A MULTILINEAR APPROACH TO HRTF PERSONALIZATION

Graham Grindlay and M. Alex O. Vasilescu

Massachusetts Institute of Technology
Media Laboratory
20 Ames St., Cambridge, MA 02139

ABSTRACT

The head-related transfer function (HRTF) captures most of the auditory cues used to discern sound source direction and therefore is essential for synthesizing convincing spatial audio. HRTFs are, however, difficult to measure and highly person-specific. This has created significant interest in alternative methods for generating custom HRTFs. We present a multilinear modeling framework for HRTFs which makes use of a tensor decomposition called the N-mode SVD [1], a multilinear extension of the conventional singular value decomposition (SVD). Regression is then used to map simple anatomical data to complex HRTF data. This mapping defines a data-driven model capable of producing entire sets of individualized HRTFs from easily obtained anatomical measurements. We show that our approach yields objectively superior results to those of a mapping based on principle components analysis (PCA).

***Index Terms*—** Audio systems, HRTF, Tensors, Multilinear algebra

1. INTRODUCTION

The head related transfer function (HRTF) and its corresponding impulse response, the head related impulse response (HRIR), are essential components of many approaches to binaurally-based spatial audio synthesis. The HRTF describes, for a particular ear and sound source direction, the frequency response that results from a complex interaction of anatomically-based reflection and diffraction effects. Along with interaural time differences and interaural level differences, these filtering effects are believed to be the primary cues used to discern sound direction [2]. Thus by filtering a sound sample with the appropriate left and right ear HRTFs, convincing virtual auditory environments can be constructed.

The HRTF is a function of two variables, sound source direction and the external morphology (head, torso, and outer ear) unique to each person. In general, a set of HRTFs recorded from one person will yield poor results when used to spatialize audio for someone else. Use of a mismatched set of HRTFs often results in front-back confusion as well as difficulty perceiving source elevation [3]. Unfortunately, it is difficult and time-consuming to measure HRTFs, making empirically-based customization unrealistic. This has given rise to significant interest in model-based approaches. Several recent papers describe approaches to HRTF customization using anthropometric data [4, 5]. Systems based on anatomical measurements are a particularly attractive approach as the data is easy to gather and computer vision systems promise to make this process even easier [6]. Other prior modeling work [7, 8] has used PCA to analyze HRTF data. However, PCA works best with functions of one variable as it can only capture the total variation present in a dataset. This inability to distinguish between variations due to sound source direction and variations due to morphology can result in PCA disposing of morphological information that is important for spatial localization.

In this paper, we introduce a nonlinear, multifactor model of HRTFs that generalizes conventional PCA. Whereas PCA employs linear (matrix) algebra, our approach exploits multilinear (tensor) algebra. Multilinear algebra, the algebra of higher order tensors, is able to learn the interactions of the multiple factors inherent to HRTFs and separately encode each of the modes. We exploit the dataset's natural factoring by using the N-mode SVD, a tensor decomposition which allows us to explicitly represent each of the factors in the dataset (frequency, people, and direction) as orthogonal vector spaces. Each of these spaces provides a representation, statistically independent of the others, that describes how the HRTF data is influenced by that particular degree of freedom.

Our approach to mapping anthropometric data to HRTF data is similar in spirit to that of Inoue et al. [9] who use a multiple regression model to map anthropometric features to the coefficients representing each HRTF after applying PCA. We use a regression model to map anthropometric features to the people vector space produced by the N-mode SVD. Then, given anatomical measurements of a person not in the database, we use the regression model to solve for that person's representation in people space. This representation is then combined with the results of the N-mode SVD to generate a custom set of HRTFs.

2. TENSOR ALGEBRA

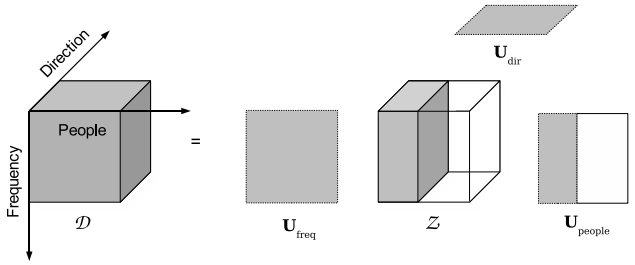


Fig. 1. N-mode SVD decomposition illustrated for a 3-mode tensor, $\mathcal{D}$. The result is core tensor $\mathcal{Z}$ and mode matrices $\mathbf{U}_{freq}$, $\mathbf{U}_{people}$, and $\mathbf{U}_{dir}$. Dimensionality reduction, shown for mode $\mathbf{U}_{people}$, is accomplished by truncating the rows of $\mathbf{U}_{people}$ as well as the corresponding portions of $\mathcal{Z}$.

