Audio Engineering Society

Convention Paper 9447

Presented at the 139th Convention
2015 October 29–November 1 New York, USA

Capturing the elevation dependence of interaural time difference with an extension of the spherical-head model

Rahulram Sridhar and Edgar Y. Choueiri

3D Audio and Applied Acoustics Laboratory, Princeton University, Princeton, NJ, 08544, USA

Correspondence should be addressed to Rahulram Sridhar (rahulram@princeton.edu)

ABSTRACT

An extension of the spherical-head model (SHM) is developed to incorporate the elevation dependence observed in measured interaural time differences (ITDs). The model aims to address the inability of the SHM to capture this elevation dependence, thereby improving ITD estimation accuracy while retaining the simplicity of the SHM. To do so, the proposed model uses an elevation-dependent head radius that is individualized from anthropometry. Calculations of ITD for 12 listeners show that the proposed model is able to capture this elevation dependence and, for high frequencies and at large azimuths, yields a reduction in mean ITD error of up to 13 microseconds (3% of the measured ITD value), compared to the SHM. For low-frequency ITDs, this reduction is up to 160 microseconds (23%).

1. INTRODUCTION

Sound originating off the median plane of the listener reaches the ipsilateral ear a short time before the contralateral one. This time difference, called the interaural time difference (ITD), is one of three cues used to localize sound. It depends on the listener's morphology, and the location and frequency content of the sound source [1].

The ITD of a listener needs to be accurately reproduced when creating 3D sound using binaural technology. For example, in virtual reality games implementing such technology, it may be important for the gamer to accurately localize sounds emanating from objects off to one side and hidden from view. Since it has been well established that the ITD is the dominant cue for localization of wideband sound sources situated laterally [2], inaccurately reproduced ITDs will lead to incorrect localization of these sources. Furthermore, as ITDs are well known to be highly idiosyncratic [3] and have been shown to vary significantly with elevation [4], it is essential that the reproduced ITDs are individualized and exhibit the correct elevation dependence.

Therefore, an approach to accurately estimate individualized, and elevation-dependent ITDs is required. The approach should also perform the estimation quickly and conveniently, without the need for specialized equipment, if it is to be used in consumer applications. This might mean, for example, that the approach must only require a small number of easily obtainable anthropometric measurements using simple, non-intrusive techniques. It would be useful, therefore, to gain some insight into how different anthropometric features influence ITD depending on the location of the sound source. This would allow developing the approach to only require those anthropometric features that are both relevant for estimating ITD, and that are easily measured.

1.1. Background and previous work

There are, in general, four approaches for estimating an individualized ITD.

The first uses a listener's measured head-related impulse responses (HRIRs) and one of many ITD estimation techniques, most of which are discussed in the works by Xie [3], and Katz and Noisterning [5]. The second uses a 3D scan of a listener's head, and an acoustic ray-tracing algorithm to compute ITD [6]. The third approximates either the azimuthal variation of ITD by a finite-term Fourier series [7–9], or the three-dimensional spatial variation of ITD by a finite-term spherical-harmonic expansion [10]. Therefore, this approach obtains a closed-form equation that is a function of one [7–9], or both [10], spherical polar angle(s). The coefficients of this equation are computed by performing, typically, a least-squares fit of the equation to measured ITDs. A relationship between the coefficients and anthropometry is then derived using multiple linear regression analysis. The fourth approach approximates the head as a simple geometrical shape, and then uses a numerical technique, or a simple closed-form equation, to estimate ITD. For example, Duda *et al.* model the head as an ellipsoid and use geometrical acoustics to develop an algorithm to compute ITD [4]. Alternatively, Algazi *et al.* develop a spherical-head model (SHM) where the head radius is computed by performing a least-squares fit of the Woodworth and Schlosberg formula [11] to measured ITDs. A relationship between the head radius and anthropometry is then derived using multiple linear regression analysis [12].

The first approach produces the most accurate ITD estimates. It is not, however, a desirable solution for individualizing ITD because it requires specialized equipment like a loudspeaker or microphone array and, ideally, an anechoic chamber, to measure HRIRs.

