Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions

Griffin D. Romigh, *Member, IEEE*, Douglas S. Brungart, *Member, IEEE*, Richard M. Stern, and Brian D. Simpson

***Abstract*—Several methods have recently been proposed for modeling spatially continuous head-related transfer functions (HRTFs) using techniques based on finite-order spherical harmonic expansion. These techniques inherently impart some amount of spatial smoothing to the measured HRTFs. However, the effect this spatial smoothing has on the localization accuracy has not been analyzed. Consequently, the relationship between the order of a spherical harmonic representation for HRTFs and the maximum localization ability that can be achieved with that representation remains unknown. The present study investigates the effect that spatial smoothing has on virtual sound source localization by systematically reducing the order of a spherical-harmonic-based HRTF representation. Results of virtual localization tests indicate that accurate localization performance is retained with spherical harmonic representations as low as fourth-order, and several important physical HRTF cues are shown to be present even in a first-order representation. These results suggest that listeners do not rely on the fine details in an HRTF's spatial structure and imply that some of the theoretically-derived bounds for HRTF sampling may be exceeding perceptual requirements.**

***Index Terms*—Head-related transfer functions (HRTFs), spherical harmonic, spatial hearing.**

I. Introduction

THE perceptual cues necessary for accurate sound source localization can be adequately modeled by a head-related transfer function (HRTF). Because the HRTF contains all of the relevant acoustic localization cues, once measured it can be used to impart directional information on any single-channel sound by filtering the sound signal with the HRTF from each ear and presenting the result binaurally over headphones. Virtual auditory displays (VADs) and other technologies built around HRTF processing provide powerful tools to investigate the perceptual mechanisms of spatial hearing and take advantage of a person's natural spatial hearing abilities. Undermining these efforts is the fact that the HRTF measured for a single ear is highly variable across individual listeners, three-dimensional space, and frequency (or time), making the generalization or customization of spatial-audio-based technology difficult. One avenue that provides hope for overcoming these challenges is the development of spatial HRTF models that are capable of representing the perceptually-relevant features of an HRTF efficiently.

A substantial amount of recent research has focused on the expansion of HRTFs onto a set of spherical basis functions called the spherical harmonics (SHs) [1]–[5]. Because spherical harmonics are spatially continuous and orthonormal, this type of expansion has the potential to provide effective solutions for several HRTF-related challenges, including efficient measurement [3], interpolation [1], [2], [4], rendering [6], compact parameterization [7], and database composition [5]. From a mathematical standpoint, these studies have shown that SH expansion can have substantial advantages over other computational models of the HRTF. However, until now these advantages have only been shown with respect to error metrics based on arithmetic differences between the spectra of the originally-measured HRTFs and those of the approximate HRTFs obtained from reduced-order SH expansions. These arithmetic differences are likely to be correlated with perceptual differences between the measured and modeled HRTFs, but without a better understanding of the specific cues listeners use to extract spatial information from the HRTF, it is difficult to make a direct prediction about the impact that a certain level of spectral error in the HRTF will have on how it is perceived by the listener. Consequently, very little is known about the impact that HRTFs produced with reduced-order SH expansion have on the perception of virtual sounds. This paper attempts to help fill this void by providing insight into the effect of spherical harmonic expansion on sound source localization in an HRTF-based virtual auditory display. Section II provides a brief overview of relevant literature necessary to explain the SH-based HRTF expansion described in Sections III and IV. Section V then presents the methodology and results of a perceptual localization experiment investigating the tradeoffs between localization accuracy and SH representation order. Finally, Section VI discusses the perceptual results in the context of previous results and remaining research questions.

II. Background

Much progress has been made in understanding how the different acoustic details contained in the HRTF impact the perceptual properties of virtual sounds. Of particular importance

Manuscript received July 15, 2014; revised December 05, 2014 and February 23, 2015; accepted March 09, 2015. Date of publication April 28, 2015; date of current version July 14, 2015. This work was supported by tuition and graduate stipend provided by the SMART scholarship program. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Sascha Spors.

G. D. Romigh and B. D. Simpson are with the 711th Human Performance Wing, Air Force Research Laboratory, Dayton, OH, 45435 (e-mail: griffin.romigh@us.af.mil; brian.simpson.4@us.af.mil).

D. S. Brungart is with Walter-Reed National Military Medical Center, Bethesda, MD 20889 USA (e-mail: douglas.s.brungart.civ@health.mil).

R. M. Stern is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: rms@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2421876

1932-4553 © 2015 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

to this study is the finding that human listeners are insensitive to the fine details in the phase spectrum of an HRTF in both localization [8] and discrimination tasks [9], as long the frequency-independent interaural time delay (ITD) at low frequencies remains intact. This means that a perceptually-valid HRTF corresponding to a single location can be encoded using only the two (right and left) monaural magnitude spectra of an HRTF and a single ITD value [8].

Other studies have also shown that listeners are insensitive to fine spectral details in the magnitude spectrum of the HRTF. These studies have used a variety of techniques for systematically varying the spectral detail in the HRTF, including techniques based on principal component analysis of the spectrum [8], techniques based on expansion of the spectrum onto a truncated Fourier basis [10], and methods that only retain the average power measured over overlapping spectral bands [11].

Less information is known about the perceptual importance of *spatial* detail in an HRTF. Some information concerning an HRTF's spatial structure has been gained through studies concerned with HRTF spatial interpolation. Spatial interpolation is an important practical challenge because while the HRTF itself is a continuous function of three-dimensional space, individualized HRTF measurements are usually and most easily measured only at discrete spatial locations (see [12] for an exception). This means that traditional HRTF measurements must be interpolated to allow for virtual sound source presentation at any non-measured spatial locations where playback is desired. While a number of techniques have been proposed for the HRTF interpolation problem, relatively few have been perceptually evaluated in terms of the resulting localization accuracy, typically relying on less perceptually-informative arithmetic metrics based on spectral reconstruction error. Some localization results have been obtained, however, including those of [13], and [14], demonstrating that localization performance remained unaffected when an HRTF is interpolated from a measurement grid with as much as 20 degrees between sample locations. This perceptually-minimal resolution stands in direct contrast to the physical resolution requirements laid out by [15], who showed that measurements must be made at a minimum resolution of approximately five degrees in the horizontal plane to capture all of the acoustic details in an HRTF.

When taken together, the results of these HRTF interpolation studies suggest that many of the spatial details contained in an HRTF are also perceptually irrelevant to localization, just as it has been shown that many of the spectral details in the magnitude and phase of the HRTF are largely irrelevant to the perceived locations of sounds. This concept is exploited by a growing class of HRTF interpolation and modeling techniques that expand the HRTF onto spherical harmonic basis functions, a set of orthonormal functions defined on the sphere. The main advantage of this technique is that a continuous HRTF at all spherical angles can be modeled with a relatively small set of SH expansion coefficients (compared to the necessity of multiple filter coefficients for each spatial location using traditional methods). One of the first applications of SH-related techniques to the interpolation of HRTFs was described by [16]. This study investigated the expansion of HRTFs first onto a set of statistically-optimal complex basis functions via Karhunen-Loeve Expansion (KLE), using a SH-based regularized fitting method (spherical thin-plate splines) to obtain KLE parameters at locations that had not been measured. While no perceptual evaluation was included in [16], [14] later adapted the technique to use PCA coefficients and obtained the interpolation results discussed previously.

