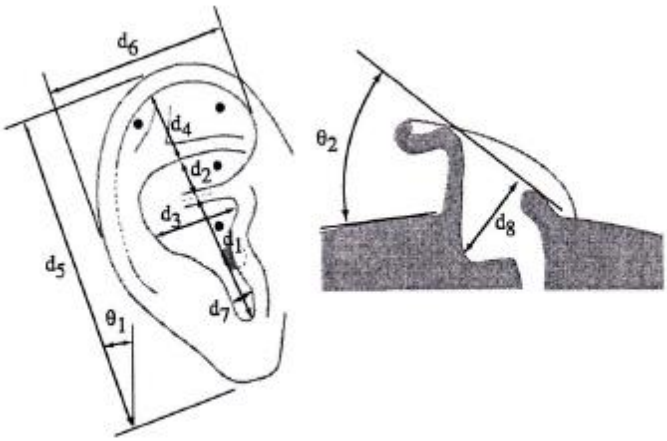

Figure 3: *Pinna measurements*

Var	Measurement	μ	σ	%
x_1	head width	14.49	0.95	13
x_2	head height	21.46	1.24	12
x_3	head depth	19.96	1.29	13
x_4	pinna offset down	3.03	0.66	43
x_5	pinna offset back	0.46	0.59	254
x_6	neck width	11.68	1.11	19
x_7	neck height	6.26	1.69	54
x_8	neck depth	10.52	1.22	23
x_9	torso top width	31.50	3.19	20
x_{10}	torso top height	13.42	1.85	28
x_{11}	torso top depth	23.84	2.95	25
x_{12}	shoulder width	45.90	3.78	16
x_{13}	head offset forward	3.03	2.29	151
x_{14}	height	172.43	11.61	13
x_{15}	seated height	88.83	5.53	12
x_{16}	head circumference	57.33	2.47	9
x_{17}	shoulder circumference	109.43	10.30	19
d_1	cavum concha height	1.91	0.18	19
d_2	cymba concha height	0.68	0.12	35
d_3	cavum concha width	1.58	0.28	35
d_4	fossa height	1.51	0.33	44
d_5	pinna height	6.41	0.51	16
d_6	pinna width	2.92	0.27	18
d_7	intertragal incisure width	0.53	0.14	51
d_8	cavum concha depth	1.02	0.16	32
θ_1	pinna rotation angle	24.01	6.59	55
θ_2	pinna flare angle	28.53	6.70	47

Table 1. Anthropometric statistics, $\% = 100(2\,\sigma/\mu)$

Correlations between these measurements may be of interest, since one might conjecture that a subject with a large head would also have large pinnae. Indeed, this is the basic assumption behind Middle brook's procedure for scaling HRTFs to account for changes in body size. In general, there are statistically significant but weak correlations between most pairs of measurements Scatter plots and correlation coefficients for four interesting examples are shown in the figure 4.

We focus on the important but difficult to measure pinna dimensions. Figure 4a shows that there is a fairly good correlation between pinna height and cavum concha height (0.45). There is also some correlation between head depth and cavum concha width (Fig. 4d, p=0.16), and about the same correlation between head height and cavum concha height p=0.17. In general, there appears to be relatively little correlation between the sizes of large and small anatomical

features, and accurate estimation of pinna dimensions from head and torso measurements is problematic.

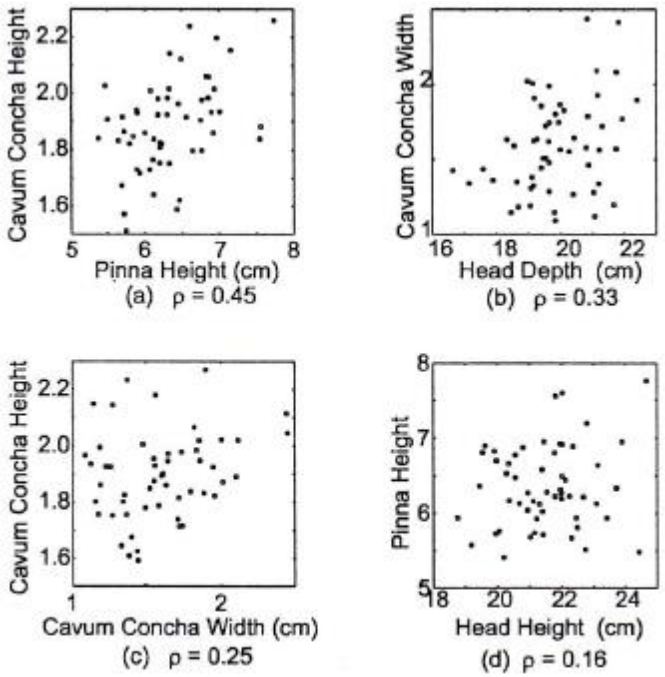

Figure 4: *Selected scatterplots*

5. HRTF Personalization

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 conha 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 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 measuring the measure over all k subjects

$$E^k = \sum_{i=1}^{7} \frac{(di - di^k)^2}{\sigma i^2}.$$

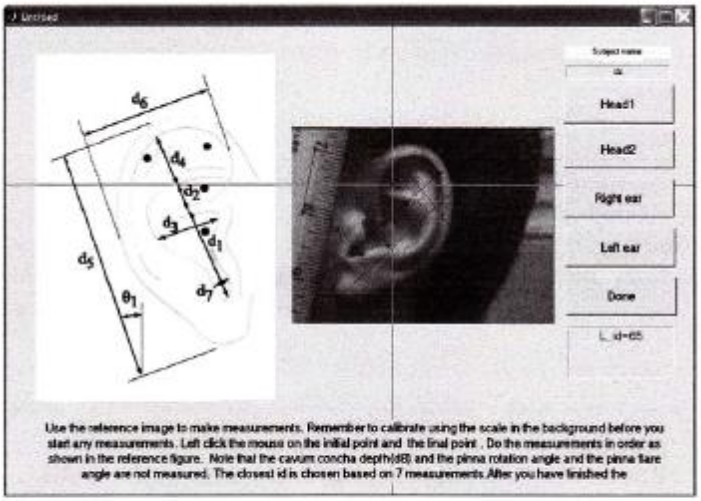

Fig. 1. Screenshot of the HRTF customization software.

