EXTRACTING SIGNIFICANT FEATURES FROM THE HRTF

Vikas C. Raykar, Ramani Duraiswami, Larry Davis

Perceptual Interfaces and Reality Laboratory
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
{vikas,ramani,lsd}@umiacs.umd.edu

B. Yegnanarayana

Speech & Vision Laboratory
Department of Computer Science and Egg.
IIT Madras, Chennai-600 036, India
yegna@cs.iitm.ernet.in

ABSTRACT

The 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 paper 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. We develop signal processing algorithms to decompose the HRTF into components, and extract the features corresponding to each component.

The support of NSF award ITR-0086075 is gratefully acknowledged

1. INTRODUCTION AND PREVIOUS WORK

Humans have an amazing ability to localize a sound source, i.e., determine the range, elevation and azimuth angles of the direction of the sound source. Interaural Time Difference (ITD) and the Interaural Level Difference (ILD) are known to provide primary cues for localization in the horizontal plane [1]. However these differences do not account for the ability to locate sound for positions in the *cone of confusion*, which have the same ITD and ILD cues. This can be explained in terms of the spectral filtering provided by the torso, head and the pinnae. This filtering can be described by a complex frequency response function called the Head Related Transfer Function (HRTF). The corresponding impulse response is called the Head Related Impulse Response (HRIR). The spectral features in the HRTF due to pinna diffraction and scattering seem to provide cues for vertical localization, i.e., elevation of the source [2]. The spectral features due to the pinna are dominant only in the high frequency (> 5 kHz) range. Psychoacoustic and perceptual studies show that there are features in the low frequency range (< 3 kHz) which seem to provide some cues for vertical localization [3]. These cues are attributed to the contributions of the head diffraction and torso reflections [3].

The HRTF varies significantly between different individuals due to differences in the sizes and shapes of different anatomical parts like the pinnae, head and torso. Applications in the creation of virtual auditory displays require individual HRTFs for perceptual fidelity. A generic HRTF would not work satisfactorily since it has been shown that non-individual HRTF results in poor elevation perception [4]. The usual customization method is the direct measurement of HRTFs, which is a time consuming and laborious process. Other approaches that have met with varying success include numerical modelling [5], frequency scaling the non-individual HRTF to best fit the listener's one [6] and database matching [7].

This paper is based on the observation that different anatomical parts contribute to different temporal and spectral features in the HRIR and HRTF respectively. If these features can be extracted, then it is possible to study the relationship between them and anthropometry. Based on these, new approaches for HRTF customization using these features can be developed.

The temporal and spectral features useful for localization have been previously studied in different ways. Contributions of different parts such as the pinna, head, torso and shoulder to these features have been studied using the KEMAR mannequin [3, 8]. Analytical solutions were obtained using simple geometrical models for the head and torso [9]. Psychoacoustic studies were made using the derived responses to show that the low frequency features due to head and torso contribute to vertical localization in some cases [3]. Studies have also been made to approximate the HRTF by using pole-zero models [10].

In all the above studies the objective was to explain the features observed in the HRIR or the HRTF for a real subject in terms of contributions from individual anatomical parts. These studies demonstrate the significance of the resulting HRTF features to the perception of localization of the source. While all these studies address the issue of how the HRTF is composed, there is no attempt to *decompose* the HRTF of a real subject into different components. Most of the these studies do not provide quantitative values for the features in the experimental data. Further, the presence of interacting multiple components makes it difficult to extract some of the features. Therefore, a decomposition of the HRIR/HRTF signal into meaningful components is needed. Analysis of the components may provide useful measurable features. If it is possible to relate these features to the anthropometry, then it may be possible to synthesize a response incorporating the perceptually significant features for any person and for any source location.

In this paper, we discuss methods to decompose the HRTF in order to derive the features corresponding to each component. In Section 2, we study the composition of the HRTF. In Section 3, we develop signal processing methods to decompose the HRTF into different components. In Section 4, we discuss how different features can be extracted using the proposed decomposition technique. In Section 5, we give a summary of the results, and discuss possible studies these results may lead to.

2. COMPOSITION OF THE HRTF

The HRTFs used for analysis in our paper were taken from the CIPIC database [11]. The CIPIC HRTF database is a public do-

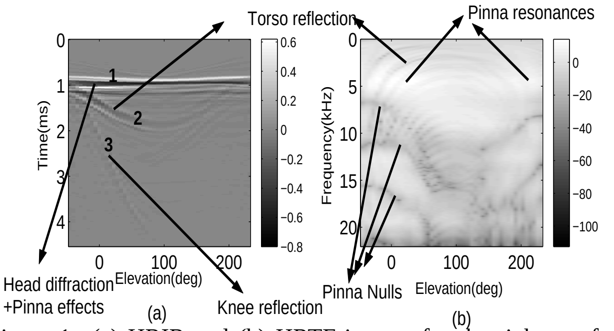

Figure 1: (a) HRIR and (b) HRTF images for the right ear for azimuth angle $\theta = 0^o$ for all elevations varying from -45^o to $+230.625^o$.

main database of high spatial resolution HRTF measurements for 45 different subjects along with the anthropometry. The coordinate system used in the database is the head-centered interaural polar coordinate system [11]. The azimuth is sampled from -80^o to 80^o and the elevation from -45^o to $+230.625^o$. 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 1 shows the HRIR and HRTF images (for all elevations) corresponding to azimuth 0^0 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 value is the magnitude of the HRTF in dB. The different features corresponding to different structural components are also marked.