A more traditional and straightforward approach is to expand the HRTF itself (in one form or another) directly onto a set of spherical harmonic basis functions. The resulting SH weights (coefficients) can then be used to calculate the HRTF at any spherical angle (as detailed in Section III). With this formulation, the spatial detail contained in an HRTF can also be systematically smoothed by a simple truncation of the expansion order. [7], one of the first studies applying spherical harmonic expansion directly to HRTFs, showed that 90% of the spatial energy in both the magnitude and phase of an HRTF could be separately modeled with only a seventh-order SH representation (64 expansion coefficients). This effort appears to have been the first to directly analyze the HRTF in the SH domain. While Evan's technique required a somewhat impractical set of measurement locations to enable direct computation of SH coefficients, similar techniques have since been investigated by [1] and [17] for fitting an HRTF's complex frequency response based on arbitrarily spaced measurements.

More recently, SH-based interpolation techniques have been introduced based on the reciprocal view of HRTFs introduced by [18]. Using the reciprocity principle, an HRTF can be measured by recording the signal arriving at a particular location in space when a signal is presented from a miniature driver placed inside a subject's ear canal, the reciprocal setup of conventional HRTF collection techniques. Within this framework, methods based on wavefield expansion can be applied to HRTFs, allowing for accurate interpolation in spherical angle, range, and frequency [2]. Because this model represents a physical wave, theoretical bounds can be placed on the problem to determine the required SH truncation order and the number of spatial locations that need to be sampled given a particular source bandwidth. [2] showed that if one assumes a 15-kHz bandwidth for a typical HRTF measurement, the theoretical bound corresponds to a 34th-order model requiring measurements at over 1200 spatial locations. These numbers are comparable to the physical resolution requirements described earlier by [15] for the horizontal plane and again highlight the large discrepancy between the number of spatial measurements required to capture all of the physical properties of the HRTF and the comparatively small number of measurements required when simple interpolation strategies are used.

The HRTF interpolation literature seems to suggest that some amount of spatial detail in an HRTF is perceptually irrelevant, a finding that SH-based HRTF models are well suited to exploit. Unfortunately, little perceptual evaluation of SH-based techniques has taken place and, without a firm understanding of the perceptual salience of spatial detail, current SH-based models seem to be exceeding the perceptual requirements for accurate localization, limiting their utility as practical models for HRTFs. The present study seeks to examine the tradeoff between the truncation of a SH-based representation and perceptual accuracy in the resulting virtual auditory display.

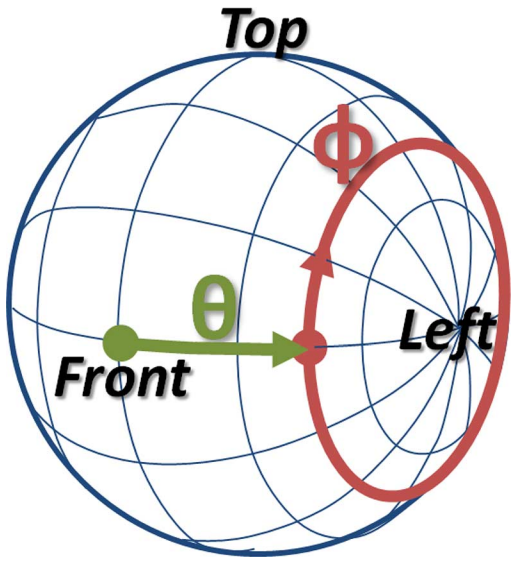

Fig. 1. The interaural polar coordinate system.

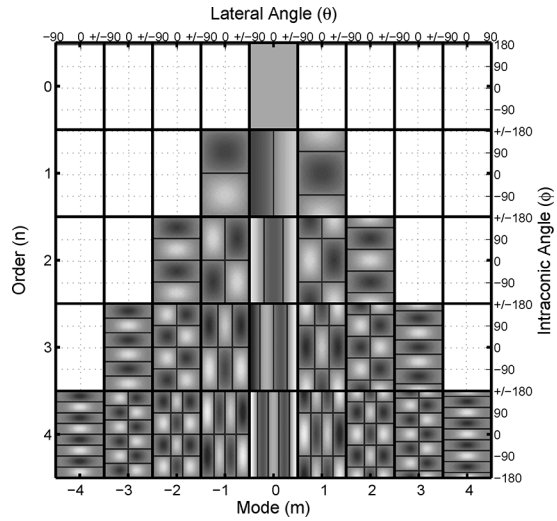

Fig. 2. Real spherical harmonic basis function for orders 0 through 4. Each panel contains the basis function of indicated order (row) and mode (column) plotted as a function of lateral angle (abscissa) and intraconic angle (ordinate). Black lines indicate the 0- value contours. Dark shading indicates regions of negative value while light regions indicate regions of positive value.

III. Spherical Harmonic Expansion

Depending on application-specific convenience and author proclivities, the precise definition of the spherical harmonic basis functions differs greatly, so even within the field of acoustics care must be taken when comparing equations across the literature. For this study, special attention is called to the fact that the spherical coordinate system over which the basis functions are defined was chosen to be the interaural-polar coordinate system common to auditory perception literature and shown in Fig. 1 rather than the conventional vertical-polar spherical coordinate system commonly employed in the field of physical acoustics. It is also worth noting that the real form of the orthonormal spherical harmonics was utilized in order to provide a model with real expansion coefficients. A detailed comparison of the real and complex spherical harmonics in the context of wave-field expansion is given in [19].