Here σi^2 is the variance of the i^{th} parameter across all subjects in the database. Subject k, k=ard $\min_k E^k$, is chosen to be the best match.

6. HAT model

Direct HRTF measurements have always been problematic at low frequencies. A simple head-and-torso model was proposed in to compensate for these deficiencies. In the model, the human body is approximated by a "snowman" consisting of a spherical head and a speherical torso, seperated by a certain distance (neck). Analytical HRTF computation is possile 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 seperable on frequency axis (an object contributes to scattering only when the object size and the wave length are comparable)

The Hat model uses two different algorithms depending on whether the sourse 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 propogation paths, to the contralateral ear are also potentially shadowed by the head. We compute HAT model HRTF $H_h(\omega)$ following in algorithm. We use the phase of $H_h(\omega)$ for all ω and gradually blend

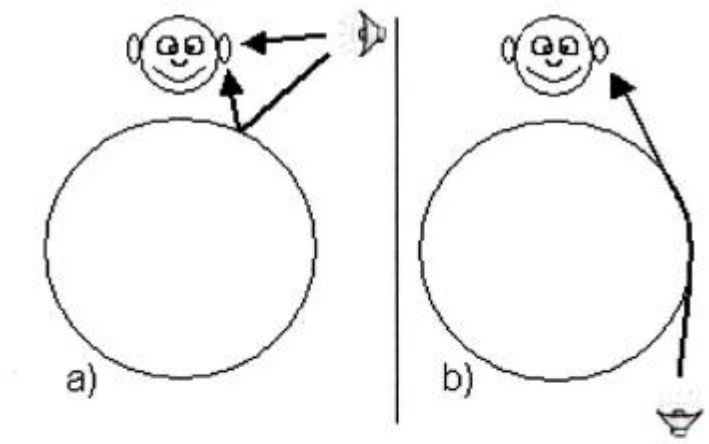

Fig. 2. a) Sound propagation path simulated by the HAT model in case of the source being outside of the torso shadow; b) source is now inside the torso shadow.

the log-magnitudes of $H_h(\omega)$ and the HRTF $H_c(\omega)$ selected from the CPIC database to get the combined HRTF $H_s(\omega)$:

$$A_s(\omega) = \begin{cases} A_h(\omega) & \omega < \omega_l \\ A_h(\omega) + \frac{A_c(\omega) - A_h(\omega)}{\omega_h - \omega_l}(\omega - \omega_l) & \omega_l < \omega < \omega_h \\ A_c(\omega) & \omega > \omega_h \end{cases}$$

$$A_s(\omega) = \log|H_s(\omega)|, \quad A_h(\omega) = \log|H_h(\omega)|,$$

$$A_c(\omega) = \log|H_c(\omega)|, \quad \omega_l = 500Hz, \quad \omega_h = 3000Hz.$$

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

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 are 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, 16.9cm 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. The system latency of the software are 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 $\varphi \in [-90^\circ, 90^\circ]$, elvation $\theta \in [-45^\circ, 65^\circ]$), 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 preceived position of the sound. The error in azimuth and in the elevation is found as a difference between true virtual source location and the pointed to lacation. 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. Two types of signals used are: 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 HRTF's.

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

In tables, s1 – s8 are the subject numbers, and colums are as follows: Mφ and Mθ are the average absolute values of the localization errors in azimuth and elevation, Eφ and Eθ are the root mean square errors in azimuth and elevation, and bφ and bθ 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 belived to be hampered most by using non-individualized HRTF), but 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 performace 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 localiztion in elevation. Now, for almost all of the subjects, incorporation of the HAT model improves performance more than personlization. For subjects 1 and 8, the localization worsens when going from GEN to PRS but improves again in PPS case (with the HAT model).

Personlization still seems to help for about half of the subjects, and for the other ones there is either no chance, or a degradation of performance. For some of the subjects, we tested GPS (generic-with-snowman) HRTF there, which also seems to improve performace 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 Mθ 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 10.36 for table 2. Indeed, it can expected that the HAT model will be important for the short bursts but will not matter much for continous sound. In the continuous sound localization, the source is essentially saying 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. 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 personlization 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 sort stimulus case, whichis probably due to the incorporation of the correctly personalized ITD cues computed by HAT model using measured head radius.

Using 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 rendering with non-personlized HRTFs.

7. HRTF Variation

One of the advantages of measuring HRTF data at high spatial resolution is that the data can represented as an image. Figure. 5a shows such an image representation of the impulse response KEMAR's right ear. Each column in this image is one impulse response at a particular azimuth, with brightness coding the strength of the response. The variation of arrival time with azimuth is clearly seen in the roughly sinusodial shape of the top envelope of the response. The weakening of the response as the azimuth approaches the opposite side of the head shows the effect of head shadow.

Figure 5b shows the spectrum for this same case. Here each column is the magnitude of the HRTF in db, after the power spetrum was smoothed by a constant-Q filter (Q=8). The generally darker appreance of the right half of the image shows the effect of head shadow. The strong response on the ipsilateral side around 5kHz corresponds to the quarter-wavelength depth resonance identified by shaw, and the weak response around 9kHz is the so-called "pinna notch." In addition, other interesting but more difficult to explain azimuth-dependent spectral features can be seen.