Tensors generalize scalars (0^{th}-order), vectors (1^{st}-order), and matrices (2^{nd}-order) to higher-order arrays. [1] A tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is said to be of *order* N. The *mode-n product* of a tensor, $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n \times \ldots, I_N}$, by a matrix, $\mathbf{M} \in \mathbb{R}^{J_n \times I_n}$, is denoted as $\mathcal{X} \times_n \mathbf{M}$. This product yields a tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots, I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N}$ whose entries are given by $y_{i_1,\ldots,i_{n-1},j_n,i_{n+1},\ldots,i_N} = \sum_{i_n} x_{i_1,\ldots,i_{n-1},i_n,i_{n+1},\ldots,i_N} m_{j_n i_n}$.

The popular singular value decomposition (SVD), which forms the basis of principle components analysis (PCA), factors a matrix $\mathbf{D} = \mathbf{U}_1 \mathbf{S} \mathbf{U}_2^T$, where $\mathbf{U}_1$ is a set of orthonormal basis vectors spanning column space, $\mathbf{S}$ is the diagonal singular value matrix, and $\mathbf{U}_2$ is a set of orthonormal basis vectors spanning row space. Using the mode-n product, this decomposition can be written as $\mathbf{D} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2$. This naturally leads to the more general N-mode SVD which decomposes a tensor of order N into its N constituent vector spaces:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_n \mathbf{U}_n \ldots \times_N \mathbf{U}_N \qquad (1)$$

Here $\mathcal{Z}$ plays a role analogous to $\mathbf{S}$ in the SVD (although note that its non-diagonal entries are not necessarily zero) and modulates the interaction between modes. The columns of each mode matrix, $\mathbf{U}_i$, span the space of that factor while the rows encode the particular instances of that factor that are present in $\mathcal{D}$. For a more detailed discussion of the N-mode SVD, see Vasilescu and Terzopoulos [1].

Perhaps one of the most useful aspects of the N-mode SVD is the ability to perform *targeted* dimensionality reduction. In contrast to traditional PCA where dimensionality reduction affects all aspects of the data, we can truncate the basis vectors of each mode matrix separately, affording a much finer degree of control. While simple truncation does provide reasonable results, significantly better results can be had using an alternating least squares algorithm [10]. Figure 1 illustrates the N-mode SVD as well as dimensionality reduction of one mode.

[1] We use uppercase calligraphic letters to denote tensors, uppercase bold letters to denote matrices, and lowercase bold letters to denote vectors.

3. DATA

We use the publically available CIPIC database [11] which contains head related impulse responses (HRIRs) recorded for both ears of 45 subjects over 1250 directions (25 azimuths $\theta \in [-80..80]$ and 50 elevations $\phi \in [-45..230.625]$ specified in interaural polar coordinates) spaced roughly uniformly over the head sphere. Each HRIR is 200 samples long (approximately 4.5 ms) and was recorded at a 44.1 kHz sampling rate in 16-bit resolution. Each HRIR was transformed into an HRTF by a 512-point FFT and then filtered to contain only frequencies between 500 Hz and 16 kHz, leaving 181 frequencies in each HRTF. We then combined left and right ear data for each direction and person by concatenating each pair of HRTFs into a single vector of length 362.

The database also contains anthropometric data for each of the subjects, although eight subjects are missing one or more morphological measurements and therefore were not used in our experiments. For the remaining 37 subjects, we used the following morphological features: head width, cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, intertragal incisure width, cavum concha depth, pinna rotation angle, and pinna flare angle. The measurements for both ears of each person were concatenated together and PCA was performed on the resulting vectors. The first 10 principle components were retained as they explained over 90% of the data's variance. A column of 37 ones was then prepended to the matrix of anthropometric coefficients to provide a constant term for the regression model. We refer to the resulting dataset as $\mathbf{A}$.