Of the remaining approaches and models, the SHM is the simplest, primarily because it has only a single parameter - the head radius. However, it has two important deficiencies. The first is that elevation dependence (and, consequently, front-back asymmetries) observed in measured ITDs are not captured by the model. This causes estimation errors that may exceed typical ITD just-noticeable-difference values for sound sources located at large azimuths, and, as a result, may adversely affect the localization of these sources. The second is that it only estimates high-frequency (> 3 kHz) ITDs. Kuhn shows that, for low frequencies (< 500 Hz), the measured ITDs are up to 1.5 times larger than corresponding high-frequency ITDs. Although Algazi *et al.* suggest multiplying the ITDs obtained using the SHM by 1.5 to obtain individualized low-frequency ITDs [12], Kuhn presents the theory showing why this solution is only accurate for sound sources located near the median plane [1]. Furthermore, since it has been shown that ITD is the dominant localization cue at low frequencies [2], we suggest that a more thorough approach to estimating an individualized low-frequency ITD is required.

The ellipsoidal-head model, and the models adopting the second and third approaches, all address the first deficiency of the SHM but not the second. Additionally, the ellipsoidal-head model, which is not individualized from anthropometry, requires the measurement of five model parameters of which two, the backward and downward offsets of the pinnae, are ambiguously defined and non-trivial to measure [4]. The model developed by Gamper *et al.* needs specialized equipment like 3D scanners [6], while those proposed by Watanabe *et al.* [7, 8], and Zhong and Xie [9], are all restricted to the horizontal plane of the listener. Finally, the more recent model developed by Zhong and Xie does not reveal how different anthropometric features might influence ITD depending on the location of the sound source [10]. Therefore, while these models satisfactorily address the first deficiency of the SHM, most appear sufficiently complicated to preclude their use in consumer applications, and none provide much insight into how different anthropometric features influence ITD depending on the location of the sound source.

1.2. Objectives and approach

The present work has two objectives: 1) to develop a model that addresses both deficiencies of the SHM

while retaining its simplicity and 2) to gain insight into the direction-dependent effects of various anthropometric features on ITD.

To achieve these objectives, we develop a spherical-head model with a head radius that varies with interaural-elevation.[1] We use this head radius in the Woodworth and Schlosberg formula to estimate high-frequency ITDs. An equally simple formula to estimate low-frequency ITDs as a function of azimuth and head radius has been derived from the low-frequency limit of the analytical solution to scattering of sound off a rigid sphere [1]. We use this formula to estimate low-frequency ITDs. Furthermore, unlike any existing anthropometry-based models, we develop a pair of multiple linear regression equations relating head radius to anthropometry, one each for the high- and low-frequency cases, with regression coefficients that are themselves functions of interaural-elevation.

1.3. Paper outline

The rest of this paper is structured as follows. In section 2, we provide our formulations of the SHM and of our proposed model. In section 3, we use measured anthropometric data from the CIPIC database [13], and both models to estimate ITD for 12 listeners. Finally, in section 4, we provide some concluding remarks.

2. MODEL FORMULATION

We define the interaural coordinate system in section 2.1, followed by formulations of the SHM and our model in sections 2.2 and 2.3, respectively.

2.1. Interaural coordinate system

We adopt the interaural coordinate system used in the CIPIC database [13], with azimuth $\theta \in [-90^\circ, 90^\circ]$ and interaural-elevation $\phi \in [0, 360^\circ)$. Figure 1 depicts the interaural coordinate system in terms of the 1250 HRIR measurement points (25 azimuths, each with 50 elevations) used in the CIPIC database. We observe that by fixing azimuth and varying elevation, we remain in a plane that is parallel to the median plane of the listener.