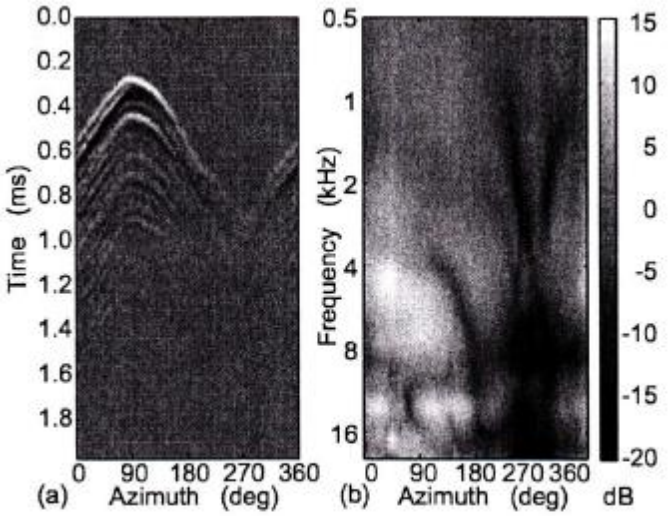

Figure 5: *Horizontal plane (a) HRIR (b) HRTF*

These images give some idea of the HRTF variability for a single subject. It is more difficult to characterizes the range of HRTF variability between subjects. However, two simple measures – the maximum interaural time difference ITD_{max}

and the pinna-notch freqeuncy f_{pn} – are simple, perceptually relevant parameters that characterize the variability that exists.

For the subjects in the database, ITD_{max} is approximately normally distributed, with μ = 646μsec and s = 33μsec, which corresponds to a +/-10.3% variation. Not surprising, ITD_{max} is strongly correlated with head size (see Fig.6), and it can be estimated quite accurately using simple linear regression. The best single predictor is the head width, with a correaltion coeffcient of p=0.78 between the estimated and the actual ITD. The best pair are the head width and the head depth, for which p=0.87. For a more detailed presentation of the estimation of the ITD from anthrometry.

In frequency-domain, most HRTFs exhibit the prominent depth resonance around 3 to 4kHz, followed by the pinna "notch". Figure 7 shows the HRTF magnitudes for $\theta = \phi = 0^\circ$ for a set of 54 subjects. The pinna notches are indicated by the black dots, and the graphs are sorted by the pinna-notch freqency f_{pn}.

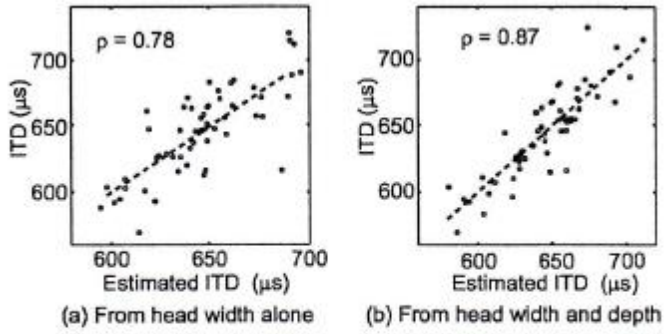

Figure 6: *Scatterplots for estimation of the ITD*

Statistically, f_{pn} is approximately normally distributed, with μ = 7600 hz and s = 1050 Hz, which corresponds to a rather large +/- 28% variation. As expected, f_{pn} is correlated with the pinna measurements, but the relationship is not strong, and linear regression is not as successful in estimating f_{pn} from anthropometry. The best single predictor of f_{pn} is the cavum concha height (p=0.33). Somewhat surprising, the best pairof predictors are two angles θ_1 and θ_2 (p=0.42), and the best triple addsto these the fossa height (p=0.51). These results reflect the fact that the scattering of incident waves by the pinna is a complex preocess related to detailed features, and that accurate estimation of f_{pn} may well require additional

concha parameters not included in our measurements. However, simple regression analysis does help identify the most significant of the measured paramenters or indicates the need for additional measurements. In our view that the effective customization of HRTFs will requires a deeper understanding of the perceptually important characteristics of the HRTF and of their dependence on detiled pinna features.

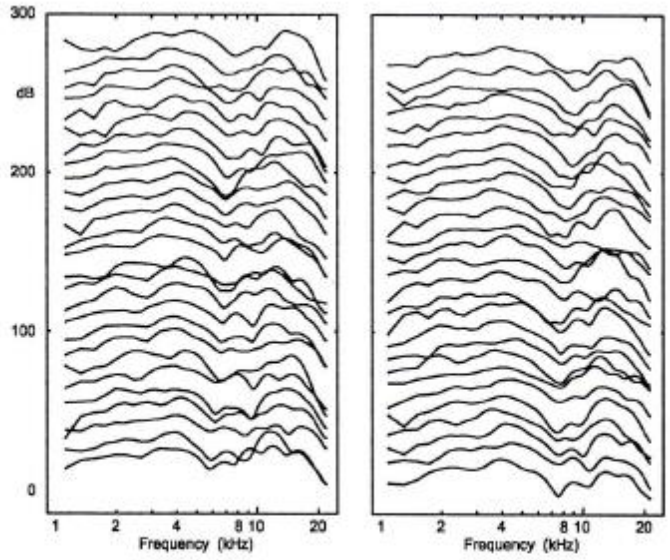

Figure 7: *HRTF magnitudes for* $\theta = \phi = 0°$

Note :- X-axis – frequency, Y-axis – azimuth

8. Composition of the HRTF

The HRTFs used for analysis were taken from the CPIC database. The CIPIC HRTF database is a public domain database of high spatial resolution HRTF