Assuming the interaural polar coordinate system as shown in Fig. 1, the real spherical harmonic basis functions are defined in (1) for integer indexes of $n : [0, \infty]$ and $m : [-n, n]$.

$$\begin{aligned} Y_{nm}(\phi,\theta) &= \sqrt{\frac{(2n+1)}{4\pi}} P_n^m(\sin(\theta)) \\ &\quad \text{for } m = 0 \\ &= \sqrt{\frac{(2n+1)}{2\pi}\frac{(n-|m|)!}{(n+|m|)!}} P_n^m(\sin(\theta))\cos(m\phi) \\ &\quad \text{for } m > 0 \\ &= \sqrt{\frac{(2n+1)}{2\pi}\frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\sin(\theta))\sin(|m|\phi) \\ &\quad \text{for } m < 0 \end{aligned} \tag{1}$$

In the equations above, P_n^m represents the associated Legendre polynomial of order n and degree m, as defined by [20]. The real spherical harmonic basis functions, shown in Fig. 2 for orders 0 through 4, form a complete orthonormal basis for any square-integrable function defined in terms of spherical coordinates. This means an arbitrary continuous spherical function $f(\phi, \theta)$ (assumed to be spatially band limited) can be expanded onto a finite set of spherical harmonics as in (2).

$$f(\phi,\theta) = \sum_{n=0}^{P}\sum_{m=-n}^{n} Y_{nm}(\phi,\theta) C_{nm} \tag{2}$$

The set of weights, C_{nm}, are known as the spherical harmonic coefficients and carry information about how the function $f(\phi, \theta)$ varies across space. The number of these coefficients is dictated by the maximum SH expansion order (a.k.a. truncation order), P, and can be shown to be equal to $(P+1)^2$. Because of the orthonormal properties of the spherical harmonics, for an arbitrary continuous spherical function $f(\phi, \theta)$, the spherical harmonic weights can be calculated directly as shown in (3).

$$C_{nm} = \int_0^{2\pi}\int_0^{\pi} f(\phi,\theta) Y_{nm}(\phi,\theta)\cos\theta \, d\phi \, d\theta \tag{3}$$

A. Least-Squares Fitting

Unfortunately, in most practical situations only samples of the underlying continuous spatial function are available. For this situation, an exact solution only exists for a select number of predefined measurement grids such as the one used by [7]. For arbitrary grids, the SH coefficients are typically estimated by forming a system of linear equations using the discretized version of (2) repeated S times, one for each spatial location $\{\phi_i, \theta_i\}_{i=1}^{S}$ [1], [17]. This system is given in matrix form by (4).

$$\mathbf{f} = \mathbf{Y}\mathbf{c} \tag{4}$$

$$\begin{aligned} \text{where } \mathbf{f} &= [f(\phi_1,\theta_1), f(\phi_2,\theta_2), \cdots, f(\phi_S,\theta_S)]^T \\ \mathbf{c} &= [C_{00}, C_{1-1}, C_{10}, C_{11}, \cdots, C_{PP}]^T \\ \mathbf{Y} &= [\mathbf{y_{00}}, \mathbf{y_{1-1}}, \mathbf{y_{10}}, \mathbf{y_{11}}, \cdots, \mathbf{y_{PP}}] \\ &\quad \text{and} \\ \mathbf{y_{nm}} &= [Y_{nm}(\phi_1,\theta_1), \cdots, Y_{nm}(\phi_S,\theta_S)]^T \end{aligned} \tag{5}$$

In this form, we can use the Moore-Penrose pseudo-inverse to find the unique least squares estimate of the coefficient vector $\mathbf{c}$ according to (6), provided we have more spatial samples than the number of spherical harmonic coefficients we are trying to calculate $(S > (P+1)^2)$.

$$\hat{\mathbf{c}} = (\mathbf{Y}^{\mathbf{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathbf{T}}\mathbf{f} \tag{6}$$

In practice, due to noise in the measurements this method can require up to twice as many spatial samples as the number of SH coefficients $(S > 2(P+1)^2)$ to obtain satisfactory results and may need additional regularization if entire portions of the sphere are not sampled, a scenario that often arises due to the absence of HRTF measurements at low elevations in typical setups [1], but not a concern for the measurement setup utilized in the present study.

IV. HRTF Collection and Modeling

At the start of each experimental session, HRTFs were measured for the subject using the procedure described in [21]. In short, the subject was outfitted with a pair of custom-made in-ear microphones and situated in the center of a 7-foot-radius geodesic speaker array housed in the large anechoic chamber at Wright-Patterson AFB, OH, as pictured in Fig. 3. Subjects were instructed to remain still but were left unconstrained while periodic chirp signals with a 200-Hz to 15-kHz bandwidth were presented sequentially from 277 loudspeaker locations and recorded binaurally. The resulting recordings were used along with the presentation signal to calculate the HRTF for each location and ear using frequency-domain division. Converting the HRTFs back to the time domain, the head-related impulse responses (HRIRs) were then windowed with 10-ms Hamming windows to reduce the effects of residual echoes within the facility, and filtered to compensate for the response of the amplifiers, speakers, and headphones. ITDs were extracted from the HRIRs by finding the slope of the best-fit line to the unwrapped phase difference between the right and left HRIRs between 300 Hz and 1500 Hz. Lastly, the HRIR for each ear and location was converted to a minimum-phase filter and truncated to 256 taps.

The minimum-phase filters that resulted were then converted to the frequency domain using a 256-point DFT, and the magnitude of each DFT coefficient was converted into decibels to better reflect the nonlinear scaling of the peripheral auditory system. Because the magnitude response was even, only $K/2+1$ magnitude coefficients of a K-point DFT were unique, meaning that the HRTF for a single location could be encoded with $K+3$ real-valued parameters, $K/2+1$ parameters representing the unique decibel-magnitude for each ear along with a single ITD value. The spatial measurement vector, $\mathbf{f}$, was then formed by one parameter (e.g., the ITD) at each location and arranging them in a column vector. In this way each HRTF parameter was treated as a spatial function and expanded independently using the least-squares spherical harmonic fitting discussed above.

From the resulting SH coefficients, the HRTF parameters were then calculated for 245 test locations (as discussed below)

Fig. 3. Auditory Localization Facility at Wright-Patterson AFB, Dayton, OH.

by using (4) with the coordinates of the S measurement locations replaced by the coordinates of the 245 test locations. For SH HRTFs of order P, only the coefficients and SH basis functions corresponding to orders 0 through P were included in the calculation. HRTF filters were then reconstructed for each ear by using the symmetry property of the HRTF magnitude, converting the magnitude to the linear domain, calculating the 256-point inverse DFT, and delaying the contralateral ear filter by the corresponding ITD. A set of baseline virtual HRTFs was also created for each test location by interpolating the measured HRTFs for each subject using the conventional nearest neighbor (NN) approach described in detail by [13] and [14].