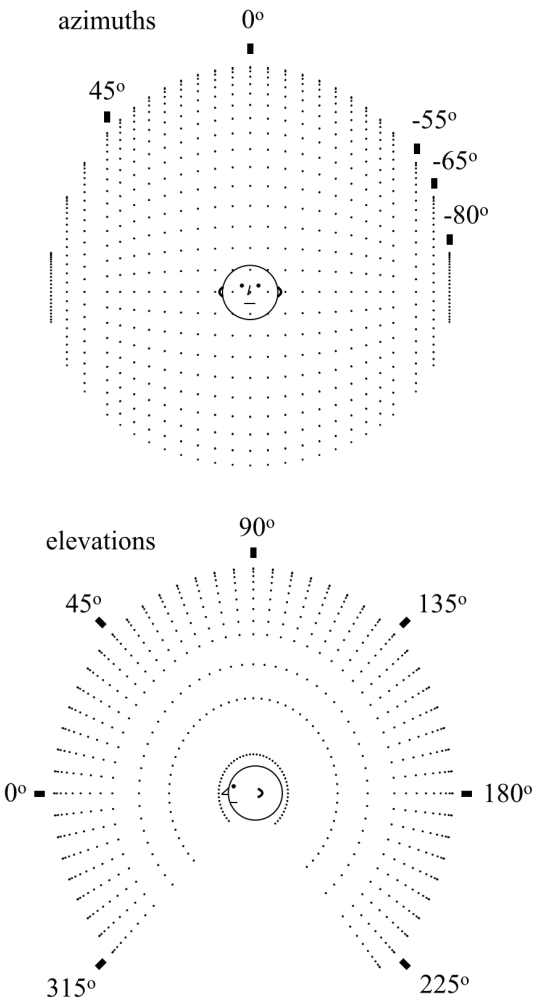

Fig. 1: Interaural coordinate system depicted in terms of actual HRIR measurement points used in the CIPIC database [13]. The 25 azimuths and 50 elevations make 1250 points in total. Azimuth spacing between $\pm 45^\circ$ is 5°; remaining azimuths are: $\pm 55^\circ$, $\pm 65^\circ$, and $\pm 80^\circ$. Elevation spacing is 5.625°. Top figure: front view; bottom figure: left-side view.

2.2. Spherical-head model

We begin by high-pass filtering, with a 1.5 kHz cut-off frequency, the HRIRs of a large number of listeners. The filtered HRIRs are then upsampled by factor of 8. We then use the thresholding technique, described, for instance, by Xie [3], with a 20% threshold value to compute measured high-frequency ITDs for all available sound source locations, for each listener. To compute the head radius, a, for each listener, we use the same least-squares minimization procedure followed by Algazi *et al.* [12]. This procedure computes a by minimizing the error between all available measured ITDs for that listener, and the corresponding estimates obtained using the

[1] Interaural-elevation is not the same as the elevation specified in a spherical polar coordinate system. We discuss this in a little more detail in section 2.1.

Woodworth and Schlosberg formula given by [11]

$$\tau_{\mathrm{W\&S}} = \frac{a(\sin\theta + \theta)}{c}, \quad (1)$$

where c is speed of sound, and $\tau_{\mathrm{W\&S}}$ is the estimated ITD. Once we have the head radii for all listeners, we perform a multiple linear regression on these head radii and the anthropometric data of the listeners using a model of the form

$$a_i = \alpha_0 + \sum_{k=1}^{3} \alpha_k x_{k,i}, \quad (2)$$

where a_i is the head radius for listener i, the αs are the model regression parameters, and the $x_{k,i}$s are the width, height, and depth of listener i's head for $k = 1$, 2, and 3, respectively. The regression yields values for the αs, which we then substitute back into Eq. (2) to produce a single equation relating head radius to the three anthropometric parameters.

Using the HRIRs and anthropometry of 25 listeners from the CIPIC database, the equation we obtain is

$$a = 0.033 + 0.283x_1 - 0.025x_2 + 0.098x_3, \quad (3)$$

where a is now specified in meters. In comparison, the corresponding equation derived by Algazi *et al.* is [12]

$$a = 0.032 + 0.255x_1 + 0.001x_2 + 0.090x_3. \quad (4)$$

The differences between Eqs. (3) and (4) are presumably because, unlike Algazi *et al.*, we do not smooth any discontinuities in the spatial variation of the measured high-frequency ITDs before computing a, and also because we may have used data for different listeners altogether, albeit from the same database. Algazi *et al.* perform smoothing because they suggest that the discontinuities are due to abrupt listener motion during HRIR measurement. We choose not to smooth the discontinuities because Katz and Noisternig show that such discontinuities still occur, even for a dummy head [5].