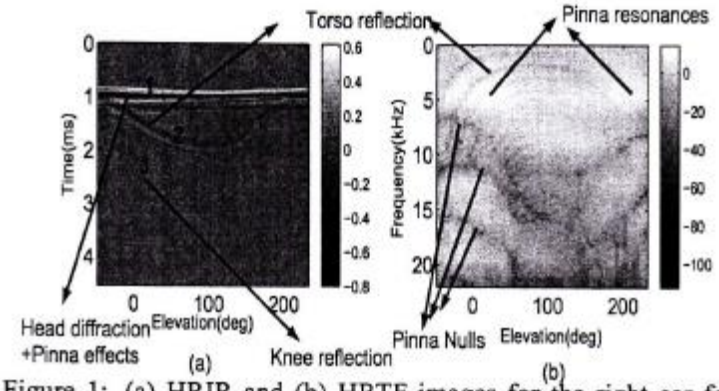

Figure 1: (a) HRIR and (b) HRTF images for the right ear for azimuth angle $\theta = 0^o$ for all elevations varying from -45^o to $+230.625^o$.

measurememnts for 45 different subjects along with the anthrometry. The coordinate system used in the database is the head-centered interaural polar coordinate system. The azimuth is sampled from -80° to 80° and the elevation from -45° to +230.625° . For any given azimuth, we form a 2-D array, where each column is the HRIR or the HRTF for a given elevation, and the entire array is displayed as an image. This method of visualization helps to identify different features and their variation with elevation. Figure shows the HRIR and HRTF images (for all elevations) corresponding to azimuth 0° for the right ear for a particular subject. In Figure 1(a) the gray scale value represents the amplitude of HRIR and in Figure 1 (b) the gray scale vlaue is the magnitude of the HRTF in dB. The different features corresponding to different structural components are also marked.

Composition of the responsed in terms of head diffraction effects, head and torso reflection, pinna effects and knee reflection can be seen both in the time domain and in the frqency domain. Most of the features marked in Figure 1, were confirmed experimentally with the KEMAR mannequin, where the responses were measured by removing and adding different components like the pinna, head and torso. Consider the HRIR image plot as shown in Figure 1(a). Three distinct ridges which are marked as1, 2 and 3 can be seen the HRIR image plot. It may be difficult to see three regions in an individual HRIR. However, the human visual system is able to perceive these three regions distinctly in the image. The first distinct feature is due to the direct acoustic wave that reaches the pinna. The difference between the time arrival for the left and the right ear is the ITD. It can be seen that the ITD

has a slight dependence on the elevation. Immediately after the direct wave the activity seen in the close vicinity is due to diffraction of the sound around the head. The corresponding diffraction pattern in the frequency domain can be explained by the Rayleigh's analytical solution for a spherical head. Also note that the diffraction is not clearly visible in the HRTF image, as the nullsand resonances caused by the pinna dominate the nulls in the diffraction pattern. The effect of head diffraction is more prominent in the contralateral HRTF than in the ipsilateral HRTF.

The second valley shaped ridge which is seen between 1 ms and 2 ms is due to the reflected wave from the torso, reaching the pinna. The delay between the direct wave and the reflected wave from the torso is maximumal above the head, and decreases on both sides. This can be explained using simple ellipsodial models for the head and torso. In the frequency domain the effect of this delay is the arch shaped comb-filter notches that run throughout the spectrum (see Figure 1(b)). The activity seen after 2 ms is due to knee reflections, since the measurements were done with the subjects seated. This is confirmed by the observation that this activity is not seen in the back. The other features that are prominent in the frequency domain, but difficult to see in the time domain are the notches above 6 kHz which are caused by the pinna. Various models have been proposed to explain the cause of these notches. They are primarily due to the scattering of acoustic wave by the pinna.

In most of the studies in the literature the effects of the indiviual parts were studied in isolation, and the responses were verified with analytical studies on simplified models. Most of the studies on the composition of the HRTFs normally do not address the problem of decomposing the measured HRTF into components. This is partly due to lack of methods to process this complex signal using available signal processing tools.

9. Decomposition

The basis for the decomposition techniques presented are spectral peaks and nulls, i.e., poles and the zeros. These poles and zeros are caused by different parts like the head, torso, knees and pinna. The challenging task is to isolate the prominent spectral nulls caused by different acoustic phenomena.

The poles can be extracted by doing a Linear Prediction (LP) Analysis. In LP analysis each sample is predicted as a linear combination of the past p samples, where p is the order for prediction. If h(n) represents the actual HRIR, then the predicted HRIR is given by

$$\hat{h}(n) = -\sum_{k=1}^{p} ak\, h(n-k) \quad \ldots(1)$$

Where, a_k are the LP coefficients obtained by LP analysis. This basically fits an all-pole model of order p to the HRIR. The HRIR cannot be completely modelled by just an all-pole model. Hence the prediction will never be perfect. The error between the actual sequence and the predicted sequence is given by

$$r(n) = h(n) - \hat{h}(n) + \sum_{k=1}^{p} ak\, h(n-k) \quad \ldots(2)$$

Where r(n) is called the LP residual. From the given HRIR signal we can compute the LP residual by passing it through the inverse filter given by

$$A(z) = 1 + \sum_{k=1}^{p} ak\, z\text{^} - k \quad \ldots(3)$$

The roots of A(z) are the locations of the poles. In LP residual the spectral nulls are preserved. LP analysis does not affect the location of the nulls significantly. Before applying the LP analysis, the HRIR signal is pre-emphasized by using a difference operation to remove the DC bias, if any, in the HRIR. If h(n) represents the actual HRIR, then the pre-emphasized HRIR $h^d(n)$ is given by