Composition of the responses 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 frequency 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 [8]. Consider the HRIR image plot as shown in Figure 1 (a). Three distinct ridges which are marked as 1, 2 and 3 can be seen in the HRIR image plot. It may be difficult to see these 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 of 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 [1]. Also note that the diffraction pattern is not clearly visible in the HRTF image, as the nulls and 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 maximal above the head, and decreases on both sides. This can be explained using simple ellipsoidal models for the head and torso [9]. 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 [12, 13]. They are primarily due to the scattering of acoustic wave by the pinna. In this paper we are mainly concerned with the extraction of these notches. Also present in the response are the resonances due to the pinna (the bright patches in the HRTF image in Figure 1 (b)), which were experimentally measured by Shaw [13].

In most of the studies in the literature the effects of the individual 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. In the next section we show that it is indeed possible to develop suitable signal processing techniques to decompose the HRIR/HRTF. While the decomposition techniques are guided by our prior knowledge of the physics of the problem, the interpretation of the features is in terms of the anthropometry.

3. DECOMPOSITION OF THE HRTF

The basis for the decomposition techniques presented in this section is that important features are present as 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

$$\hat{h}(n) = -\sum_{k=1}^{p} a_k h(n-k) \tag{1}$$

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

$$r(n) = h(n) - \hat{h}(n) = h(n) + \sum_{k=1}^{p} a_k h(n-k) \tag{2}$$

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

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3}$$

The roots of $A(z)$ are the locations of the poles. We show in Section 4 that these poles correspond to resonances of the pinna, reported by Shaw [13]. We computed the pole locations for different orders of LP analysis and found that the pole locations do not change significantly, confirming the fact that their origin is indeed due to resonances. In the 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

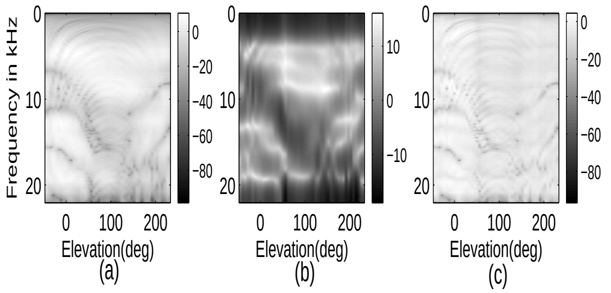

Figure 2: (a) The pre-emphasized HRTF image for right ear at azimuth 0^o, (b) frequency response of a 12^{th} order all-pole model and (c) the frequency response of the corresponding LP residual.

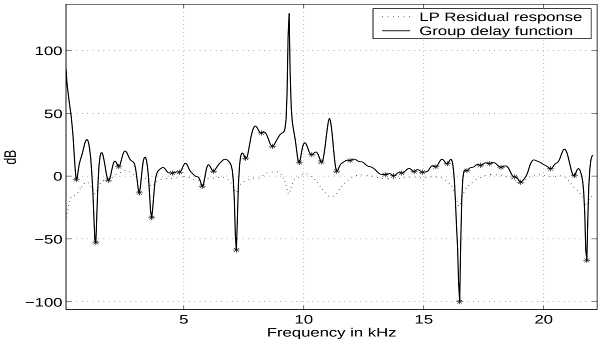

Figure 3: Frequency response of the HRIR residual for elevation 0^o and azimuth 0^o and the corresponding group delay spectrum.The nulls are also marked.

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

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

Figure 2 (a) shows the pre-emphasized HRTF image for azimuth 0^o. The corresponding frequency response of a 12^{th} order all-pole model and the frequency response of the LP residual are 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 [14]. The group delay function is the negative of the derivative of the phase of the frequency response of a signal. If $H(\omega)$ is complex frequency response of a HRIR $h(n)$, then the group delay is given by [14]

$$h_g(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{5}$$

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. Figure 3 shows the frequency response of the LP residual of the HRIR for elevation 0^o and azimuth 0^o, 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 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 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 when printed on paper. The zero thresholded group delay function of the scaled HRIR (multiplied by r^n) LP residual shows the nulls corresponding to different components.