In the methods of Section III, the spatial function, $f(\phi,\theta)$, was left as an arbitrary function to illustrate that the methods behind spherical harmonic expansion do not depend on what spherical function is being expanded. As mentioned above, [7] proposed using real SH basis functions and expanding the HRTF magnitude and phase independently, while [1] and [17] expanded the complex frequency response onto complex SH basis functions. The choice of HRTF parameters utilized here (right and left decibel magnitude spectra and the corresponding ITDs) were chosen because the current work is focused on defining a perceptually-efficient HRTF representation (rather than one that is focused on physical accuracy), and this form of representation has been shown to preserve all of the perceptually-relevant information contained in a measured HRTF for a single location [8] while providing a saving in terms of the number of parameters describing an HRTF at a single location. Additionally, the decibel-magnitude HRTF parameters used here seem to better preserve contralateral spectral structure when the SH representation is truncated to low orders. Fig. 4 shows the RMS spectral error for left-ear HRTFs along the horizontal plane when modeled using in-house implementations of the methods proposed by [1] and [7], and those presented here. All three methods produced reasonable results both in terms of modeling error and in casual listening tests for

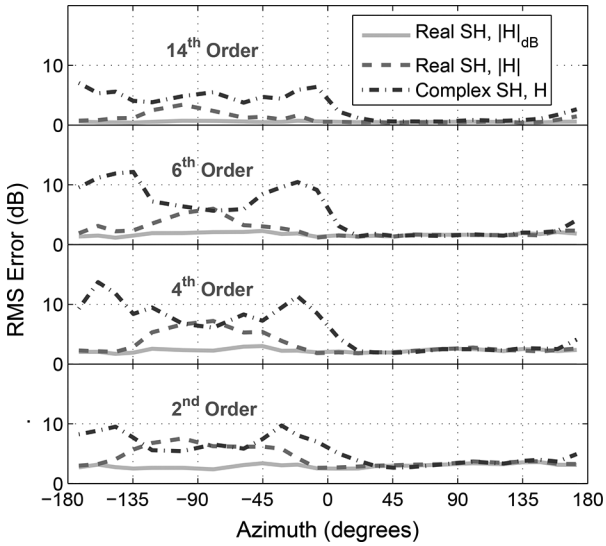

Fig. 4. RMS modelling error for left HRTF magnitude spectra on the horizontal plane using three different spherical harmonic representation schemes found in the literature for various SH orders.

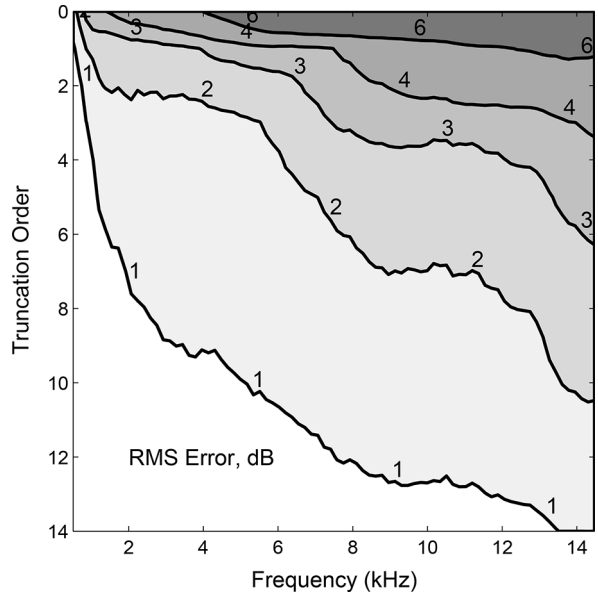

Fig. 5. RMS modeling error (in dB) using the proposed SH technique as a function of truncation order (ordinate) and frequency (abscissa). Black lines follow the indicated equal-value contour lines.

high-order representations. However, as the truncation order was reduced, the current implementation (Real SH, $|H|_{dB}$) seemed to do the best job preserving contralateral HRTFs (negative azimuths in Fig. 4).

With the current spherical harmonic representation, the amount of spatial detail contained in the HRTF can be systematically reduced by decreasing the SH truncation order P. Because the higher-order SH basis functions capture more rapid spatial variation, truncation of the expansion leads to progressively smoother spatial representations of the HRTF. While the progressive truncation is beneficial to modeling efficiency (lower-order models require fewer SH coefficients), additional smoothing will inherently increase the amount of error between the HRTF model and the underlying measurements. Fig. 5 shows the average RMS spectral error (in decibels) for HRTFs modeled with the indicated truncation order as a function of frequency. As can be seen, the greatest decrease in spectral error occurs in going from a zeroth-order to a first-order representation. Because the zeroth-order basis function contains no spatial variation, these coefficients actually represent the average or "diffuse field" HRTF, and the first-order coefficients are the first to provide any actual spatial detail. In general, the amount of modeling error for a given truncation order increases as a function of frequency. However, for the 15-kHz bandwidth used in the current investigation, modeling error is always within approximately one dB for a 14th-order representation.

V. Perceptual Evaluation of Truncated Spherical Harmonic HRTFs

While previous research suggests that high-order SH representations contain sufficient detail to preserve localization accuracy found with measured HRTFs [2], it is unclear how the modeling errors apparent in the current low-order truncated SH model will affect perception. In order to examine the effects of SH model truncations and to identify the minimal SH truncation order that preserves localization accuracy, a virtual localization task was conducted in which performance using full-resolution individualized HRTFs was compared with performance with HRTFs that had been expanded using SH models of various truncation orders following the method described above.

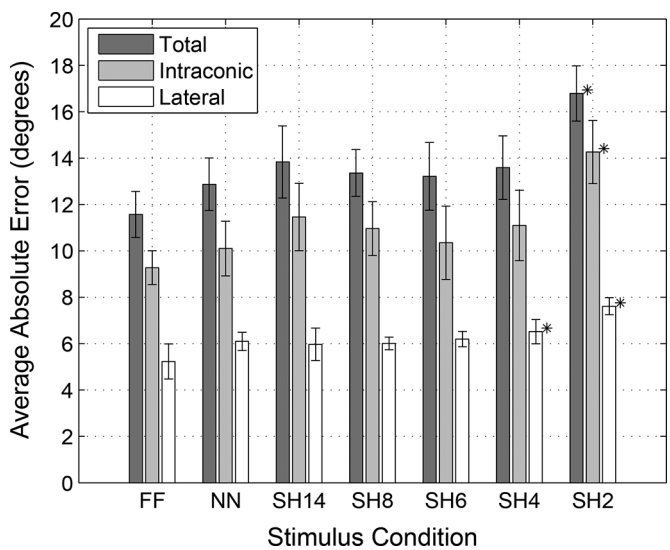

Fig. 6. Localization results averaged over all subjects and locations for each experimental test condition. Results are given in terms of total absolute angular error (Total), the average error component which fell along the interaural axis (Lateral), and the average error component which fell along a cone of confusion (Intraconic). Error bars represent 95% confidence intervals, and asterisks represent results significantly different than free-field (FF).