$$h^{d(n)} = h(n) - h(n-1) \quad \ldots(4)$$

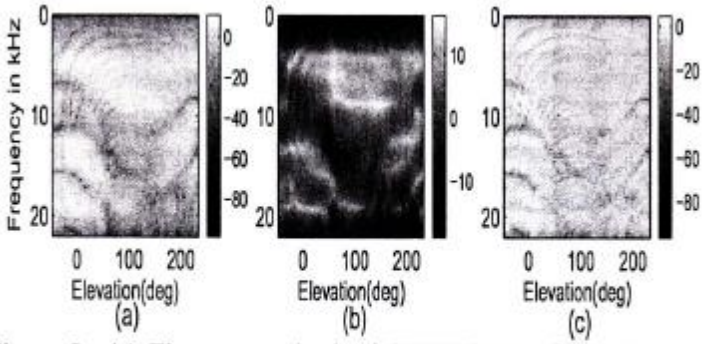

Figure 2: (a) The pre-emphasized HRTF image for right ear at azimuth 0^o, (b) frequency response of a 12^{th} order all-pole model and (c) the frequency response of the corresponding LP residual.

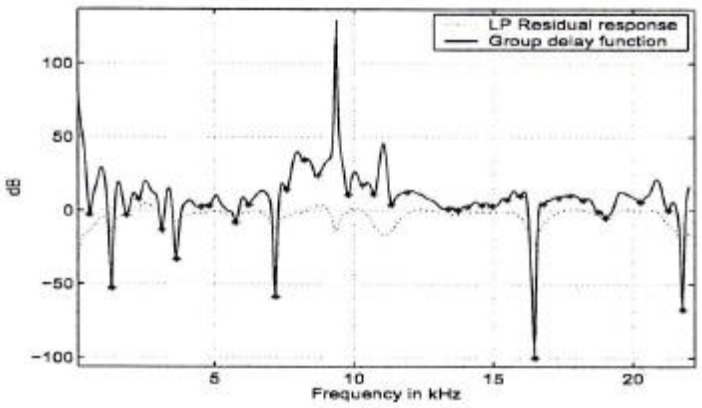

Figure 3: Frequency response of the HRIR residual for elevation 0^o and azimuth 0^o and the corresponding group delay spectrum.The nulls are also marked.

Figure 2(a) shows the pre-emphasized HRTF image for azimuth 0° . The corresponding frequency response of a 12^{th} order all-pole model and the frequency response of the LP residual are in shown in Figure 2(b) and (c) respectively. It can be seen that the spectral nulls are preserved in the frequency response of the LP residual.

Once we have the LP residual, in order to emphasize the spectral peaks and nulls, we compute it's group delay function. The group delay function is the negative of the derivative of the phase of the frquency response of the signal. If $H(\omega)$ is complex frequency response of a HRIR h(n), then the group delay is given by

$$h_g(\omega)=\frac{d\theta(\omega)}{d\omega} \quad \ldots (5)$$

where ω is the angular frequency and $\theta(\omega)$ is the phase angle of $H(\omega)$. The peaks and the valleys are sharper in the group delay spectrum and they typically correspond to significant poles and zeros. Figure3 shows the frequency response of the LP residual of the HRIR for elevation 0° and azimuth 0° , and the corresponding group delay function. It can be seen that the nulls show up as very sharp valleys in the group delay function. Most of the spectral nulls are due to the combined effects of the head and torso reflection, the knee reflection, the head diffraction effects and the pinna effects. The task is to separate the nulls due to the individual effects. In order to highlight nulls due to different components of the

HRIR we multiply the LP residual of the HRIR by r^n where n is the sample index and r is a constant to be chosen between 0.9 and 1.0. Before doing this, it is

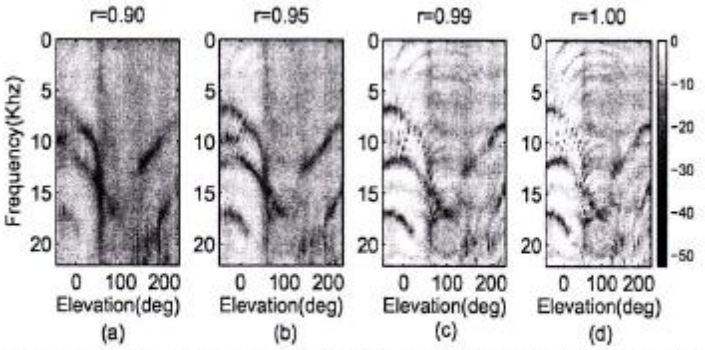

Figure 4: Frequency response of the time aligned and scaled LP residual for (a) $r = 0.90$, (b) $r = 0.95$, (c) $r = 0.99$ and (d) $r = 1.00$

necessary to time align all the HRIRs. The initial onset time can be found by computing the average delay, i.e., the slope of the phase response. Once the onset time is found, the HRIR residual can be multiplied by r^n with the index starting from the instant corresponding to the first onset. By varying r between 0.9 and 1.0, one can decompose the HRIR into different components. When r is less say around 0.9, since the function r^n decays very rapidly only the initial part of the HRIR which is mainly due to the head diffraction and pinna effects is emphasized and the rest of the HRIR is suppressed. As r is further increased the function r^n decays more slowly and hence the later part of the HRIR also gets significant emphasis. When r=1.0 the complete HRIR gets equal emphasis. Figure 4 shows the frequency response of the time aligned and scaled LP residual for different values of r. For r=0.90 only the nulls due to the pinna are prominent and the other features which we mentioned earlier are significantly reduced. At about r=0.99 the ridges due to torso reflection become prominent. Also note that the effects due to the pinna nulls are still there but since we already know the locations of the pinna nulls it is possible to isolate the nulls due to torso reflection only. Increasing r further will bring out the nulls due to knee reflection. The knee reflection is fainter than the torso reflection and the delay is large. As a result the fine ridges in the HRTF image due to torso reflection are not clearly visible in the HRTF image due to torso reflection are not clearly visible in the HRTF image when printed on paper. The zero thresholded group delay function of the scaled HRIR (multipled by r^n) LP residual shows the nulls corresponding to different components.