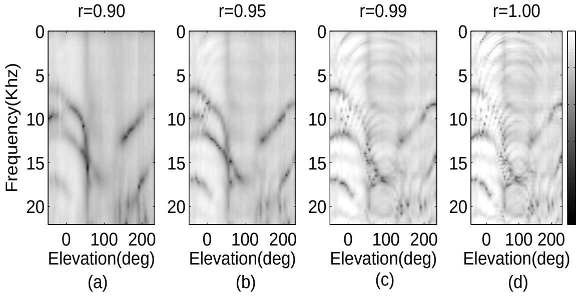

Figure 4: Frequency response of the time aligned and scaled LP residual for (a) $r = 0.90$, (b) $r = 0.95$, (c) $r = 0.99$ and (d) $r = 1.00$

4. 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 reported by Shaw [13]. Shaw described six modes of resonances under the blocked meatus condition based on experimental measurements for 10 subjects [13]. Figure 5 shows the frequency response of the 12^{th} order all-pole model for the subject 10 for azimuth 0^o 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

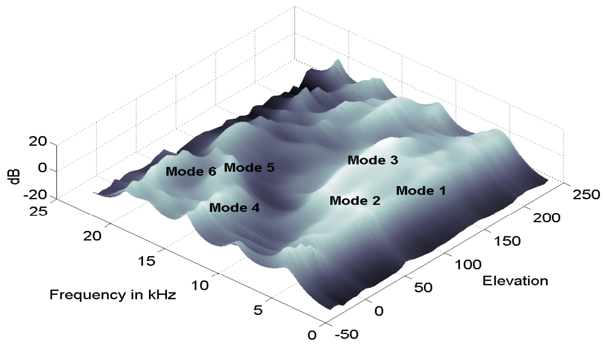

Figure 5: Frequency response of the 12^{th} order all pole model for azimuth 0 as a function of different elevations. The six modes are approximately marked.

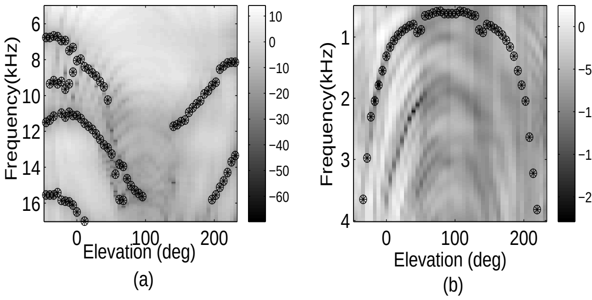

Figure 6: (a) Extracted pinna nulls and (b) extracted first null due to torso reflection.

are mostly due to the pinna. We can find the frequencies of these nulls by finding the local minima. Figure 6 (a) shows the 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 approach is a bit trickier since the ridges due to knee reflection are very faint and the frequency spacing is very less. Currently we are also working on time-domain approaches to extract these delays. We were able to extract these features for all the subjects in the CIPIC database but due to space constraints the results are shown for only one subject. For different subjects we need to tune the value of r slightly to get the desired features.

5. CONCLUSION AND FUTURE WORK

The main contribution of this paper is the 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 perceptually 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.

6. REFERENCES

[1] J. W. Strutt (Lord Rayleigh), “On our perception of sound direction,” *Phil. Mag.*, vol. 13, pp. 214–232, 1907.

[2] F. L. Wightman and D. J. Kistler, “Monaural sound localization revisited,” *Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1050–1063, Feb. 1997.

[3] V. R. Algazi, C. Avendano, and R. O. Duda, “Elevation localization and head-related transfer function analysis at low frequencies,” *Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1110–1122, Mar 2001.

[4] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, “Localization using nonindividualized head-related transfer functions,” *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[5] Y. Kahana, P. A. Nelson, M. Petyt, and S. Choi, “Numerical modelling of the transfer functions of a dummy-head and of the external ear,” in *AES 16th Int. Conf. Spatial Sound Reproduction*, Rovaniemi, Finland, 1999.

[6] J. C. Middlebrooks, “Virtual localization improved by scaling non-individualized external-ear functions in frequency,” *Journal of the Acoustical Society of America*, vol. 106(3), pp. 1493–1509, 1999.

[7] D. N. Zotkin, R. Duraiswami, L. Davis, A. Mohan, and V. Raykar, “Virtual audio system customization using visual matching of ear parameters,” in *International Conference on Pattern Recognition*, August 2002.

[8] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, “Structural composition and decomposition of HRTF’s,” in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, NY, Oct. 2001.

[9] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, “Approximating the head-related transfer function using simple geometric models of the head and torso,” *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2053–2064, Nov 2002.

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

[11] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRTF database,” in *Proc.2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, NY, Oct.21-24 2001, pp. 99–102.

[12] E. A. Lopez-Poveda and R. Meddis, “A physical model of sound diffraction and reflections in the human concha,” *Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3248–3259, Nov. 1996.

[13] E. A. G. Shaw, “Acoustical features of the human ear,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. B. Anderson, Eds., pp. 25–47. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

[14] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan, “Significance of group delay functions in signal reconstruction from spectral magnitude or phase,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 3, pp. 610–623, 1984.