A. Stimuli

The stimuli were 250-ms bursts of white noise, which had been bandpass filtered between 500 Hz and 15 kHz and windowed with 10-ms onset and offset ramps. On each trial, a target stimulus was presented at a location that corresponded to one of 245 speaker locations above -45 degrees in elevation. Low

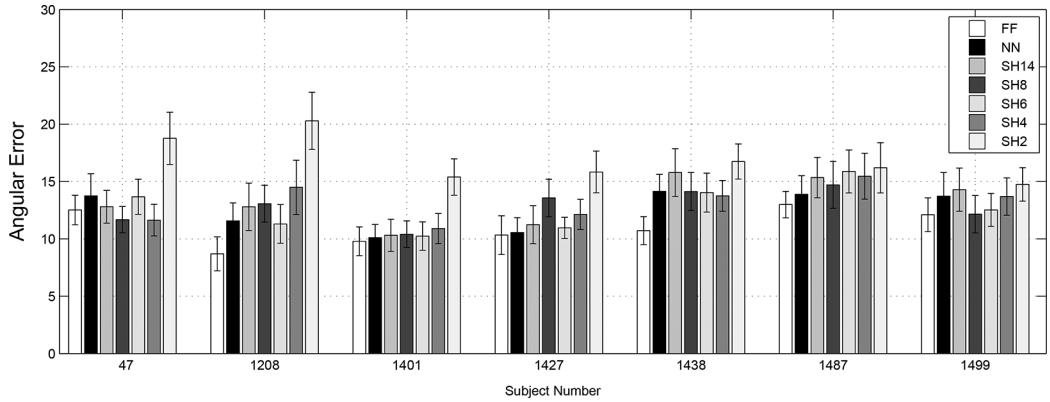

Fig. 7. Localization results averaged over all locations for each experimental test condition and subject. Results are given in terms of total absolute angular error. Error bars represent 95% confidence intervals.

elevations were excluded from testing because of interference from the subject platform contained in the Auditory Localization Facility (ALF). Virtual stimuli were convolved with the appropriate HRTF based on the test condition and location and presented to the subject through a pair of custom earphones (described in [21]). Within a block of trials, only one HRTF condition was tested, and the order in which the subjects completed the blocks was randomized. Prior to the virtual localization task, subjects also completed three 60-trial blocks in which stimuli were presented from the physical speaker locations.

B. Experimental Procedure

Seven subjects with normal hearing (standard audiometric thresholds within 20-dB HL at all frequencies) participated in experimental sessions over the of course four weeks. At the beginning of each 30-minute experimental session, a set of HRTFs and the corresponding headphone correction were measured using the procedure outlined above. This overall process from microphone fitting to the end of collection took approximately 5–6 minutes, after which the subject was asked to complete three 60-trial blocks of a localization task. On each trial the subject was presented with a stimulus and asked to indicate the perceived direction by orientating a 6-DOF tracked wand toward the perceived location and pressing a response button. The subject was made aware of the wand's orientation because it illuminated LED clusters mounted on each speaker as the subject pointed to that speaker. In this way, the subject was in fact making a discrete response confined to the locations of actual speaker locations within the ALF. The correct location was then presented to the subject by illuminating the LEDs on the target speaker location, which had to be acknowledged by the subject via a button press. Subjects were required to reorient toward the speaker directly in front of the sphere before they could initiate the start of the next trial by again pressing the button.

C. Results

Fig. 6 shows the average absolute angular localization errors averaged over all seven subjects for each test condition. This figure also shows the average results in terms of the error component that fell along the interaural axis (Lateral) and the component along a cone of confusion (Intraconic). For each error type, the asterisk indicates test conditions that were significantly different from the free-field conditions (FF). Statistical significance was determined using a paired t-test with Tukey-Kramer correction for multiple comparisons at a 95% confidence level. The general trend shows a slight increase in total angular error as the spherical harmonic order is decreased from around 12° for the free-field condition to approximately 17° for the 2nd-order SH condition, a span of only 5°. Similar trends are seen when the total error is broken down into its lateral and intraconic components, and for all conditions the largest component of the total error occurred in the intraconic dimension. Results for the 2nd-order SH condition was significantly different from free-field in terms of all three types of error, while the 4th-order SH condition showed significant difference only in the lateral dimension.

Fig. 7 shows the total angular error results broken out by subject. Excluding Subject 47, all subjects achieved their lowest average localization errors with the free-field sources (FF). Subjects 1438 and 1208, showed a large difference between the free-field conditions and even the best virtual conditions, possibly indicating a poor baseline HRTF. Despite this, all subjects showed a general increase in localization error as the spherical harmonic truncation order was reduced, and the greatest amount of error occurred with a 2nd order SH representation (SH2).

To further investigate how the SH model truncation affects response distributions the localization results were also analyzed in terms of response bias and response blur. Here the response bias is defined as the average absolute angular separation between a target location and the response centroid, while the response blur is defined as the average absolute angular distance between a response location and the response centroid. All centroids were calculated separately for each subject, stimulus condition, and location. The resulting angular errors were averaged over subject and location and presented in Fig. 8. Despite the spatial smoothing properties of SH model truncation, the relative contributions of response bias and response blur tend to remain fairly consistent as the SH model order is reduced. Moreover, a significant difference is seen for the 2nd order SH condition both in terms of response blur *and* response bias, indicating that reducing the model order did not just result in an increase in response variance.

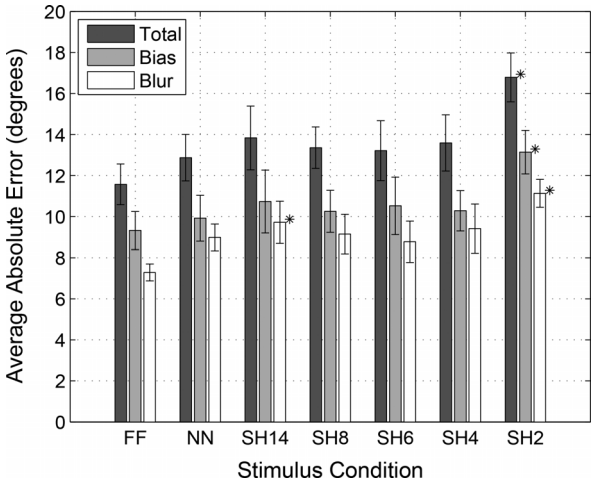

Fig. 8. Localization results averaged over all subjects and locations for each experimental test condition. Results are given in terms of total absolute angular error (Total), the average angle between the target and response centroid (Bias) and the average angle between a response and the response centroid (Blur). Error bars represent 95% confidence intervals, and asterisks represent results significantly different than free-field (FF).

VI. Discussion

The most surprising result of this study is the extent to which the listeners' localization responses were robust to the spectral errors introduced by low-order SH reconstruction of the HRTF. Even in the SH2 condition, where the spatial reconstruction was restricted to only a 2nd-order representation, listeners were able to localize with a mean absolute angular error of just 17°, which is only 5° worse than the 12° error obtained in the free-field condition of the experiment. It is worth noting that localization studies conducted in the same setup used in this experiment but with non-individualized HRTFs (measured on a KEMAR manikin) have produced overall localization errors on the order of 28° [22], which is *much* worse than performance in the SH2 condition tested in this experiment.

To put these results in perspective, it is helpful to compare them to those of other studies that have compared free-field and virtual localization under similar conditions. While strictly speaking, the discrete response grid of the current task makes it one of location identification rather than pure localization, the current loudspeaker array seems sufficiently dense to accurately capture aggregated response data consistent with true localization tasks reported in the literature. The overall free-field localization error of 12° appears quite reasonable, and it is definitely comparable to, or better than, the free-field errors reported in other studies examining the localization of a brief sound stimulus that could be distributed anywhere on the surface of the sphere [21], [23], [24].