10. Feature Extraction

In this section we show how features like pinna resonant frequencies, pinna nulls and the delay due to torso and knee reflection can be extracted using the above decomposition technique. Previous studies have shown that these features are perceptually important for localization. Knee reflections appear because the measurements were made with the subject seated. It is not known whether it has any significance for localization. The poles extracted by LP analysis appear to correspond to the resonances of the pinna. Figure 5 shows the frequency response of the 12^{th} order all-pole model for the subject 10 for azimuth 0° as a function of different elevations as a mesh plot. These six modes are marked in the plot. As discussed earlier, depending upon the value of r, different effects get emphasized in the group delay function. For r=0.90 the effects due to the pinna are emphasized, and the effect due to the head and torso reflection is reduced. Therefore the prominent nulls in the group delay function are mostly due to pinna. We can find the frequencies of these nulls by finding the local minima. Figure 6(a) shows the

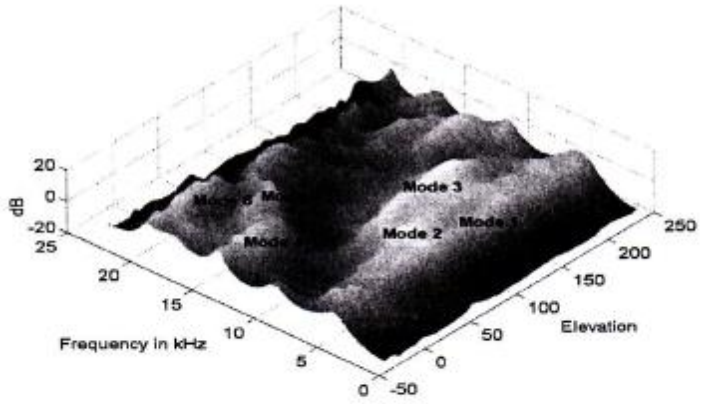

Figure 5: Frequency response of the 12^{th} order all pole model for azimuth 0 as a function of different elevations. The six modes are approximately marked.

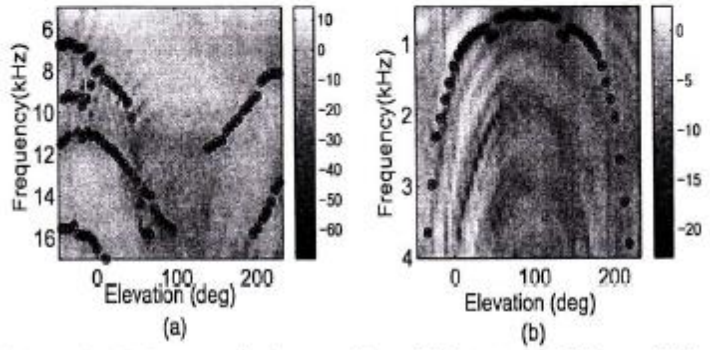

Figure 6: (a) Extracted pinna nulls and (b) extracted first null due to torso reflection.

extracted pinna nulls. The effect of torso reflection delay in the frequency domain is the appearance of periodic comb-filter nulls. The objective is to extract the frequencies at which these notches appear and derive analytical expressions for the frequency spacing and hence the time delay. For r=0.99, the effects due to the torso reflection are emphasized. The nulls can be extracted by finding the local minima of the zero thresholded group delay function. Figure 6(b) shows the extracted first null due to torso reflection. Extracting the delays due to knee reflection using the above approch is a bit tricker since the ridges due to knee reflection are very faint and the frequency spacing is very less.

11. Result

High-spatical-resolution HRTF measurements clarify the physical sources of HRTF behavior. A uniform database of HRTF's enables the study of person-to person differences and of the relation of temporal and spectral characteristics of the HRTF to the anthropometric data.

The composition and decomposition of the HRTF into different components, and extraction of features which could be perceptually important for sound source localization. Previous studies have confirmed that the features we extract are perpectaully significant for localization. So instead of using the complete HRTF, one could build simplified models based on the features extracted. Further, current HRTF interpolation methods do not take the perceptual importance of different features into consideration and lose these features by incorrect interpolation. Using the features extracted interpolation can be done in the feature domain. Also, these features can be related to the physical dimensions of the human anatomy and the pinna so that the HRTF could be customized.