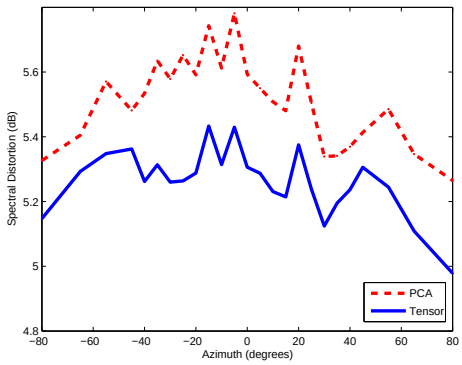
4. HRTF CUSTOMIZATION

Using the HRTF data described above, we define a data tensor, $\mathcal{D} \in \mathbb{R}^{F \times P \times D}$ where F is the number of frequencies (362), P is the number of people (37), and D is the number of directions (1250). Using the N-mode SVD, we decompose the tensor as follows (see Figure 1):

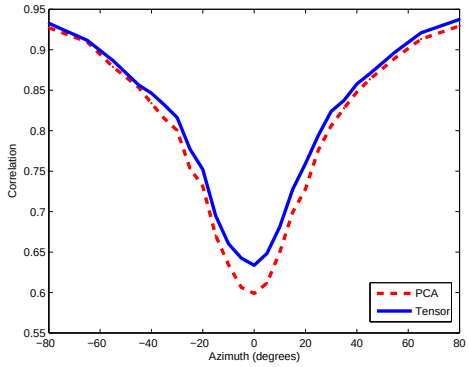
$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_F \times_2 \mathbf{U}_P \times_3 \mathbf{U}_D \qquad (2)$$

In our experiments, we then used the ALS algorithm mentioned in Section 2 to reduce the dimensionality of the people mode, $\mathbf{U}_P$, from 37 to 5 (chosen ad-hoc), giving $\hat{\mathbf{U}}_P$. We now wish to find a mapping $\mathbf{A} \mapsto \hat{\mathbf{U}}_P$. This can be done with the following regression model,

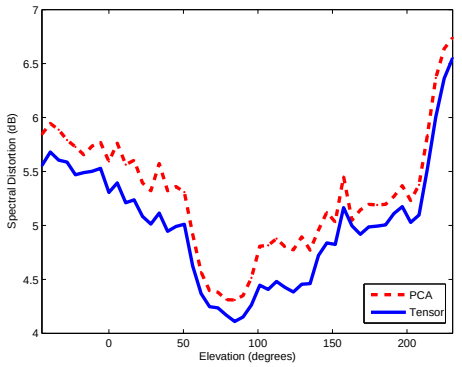
$$\hat{\mathbf{U}}_P^T = \mathbf{B}\mathbf{A}^T \qquad (3)$$



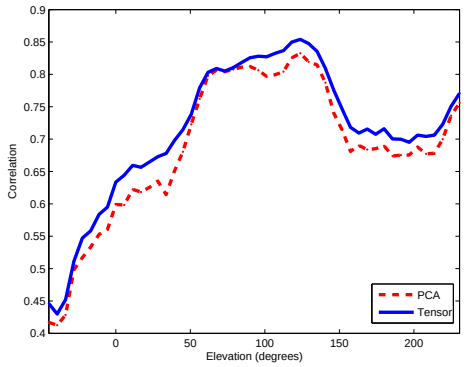
(a) Spectral distortion values for $\phi = 0$, $\theta = [-80..80]$



(b) Correlation values for $\phi = 0$, $\theta = [-80..80]$



(c) Spectral distortion values for $\theta = 0$, $\phi = [-45..230.625]$



(d) Correlation values for $\theta = 0$, $\phi = [-45..230.625]$

Fig. 2. Comparison of representative tensor and PCA modeling results for one fixed elevation and one fixed azimuth. Both spectral distortion scores as well as correlation values between synthesized and real HRTFs are shown.

where $\mathbf{B}$ is a matrix of coefficients. Using $\mathbf{A}^+$ to denote the pseudoinverse of $\mathbf{A}$, we can easily solve for $\mathbf{B}$ as:

$$\mathbf{B}^T = \mathbf{A}^+ \hat{\mathbf{U}}_P \tag{4}$$

Given a vector of anthropometric measurements, $\mathbf{a}_{new}$ for a person not in the database, we solve for the vector of coefficients, $\hat{\mathbf{u}}_{p_{new}}$, that correspond to the new person in people space.

$$\hat{\mathbf{u}}_{p_{new}}^T = \left[1\ \mathbf{a}_{new}^T\right] \mathbf{B}^T \tag{5}$$