The virtual localization errors in the nearest neighbor condition (NN, ≈13°) were also quite good, and in fact are among the best reported for virtual sound localization [25], [23], [14]. Of note here is that neither the HRTF collection or the localization task included head fixation, which means that the HRTFs used in the baseline NN condition were in fact interpolated to correspond to a head-relative loudspeaker position at the time of presentation. These results ensure that the drop in localization accuracy observed for low-order HRTFs are in fact due to the SH order reduction and not artifacts found in the HRTF measurements before SH analysis was applied.

Thus, it appears that the SH conditions of the experiment, and in particular those of the SH4-SH14 conditions, effectively capture almost all of the relevant individual-specific spatial information contained in the HRTFs. This implies that the perceptual system does not rely on the finer spatial structure contained in measured HRTFs for localization judgments. This result is in agreement with the HRTF interpolation work of [13], and [14], and further highlights the fact that the theoretical bounds placed on physical reconstruction of HRTFs like those of [2] and [15] far exceed the requirements of the auditory system.

An important note here is that these results apply to localization performance. Other perceptual factors such as perceived source width, source distance, and overall subjective quality might be affected by these low order representations. Preliminary results from subsequent experiments completed in our laboratory indicate that subjective quality ratings of these other factors may begin to decline more rapidly than localization accuracy when the SH order is reduced. It is currently unclear whether these early findings relate to the process of SH truncation and spatial smoothing, or whether they are related to the removal of echoic features in the original measured HRTFs. For applications where these subjective spatial attributes may be important, additional testing will be needed to determine the effects of this type of SH modeling.

Perhaps the most surprising result of the experiment is the high level of performance that was achieved with just the 2nd-order (SH2) representation of the HRTF. This seems to imply that a large amount of reliable localization information is present in the first and 2nd-order coefficients. In his discussion, [7] noted that the three first-order spherical harmonic coefficients might be particularly interesting from an analysis perspective. Due to their simple shapes and orientations, the first-order basis functions capture spatial variations in the underlying HRTF, which correspond to purely up-down, left-right, or front-back differences in the HRTF (these are also the three basis functions used in traditional ambisonic techniques [26]). Further analysis of these first-order coefficients shows that a number of existing phenomena seen in spatial auditory perception literature could have physical roots described by these first-order coefficients.

Fig. 9 shows the three spectra that result from taking the RMS energy for each of the three first order SH coefficients across the seven subjects in the current study at each frequency for the left ear. Not surprisingly, the largest amount of energy is contained in the coefficient that captures the left-right variation. This coefficient also clearly shows a frequency dependence, which is likely attributed to the left-right level difference caused by the head shadow. Both the up-down and the front-back coefficients show isolated spectral regions in which they contain a peak energy level. The center frequencies of these peaks align quite well with the perceptually derived "directional bands" described by [27] for the "forward" region around 4 kHz as well as the "overhead" region around 8 kHz. The second smaller peak in the up-down coefficient energy above 10 kHz could also be linked to the perceptual observation by [28] that the rate of front-back

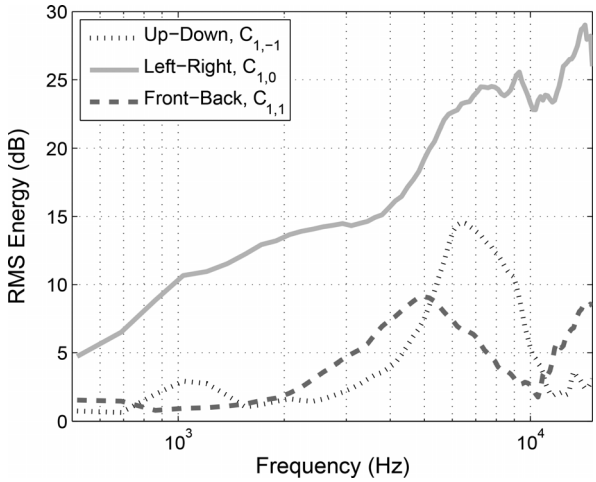

Fig. 9. The energy spectra (dB) for the first three SH coefficients averaged over all subjects.

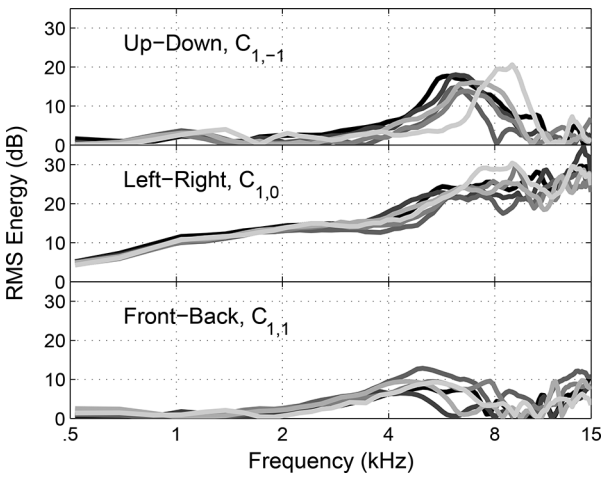

Fig. 10. The energy spectra (dB) for the first three SH coefficients plotted separately for each subject.

confusions increased dramatically when this region was eliminated by low-pass filtering the stimuli.

Some amount of individuality is also retained in the first-order coefficients. Fig. 10 shows the energy of the three first-order coefficients for the experimental subjects. While overall spectral shapes seem similar between subjects, differences can be seen, especially for the up-down coefficient plotted in the top panel. Interestingly, by color-coding the energy curves according to the subject's height (dark is taller, light is shorter) there seems to be a general trend for frequency of the maximum up-down energy to increase with decreasing subject height. This would agree with the findings of [29] who showed vertical localization errors could be reduced by frequency scaling a subject's HRTF based on a parameter related to subject height. This relationship is highlighted more formally in Fig. 11, where for each subject, the frequency of maximum energy for the up-down coefficient was plotted against the subject's height.

While these somewhat casual observations based on first-order coefficients may help explain why localization accuracy remains reasonable for the 2nd-order representation, they cannot explain why the departure from free-field accuracy occurs between 42th and 2nd order. Additional insight

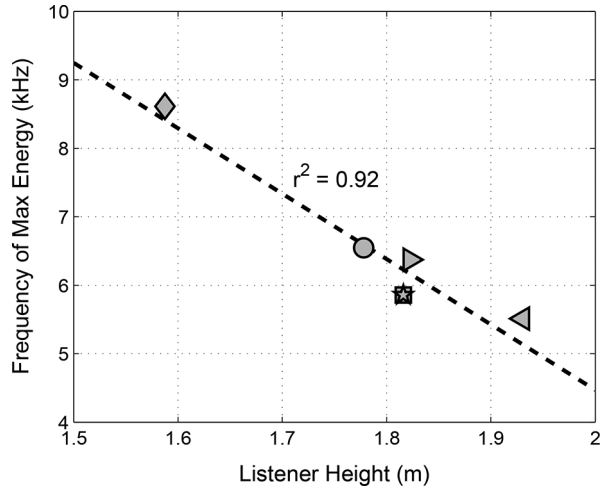

Fig. 11. For each subject (symbols) the frequency of maximum energy for the up-down coefficient $(C_{1,-1})$ plotted against the subject's height along with a best-fit linear regression line. The indicated r^2 represents the coefficient of determination for the best-fit line.