2.3. Proposed spherical-head model

We begin by defining interaural-elevation half-planes by considering all available azimuths at each elevation (see, for example, Fig. 1). For each listener, we then compute, using the same measured high-frequency ITDs as in section 2.2, a list of head radii, b_j, where j is the index of each half-plane. To do so, we follow the same least-squares minimization procedure described in section 2.2, one half-plane at a time, with a in Eq. (1) replaced by b_j. We then construct a matrix, $\boldsymbol{B}$, of head radii, with L rows, one for each listener, and E columns, one for each half-plane.

Using the head radius data in $\boldsymbol{B}$ one column at a time, we perform a sequence of E multiple linear regressions on these radii, and the anthropometric data of the L listeners to which they correspond, using a model of the form

$$b_{i,j} = \beta_{0,j} + \sum_{k=1}^{3} \beta_{k,j} y_{k,i}, \quad (5)$$

where $b_{i,j}$ is the head radius of the i^{th} listener at the j^{th} half-plane, the $\beta_{k,j}$s are the model regression parameters, and the $y_{k,i}$s are the width and depth of the head ($k = 1$, 2, respectively), and the pinnae mean arc length[2] ($k = 3$) of each listener. The regression yields values for the βs which we then substitute back into Eq. (5), one half-plane at a time, to produce a list of E equations relating head radius to these three anthropometric parameters.

We can now use these equations to estimate head radii from anthropometry for a given listener. These head radii are then used in place of a in Eq. (1), one half-plane at a time, to estimate high-frequency ITD for that listener. While the SHM requires a total of three multiplications to arrive at a single head radius from the anthropometry of a listener, our proposed model requires $3 \times E$ multiplications to produce E head radii for a listener. This is the extent of complexity our proposed model adds to that of the SHM.

We have just described the formulation of our model for estimating high-frequency ITD. To estimate low-frequency ITD, we use the same formulation as above, but with the following modifications.

First, we compute measured low-frequency ITDs as an interaural phase delay (IPD) difference averaged from 0 to 500 Hz. The IPD difference technique for computing measured ITD is described, for instance, by Xie [3]. We use the aforementioned averaging range because Kuhn shows that measured ITDs are generally frequency-independent below approximately 500 Hz [1].

[2] If we define the flare angle of the pinna as the angle made by the pinna with the side of the head (when viewed from above), we then define the pinnae mean arc length as the product of the mean of the maximum width of the left and right pinnae, and the mean of their flare angles in radians.

Next, in place of the Woodworth and Schlosberg formula, we use the equation corresponding to ITD evaluated from the interaural phase delay that results when low-frequency plane waves scatter off a rigid sphere. This equation is given by [1]

$$\tau_{\mathrm{LF}} = \frac{3\,b\sin\theta}{c}, \tag{6}$$

where b is head radius, and τ_{LF} is estimated ITD. We use this equation because Kuhn shows that the resulting ITD is a good approximation of the measured low-frequency ITD [1].

Finally, instead of using Eq. (5) as the multiple linear regression model, we use a model of the form

$$b_{i,j} = \beta_{0,j} + \sum_{k=1}^{2} \beta_{k,j}\, y_{k,i}, \tag{7}$$

where the variables have the same meanings as before. The difference between Eqs. (5) and (7) is in the number of anthropometric parameters and, consequently, model regression parameters used. In this case, the $y_{k,i}$s are the width and depth of the head, for $k = 1$ and 2, respectively, of listener i.

Using the HRIRs and anthropometry of the same 25 listeners used in section 2.2, we compute two separate sets of $E = 50$ equations relating head radius to anthropometry. In Fig. 2, we plot, as a function of elevation, the contributions, $\beta_k y_k$, of the different anthropometric features towards the head radii for a randomly selected listener (from those 25) for each frequency regime. The anthropometry of the chosen listener are $y_1 = 0.15\,\mathrm{m}$, $y_2 = 0.21\,\mathrm{m}$, and $y_3 = 0.018\,\mathrm{m}$.