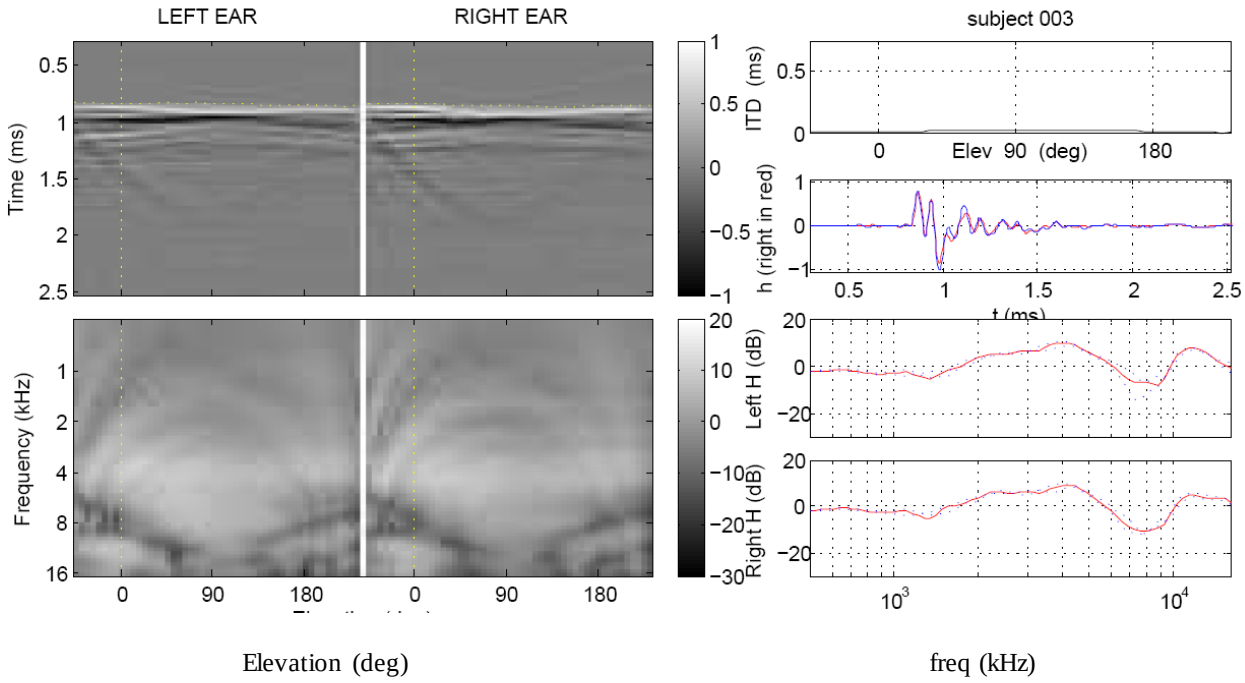

Figure 1. Subject 003, Azimuth 0° , Elevation 0°

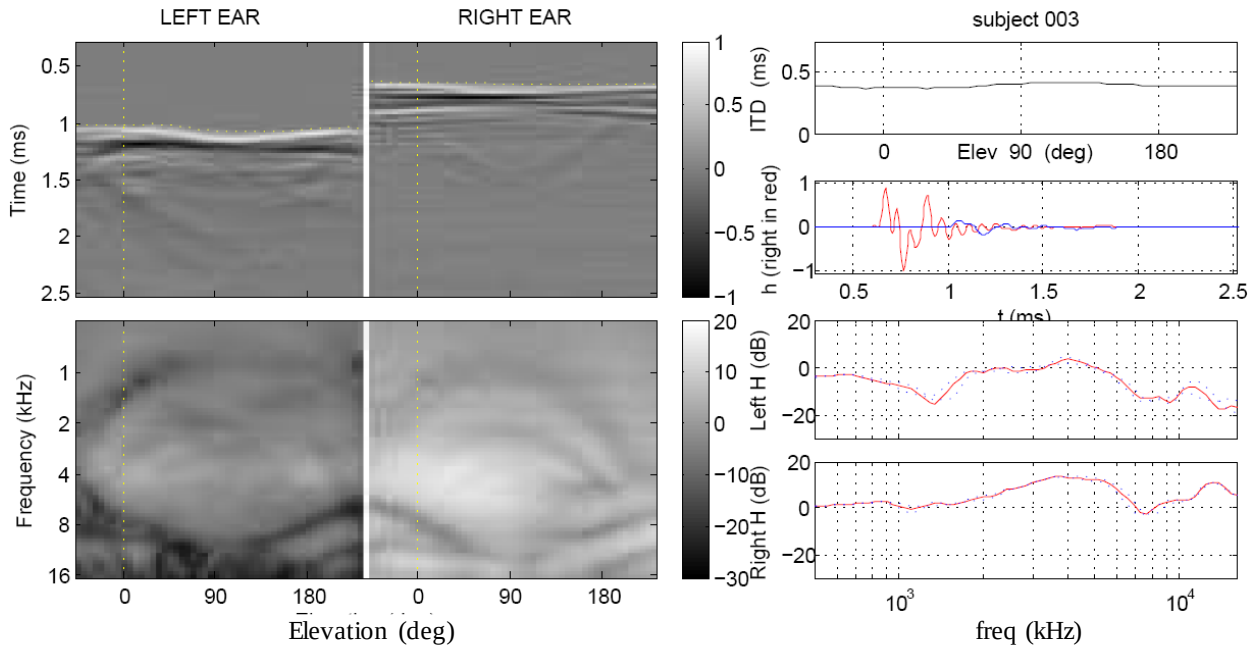

Figure 2. Subject 003, Azimuth 45° , Elevation 0°

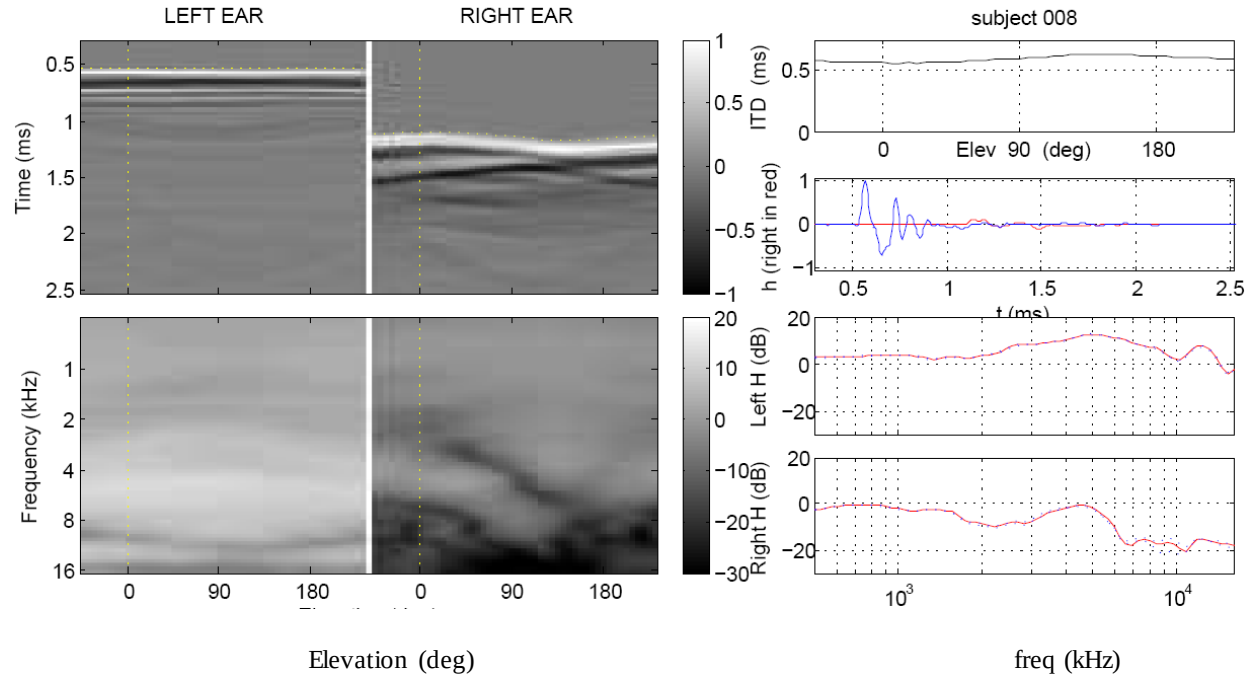

Figure 3. Subject 008, Azimuth -80° , Elevation 0°

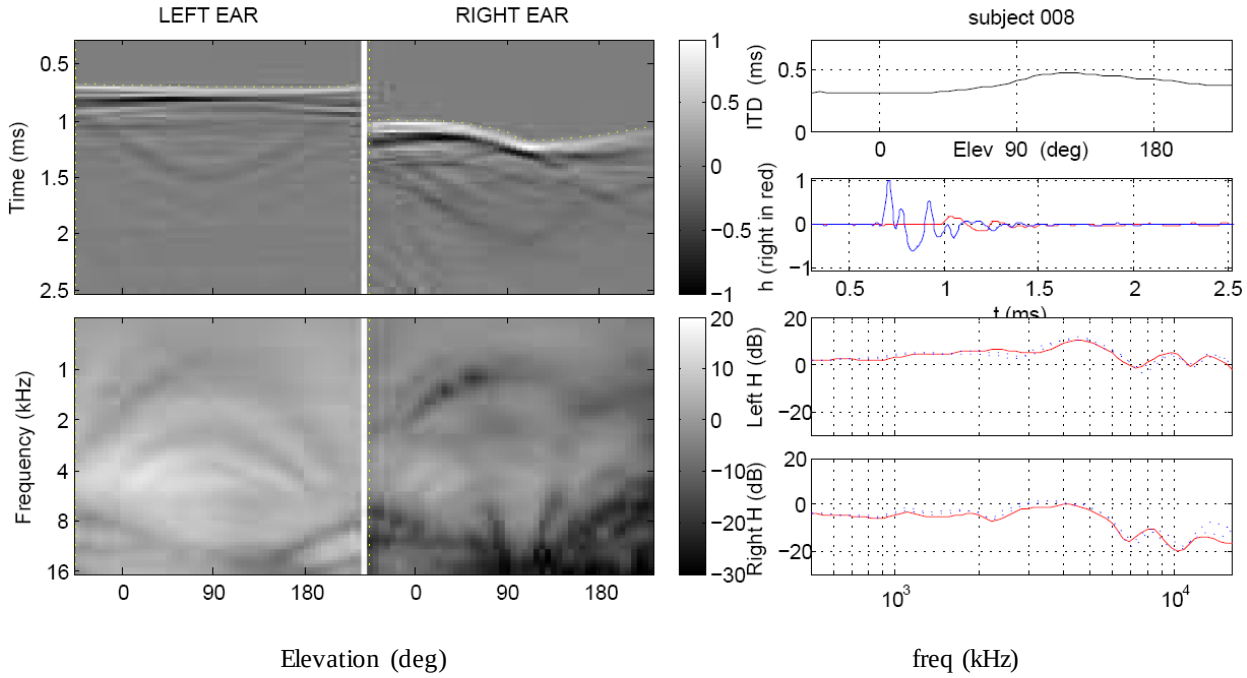

Figure 4. Subject 008, Azimuth -45° , Elevation -45°

The examples are shown in the above figures. Figure 1 shows subject 003 at azimuth 0° , elevation 0° , Figure 2 shows Subject 003 at azimuth 45° , elevation 0° , Figure 3 shows Subject 008 at azimuth -80° , elevation 0° and figure 4 shows Subject 008 at azimuth -45° , elevation -45° .

12. References

[1] V.R. Algazi, R. O. Duda and D. M. Thompson, C. Avendano, "The CIPIC HRTF Database" October 21-24 2001, New Paltz, New York.

[2] Vikas C. Raykar, Ramani Duraiswami, Larry Davis, B. Yegnanarayana, "Extracting Significant Features from the HRTF" July 6-9 2003, Boston, MA

[3] Dmitry N. Zotkin, Jane Hwanf, Ramani Duraiswamy, Larry S. Davis, "HRTF Personalization using Anthropometric Measurements" College park, MD