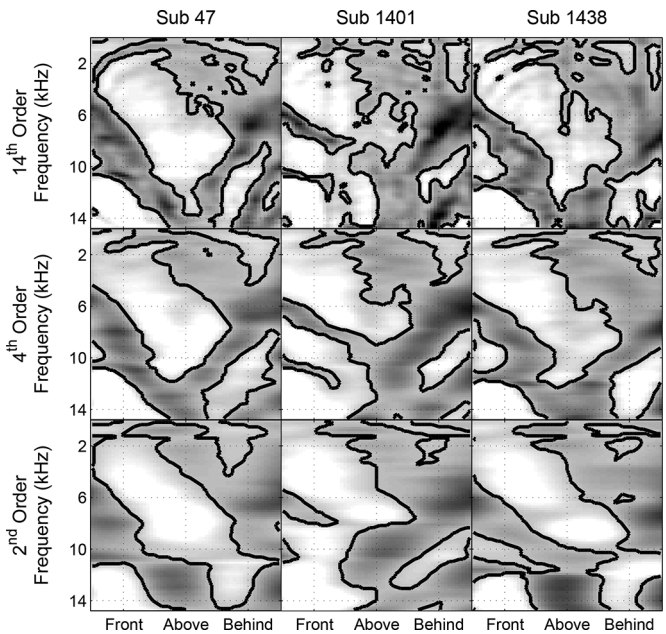

Fig. 12. Magnitude spectra (in dB) of modeled HRTFs on the median plane for three subjects (columns) and three different truncation orders (rows) normalized by the average HRTF spectrum on the median plane for each subject and order. Black lines indicate the 0-dB contour. Dark shading indicates regions of lower energy (less than 0 dB) while light regions indicate regions of high energy (greater than 0 dB).

into the physical features responsible for these results may be gleaned from further analysis of the resulting modeled HRTFs. Fig. 12 shows reconstructed magnitude spectra for the left ears of three representative subjects (columns) and three different truncation orders (rows). In each panel the median-plane HRTF magnitude (in dB) is shown as a function of frequency (ordinate) and angle around the median plane (abscissa). The median-plane HRTF shown here is the HRTF magnitude response in decibels minus the average magnitude response taken over all median plane locations. Shading is used to represent the level of each angle-frequency bin, and the contour line separates regions with positive (light) and negative (dark) values.

Considering first the 14th-order representation, each subject clearly shows the intricate and individualized set of spectral patterns, which are widely accepted to be the physical features utilized for accurate localization within the median plane. As the spherical harmonic order is further reduced, the spatial smoothing that results from model truncation becomes evident as smaller, more localized features begin to disappear in the 4th-order representations, and are completely replaced by large global regions in the 2nd-order models. Combining these observations with the perceptual results indicates that the removal of small isolated spectro-spatial features apparent in going from 14th to 4th order have little impact on localization accuracy, while disruption of larger features (like the removal of the rear 8-kHz elevation notch for Subject 47) begins to cause more dramatic losses of localization accuracy. While definite conclusions cannot be drawn from these observations alone, they seem to support previous hypotheses linking the presence of "macroscopic" features in the median plane HRTF to localization accuracy [30].

In terms of modeling efficiency, the fact that only a 4th-order model needs to be used means that an HRTF can be represented with $25 * (K + 3)$ real-valued parameters which, assuming $K = 256$ taps in each FIR HRTF filter and $S = 300$ spatial measurements, means that a perceptually valid HRTF can be stored with just over 4% of the parameters required to represent the original, measured HRTF. This puts the truncated SH model on par with the traditional spectral-based PCA approaches such as [8], which require approximately 2.5% of the number of coefficients in the measured form. Perhaps the clearest comparison can be made to the spatial PCA approach introduced by [31], who found accurate reconstruction performance could be achieved with 35 spatial basis functions which had been derived optimally from a database of HRTFs. While the SH-based method presented here required fewer spatial basis functions (25 versus 35) it is not known whether adequate performance could be achieved with a reduced basis since no localization tests were included.

The localization results for the 4th-order model would also imply that an HRTF that preserves localization accuracy could be estimated with as few as 26 spatial measurement locations since this is all that is required to ensure a unique solution to (4) with $P = 4$. Additionally, as shown in Fig. 5, modeling error with low-order models was much smaller at low frequencies, implying that increased modeling efficiency may be attainable with a frequency-dependent truncation scheme. An even larger gain could potentially be accomplished if the technique above were integrated with a scheme to approximate the spectral resolution of the auditory system.

VII. Conclusion

The current study introduced a new model for representing head-related transfer functions that permits the systematic removal of spatial detail. It was found that a large amount of spatial detail could be eliminated without affecting localization accuracy. This finding implies that theoretical bounds for minimal HRTF representations derived from purely acoustic considerations may be grossly overestimating the requirements from a perceptual standpoint.

The current SH-based HRTF model was also shown to produce model HRTFs with a spectral distortion of less than 1 dB across the frequency range from 300 Hz to 14 kHz when a 14th-order representation was used. Free-field comparable localization performance could be retained with model orders as low as 4, giving the technique similar efficiency to previous statistically-based approaches for HRTF representation. This finding has implications for the way HRTFs are modeled, stored, and used to render virtual auditory displays, as well as how soundscapes in general are captured for eventual playback.

References

[1] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF fitting using spherical harmonics," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'09)*, 2009, pp. 257–260.

[2] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights into head-related transfer function: Spatial dimensionality and continuous representation," *J. Acoust. Soc. Amer.*, vol. 127, pp. 2347–2357, 2010.

[3] P. Guillon and R. Nicol, "Head-related transfer function reconstruction from sparse measurements considering a priori knowledge from database analysis: A pattern recognition approach," in *Proc. 125th Conv. Audio Eng. Soc.*, 2008.

[4] W. Zhang, R. A. Kennedy, and T. D. Abhayapala, "Efficient continuous HRTF model using data independent basis functions: Experimentally guided approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 819–829, May 2009.

[5] M. Aussal, F. Alouges, and B. Katz, "A study of spherical harmonics interpolation for HRTF exchange," in *Proc. Meetings Acoust.*, 2013.