We observe that the contribution of the pinnae mean arc length to head radius increases for sound sources located behind the listener. This is likely due to the effective increase in path length to the contralateral ear caused by the pinnae flare angle, for high-frequency sounds originating behind the listener. We also observe that the contribution of head depth for these directions decreases significantly when estimating high-frequency ITD, but remains relatively constant when estimating low-frequency ITD. In general, we observe that head width is the dominant feature in determining high-frequency ITD, whereas head depth dominates at low frequencies. Finally, we observe that the contribution of head width decreases sharply just beyond 90° when estimating high-frequency ITDs.

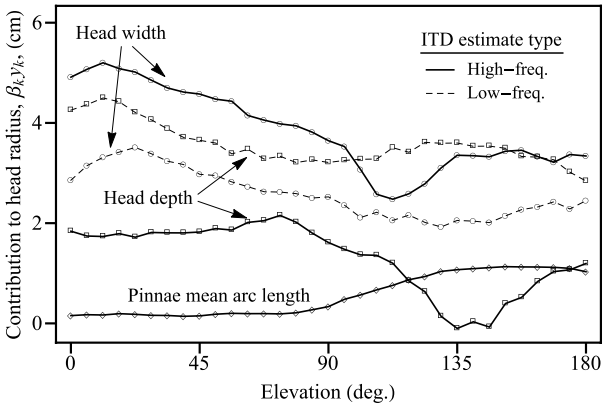

Fig. 2: Contribution of the different anthropometric features to head radius as a function of elevation. The solid and dashed lines show contributions to head radius when estimating high- and low-frequency ITDs, respectively.

One anthropometric parameter commonly thought to be necessary to include in an ITD model is the offset of the pinnae towards the back of the head [4]. Although we do not explicitly consider this parameter, it is implicitly included in our model by individualizing the head radius one half-plane at a time. This also simplifies the use of our model because measuring ear offsets accurately is not a straightforward task [4].

3. MODEL EVALUATION

We begin with the measured anthropometry of 12 new listeners from the CIPIC database and estimate their ITDs using both models. We also compute the measured ITDs for these listeners using the same techniques described in section 2 to serve as a benchmark against which the estimates are compared. We then compute an *absolute ITD error*, ε, for each of the 1250 sound source locations (see Fig. 1) and all 12 listeners, for both models. We define ε as the absolute value of the difference between measured and estimated ITD for a given sound source location. Using ε, we compute a *mean absolute ITD error*, ζ, for a given model, by averaging the εs first over all 50 elevations, and then all 12 listeners.

The ζs are plotted, in Fig. 3, as a function of azimuth, along with error bars representing one standard deviation of the mean (across listeners). We observe that, at large azimuths ($|\theta| \geq 45°$), the ζs for the proposed model are smaller in magnitude than those for the SHM, whereas at small azimuths ($|\theta| \leq 20°$), they are approximately the same. At these large azimuths, for high-frequency ITD, the mean absolute ITD errors generated by the proposed

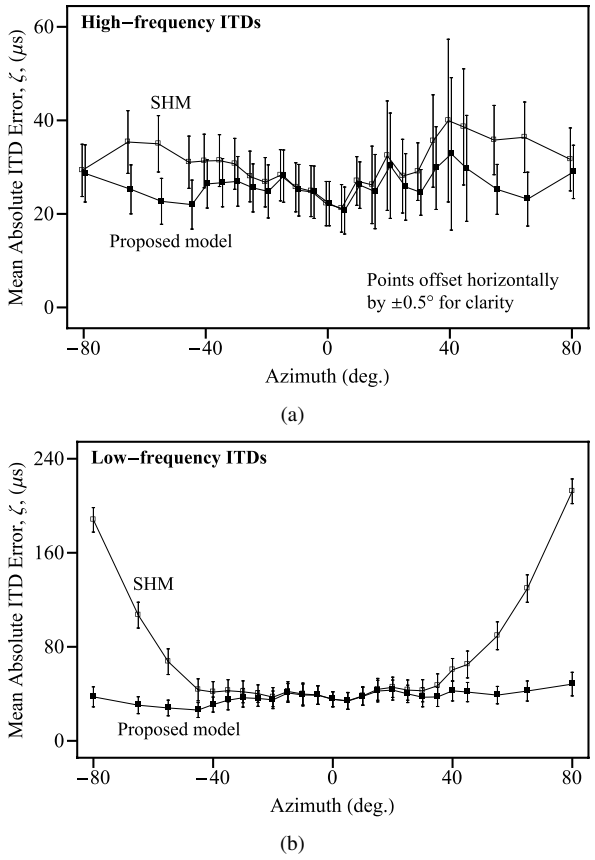

Fig. 3: Variation of mean absolute ITD error, ζ, with azimuth, for estimates of (a) high-frequency ITDs, and (b) low-frequency ITDs. The error bars represent one standard deviation of the mean computed when averaging over 12 listeners.