We can now solve for the complete set of HRTFs, $\mathcal{D}_{new}$, using $\hat{\mathbf{u}}_{p_{new}}$:

$$\mathcal{D}_{new} = \mathcal{Z} \times_1 \mathbf{U}_F \times_2 \hat{\mathbf{u}}_{p_{new}}^T \times_3 \mathbf{U}_D \tag{6}$$

5. EXPERIMENTS

To test the effectiveness of our approach, we conducted a suite of experiments comparing our multilinear approach to PCA. For the PCA approach, we considered the HRTFs across subjects for each direction separately. For each direction, d, we performed PCA on the corresponding HRTFs of all subjects, reducing their dimensionalities down to 5 (same reduction as $\mathbf{U}_P$). This yielded a matrix, $\hat{\mathbf{C}}_d$, containing the coefficient vectors for direction, d, for each subject. As with the tensor framework, regression was then applied to map the anthropometric data to $\hat{\mathbf{C}}_d$. Note, however, that the PCA approach requires a separate regression model for each direction (1250 in total), while the tensor model requires only one.

Performance was evaluated using a cross-validation approach. For each of the 37 subjects we constructed both a tensor as well as a PCA-based mapping using the other 36

subjects' data. These models were then used to synthesize the held-out subject's HRTFs from their anthropometric data. We then computed the reconstruction error and correlation between each of the held-out subjects' synthesized and real HRTFs. These correlation and error values were then averaged across subjects. The spectral distortion was used as an error metric between the real and synthesized HRTF data:

$$SD\left(H,\hat{H}\right)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(20\log\frac{|H_i|}{|\hat{H}_i|}\right)^2} \qquad (7)$$

where $|H_i|$ is the magnitude of the i^{th} frequency of the true HRTF, $|\hat{H}_i|$ is the magnitude of the i^{th} frequency of the synthesized approximation, and N is the number of frequencies.

Figure 2 shows a pair of representative experimental results. Spectral distortion and correlation values are shown for all azimuths with a fixed elevation (Figure 2a and Figure 2b) and for all elevations with a fixed azimuth (Figure 2c and Figure 2d). In both cases 0 was arbitrarily chosen for the fixed value.

As can be seen in Figure 2, the tensor approach resulted in lower spectral distortion and higher correlation values in a variety of experimental conditions. In fact, in all 1250 directional cases, the spectral distortion was lower and the correlation was higher for the tensor model than for the PCA model.

6. CONCLUSIONS

We have presented a novel tensor framework for modeling person-specific HRTFs. Our generative model is fully data-driven and capable of producing individualized sets of HRTFs from easily obtained anatomical measurements. We have demonstrated that our approach compares favorably to a PCA-based framework, achieving consistently lower spectral distortion scores and consistently higher correlation values across all directional conditions. Furthermore, our system maps anthropometrics to HRTFs using a single regression model while the PCA approach requires a separate regression model for each spatial direction.

We are currently planning listening experiments to provide a subjective measure of our system's abilities. In addition, we are investigating how the tensor framework might be used in other areas of HRTF research, such as spatial interpolation and source position estimation [12].

7. REFERENCES

[1] M. A. O. Vasilescu and D. Terzopoulos, "TensorTextures: Multilinear image-based rendering," in *Proceedings of ACM SIGGRAPH 2004 Conference*, 2004, pp. 334–340.

[2] J. Middlebrooks and D. Green, "Sound localization by human listeners," *Annual Review of Psychology*, vol. 42, pp. 135–159, January 1991.

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

[4] D. Zotkin, J. Hwang, R. Duraiswaini, and L. Davis, "HRTF personalization using anthropometric measurements," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2003, pp. 157–160.

[5] D. Zotkin, R. Duraiswami, and L. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 553–564, 2004.

[6] R. Duraiswami et al., "Creating virtual spatial audio via scientific computing and computer vision," in *Proceedings of the 140th Meeting of the ASA*, Newport Beach, CA, December 2000, p. 2597.

[7] D. Kistler and F. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, March 1992.

[8] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, "Enabling individualized virtual auditory space using morphological measurements," in *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, December 2000.

[9] N. Inoue, T. Nishino, K. Itou, and K. Takeda, "HRTF modeling using physical features," in *Forum Acusticum 2005*, Budapest, Hungary, 2005, pp. 199–202.

[10] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-$(r_1, r_2, ..., r_n)$ approximation of higher-order tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.

[11] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2001, pp. 99–102.

[12] K. Martin, "Estimating azimuth and elevation from interaural differences," in *IEEE Mohonk workshop on Applications of Signal Processing to Acoustics and Audio*, October 1995.