[6] R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis, "High order spatial audio capture and its binaural head-tracked playback over headphones with hrtf cues," in *Proc. AES 119th Conv.*, New York, NY, USA, 2005.

[7] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Anlaysing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 104, pp. 2400–2411, 1998.

[8] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Amer.*, vol. 91, pp. 1637–1647, 1992.

[9] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *J. Acoust. Soc. Amer.*, vol. 105, pp. 2821–2840, 1992.

[10] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, pp. 747–749, 1998.

[11] J. Breebaart, F. Nater, and A. Kohlrausch, "Spectral and spatial parameter resolution requirements for parametric, filter-bank-based HRTF processing," *J. Aud. Eng. Soc.*, vol. 58, no. 3, pp. 126–140, 2010.

[12] G. Enzner, M. Krawczyk, F.-M. Hoffmann, and M. Weinert, "3d reconstruction of HRTF-fields from 1d continuous measurements," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'11)*, 2011, pp. 157–160.

[13] R. Martin and K. McAnally, "Interpolation of head-related transfer functions," Air Operations Division Defence Science and Technology Org., Tech. Rep., 2007.

[14] S. Carlile, C. Jin, and V. van Raad, "Continuous virtual auditory space using HRTF interpolation: Acoustic & psychophysical errors," in *Proc. Int. Symp. Multimedia Inf. Process.*, 2000.

[15] T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli, "Sound field analysis along a circle and its applications to HRTFs interpolation," Audiovisual Communications Lab., EPFL, Lausanne, Switzerland, Tech. Rep., 2008.

[16] J. Chen, B. D. V. Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *J. Acoust. Soc. Amer.*, vol. 97, pp. 439–452, 1995.

[17] Q. Huang and Y. Fang, "Interpolation of head-related transfer functions using spherical Fourier expansion," *J. Electron. (China)*, vol. 26, pp. 571–576, 2009.

[18] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerovd, "Fast head-related transfer function measurement via reciprocity," *J. Acoust. Soc. Amer.*, vol. 120, pp. 2202–2215, 2006.

[19] M. Poletti, "Unified description of ambisonics using real and complex spherical harmonics," in *Proc. Ambisonics Symp.*, 2009.

[20] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. San Diego, CA, USA: Academic, 1999.
[21] D. S. Brungart, G. D. Romigh, and B. D. Simpson, "Rapid collection of HRTFs and comparison to free-field listening," in *Proc. Int. Workshop Principles Applicat. Spatial Hearing*, 2009.
[22] D. S. Brungart and G. D. Romigh, "Spectral HRTF enhancement for improved vertical-polar auditory localization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2009.
[23] R. Martin, K. McAnally, and M. Senova, "Free-field equivalent localization of virtual audio," *J. Aud. Eng. Soc.*, vol. 49, pp. 14–22, 2001.
[24] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *J. Acoust. Soc. Amer.*, vol. 92, pp. 2607–2624, 1992.
[25] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Amer.*, vol. 85, pp. 868–878, 1989.
[26] M. A. Gerzon, "Practical periphony: The reproduction of full-sphere sound," in *Proc. Audio Eng. Soc.; 65th Conv.*, London, U.K., 1980.
[27] J. Blauert, *Spatial Hearing*. Cambridge, MA, USA: MIT Press, 1997.
[28] R. B. King and S. R. Oldfield, "The impact of spatial bandwidth on auditory localization: Implications for the design of three-dimensional audio displays," *Human Factors*, vol. 39, pp. 287–295, 1997.
[29] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Amer.*, vol. 106, pp. 1493–1510, 1999.
[30] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoust. Soc. Amer.*, vol. 88, pp. 159–168, 1990.
[31] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *J. Acoust. Soc. Amer.*, vol. 132, pp. 282–294, 2012.

Griffin D. Romigh received the B.S. degree in biomedical engineering from Wright State University in 2009, an M.S. in 2011, and Ph.D. in 2012, both from Carnegie Mellon University in electrical and computer engineering. He is currently an Electrical Engineer and Program Manager for the Battlespace Acoustics Branch at the Air Force Research Laboratory, Wright-Patterson Air Force Base, OH.

His research focuses on the application of signal processing and machine learning techniques to better understand human spatial auditory and speech perception, particularly how to efficiently model and estimate individualized head-related transfer functions. He also conducts applied research on the application of spatial audio and language technology to improve communication and situation awareness in complex military environments. Dr. Romigh is a member of IEEE, Acoustical Society of America, and former U.S. Department of Defense SMART Scholar.

Douglas S. Brungart received the B.A. degree from Wright State University in 1993, an S.M. in 1994, and Ph.D. in 1998, both from Massachusetts Institute of Technology in electrical engineering all in computer science and electrical engineering. He is currently the Chief Scientist at the Army Audiology and Speech Center at Walter Reed NMMC and Director of Research for the U.S. Dept. or Defense Hearing Center of Excellence.

His research addresses aspects of basic and applied research in the areas of spatial hearing, hearing impairment, speech perception, and hearing protection. He is most well-known for his work on informational masking in multi-talker speech displays and his work on near-field spatial hearing. Dr. Brungart is a member of IEEE and a Fellow of the Acoustical Society of America.

Richard M. Stern received the S.B. degree from the Massachusetts Institute of Technology in 1970, the M.S. from the University of California, Berkeley, in 1972, and the Ph.D. from MIT in 1977, all in electrical engineering. He has been on the faculty of Carnegie Mellon University since 1977, where he is currently a Professor in the Electrical and Computer Engineering Department, the Computer Science Department, and the Language Technologies Institute. He is also a Lecturer in the School of Music.

Much of his current research is in spoken language systems, where he is particularly concerned with the development of techniques with which automatic speech recognition can be made more robust with respect to changes in environment and acoustical ambience. He has also developed sentence parsing and speaker adaptation algorithms for earlier CMU speech systems. In addition to his work in speech recognition, he also maintains an active research program in psychoacoustics, where he is best known for theoretical work in binaural perception. Dr. Stern is a Fellow of the IEEE, the Acoustical Society of America, and the International Speech Communication Association (ISCA), and he was a recipient of the Allen Newell Award for Research Excellence in 1992. He served as the General Chair of Interspeech 2006 and as the 2008-2009 ISCA Distinguished Lecturer. He is also a member of the Audio Engineering Society.

Brian D. Simpson received his A.B. in psychology for Washington University in 1995, and his M.S. and Ph.D. in psychology with an emphasis on Human Factors from Wright State University in 2002 and 2011, respectively. He is currently a Research Psychologist and the Technical Advisor for the Battlespace Acoustics Branch at the Air Force Research Laboratory, Wright-Patterson Air Force Base, OH.

Dr. Simpson's research has focused on the investigation of peripheral and central processes that mediate speech perception and spatial hearing in multisource acoustic environments, spatial auditory attention, and the development of auditory displays to support performance in complex task environments. He is a member of the Acoustical Society of America and the Human Factors and Ergonomics Society.