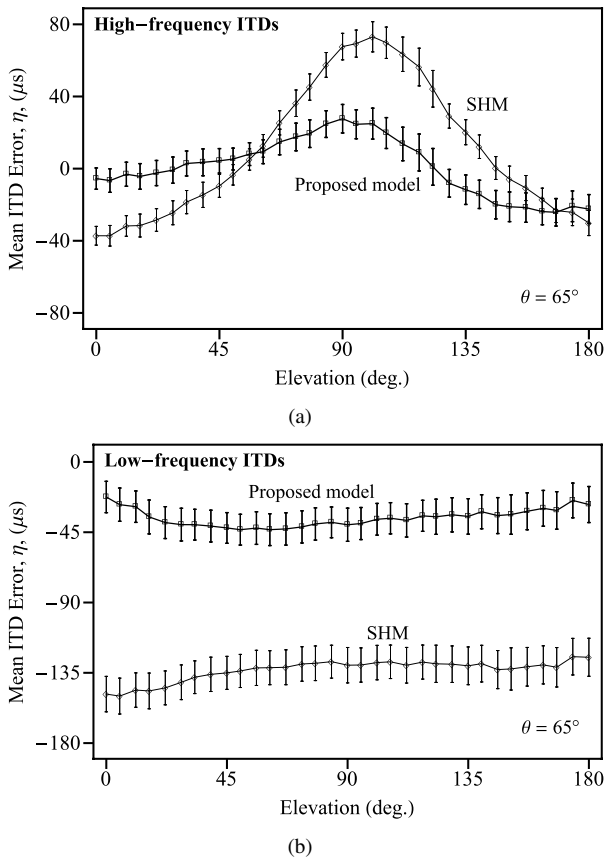

Fig. 4: Variation of mean ITD error, η, with elevation for $\theta = 65°$. (a) and (b) show high- and low-frequency ITDs respectively. The error bars show one standard deviation of the mean.

model are reduced by up to $13\,\mu$s (3% of the measured ITD value), compared to the SHM. For low-frequency ITDs, this reduction is up to $160\,\mu$s (23%).

We also define, for each model, a *mean ITD error*, η, as the average, over all listeners, of the difference between measured and estimated ITD for a given sound source location. Unlike for ζ, we do not take the absolute value of the difference before averaging. This means that negative η indicates that the model overestimates the measured ITD, and vice versa.

We plot, in Fig. 4, the variation, with elevation, of mean ITD error, for $\theta = 65°$. We observe, from Fig. 4(a), that for most sound sources in front of the listener, the SHM significantly overestimates measured high-frequency ITD, while for most sources above the listener, it significantly underestimates this ITD. In contrast, although the proposed model also underestimates ITD for sources above, and overestimates ITD for those behind the listener, the mean ITD error is almost always smaller in magnitude compared to the SHM. However, we also observe that η for both models are comparable for sources located behind the listener close to the horizontal plane. This can be explained by observing that the shape of the back of most heads, looking from above, is largely spherical between the two pinnae (see, for example, Fig. 6 in the work by Ball *et al.* [14]). Finally, we observe, from Fig. 4(b), that both models consistently overestimate measured low-frequency ITD, albeit by noticeably different amounts.

We also point out, that both Figs. 3(b) and 4(b) provide experimental evidence that multiplying high-frequency

ITD by 1.5 to estimate low-frequency ITD [12] is not accurate for sound sources at large azimuths and any elevation. Figure 3(b), however, shows that this approximation is sufficiently accurate near the median plane.

4. CONCLUSIONS

The spherical-head model (SHM) developed by Algazi *et al.* [12] provides a simple equation relating head width, height, and depth to head radius (see Eqs. 3 and 4), which is then used in the Woodworth and Schlosberg formula to estimate individualized high-frequency ITD. However, we showed that it does not capture the elevation dependence (and, consequently, front-back asymmetries) observed in measured ITDs (see Fig. 4(a)). As a result, the SHM produces inaccurate ITD estimates, especially at large azimuths (see Fig. 3(a)). We also showed experimentally that the suggestion by Algazi *et al.* to multiply the individualized high-frequency ITD by 1.5 to estimate a low-frequency ITD produces inaccurate estimates at large azimuths (see Fig. 3(b)).

To address the deficiencies of the SHM, we developed a spherical-head model with a head radius that varies with interaural elevation. We also derived two sets of relationships between head radius and anthropometry, one each for estimating high- and low-frequency ITD, respectively. Figure 4(a) shows that the proposed model is able to capture the elevation dependence in measured ITDs, resulting in smaller errors at large azimuths compared to the SHM (see also, Fig. 3). Furthermore, since our model allows the contributions of the anthropometric parameters to head radius to also vary with elevation, we observed how the influence of various anthropometric features on ITD depends on the direction of the sound source. For instance, we observed that the flare angle of the pinnae affects high-frequency ITD only for sound sources located behind the listener, whereas head depth primarily affects this ITD for sound sources located in front (see Fig. 2). Additionally, we found the head width to be the dominant feature influencing ITD at high frequencies, whereas head depth dominates at low frequencies.

ACKNOWLEDGEMENTS

This research was conducted under a contract from the Sony Corporation of America. The authors thank Joseph G. Tylka for valuable discussions, and recommendations regarding the presentation of the content of this paper.

5. REFERENCES

[1] G. F. Kuhn. "Model for the interaural time differences in the azimuthal plane". *The Journal of the Acoustical Society of America*, 62(1):157–167, 1977.

[2] F. L. Wightman and D. J. Kistler. "The dominant role of low-frequency interaural time differences in sound localization". *The Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992.

[3] B. Xie. *Head-related transfer function and virtual auditory display*. J Ross, 2013.

[4] R. O. Duda, C. Avendano, and V. R. Algazi. "An adaptable ellipsoidal head model for the interaural time difference". In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 2, pages 965–968. IEEE, 1999.

[5] B. F. G. Katz and M. Noisternig. "A comparative study of interaural time delay estimation methods". *The Journal of the Acoustical Society of America*, 135(6):3530–3540, 2014.

[6] H. Gamper, M. R. P. Thomas, and I. J. Tashev. "Estimation of multipath propagation delays and interaural time differences from 3-D head scans". Microsoft Research, 2015.

[7] K. Watanabe, Y. Iwaya, J. Gyoba, Y. Suzuki, and S. Takane. "An investigation on the estimation of interaural time difference based on anthropometric parameters". *Trans. the Virtual Reality Soc. of Jpn*, 10:609–617, 2005.

[8] K. Watanabe, K. Ozawa, Y. Iwaya, Y. Suzuki, and K. Aso. "Estimation of interaural level difference based on anthropometry and its effect on sound localization". *The Journal of the Acoustical Society of America*, 122(5):2832–2841, 2007.

[9] X. Zhong and B. Xie. "A novel model of interaural time difference based on spatial Fourier analysis". *Chinese Physics Letters*, 24(5):1313, 2007.

[10] X. Zhong and B. Xie. "An individualized interaural time difference model based on spherical harmonic function expansion". *Chinese Journal of Acoustics*, 3:10, 2013.

[11] R. S. Woodworth and H. Schlosberg. *Experimental psychology*. Holt, 1954.

[12] V. R. Algazi, C. Avendano, and R. O. Duda. "Estimation of a spherical-head model from anthropometry". *Journal of the Audio Engineering Society*, 49(6):472–479, 2001.

[13] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. "The CIPIC HRTF database". In *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*, 2001.

[14] R. Ball, C. Shu, P. Xi, M. Rioux, Y. Luximon, and J. Molenbroek. "A comparison between Chinese and Caucasian head shapes". *Applied Ergonomics*, 41(6):832–839, 2010.