A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features

Felipe Grijalva, *Student Member, IEEE*, Luiz Martini, Dinei Florencio, *Senior Member, IEEE*,
and Siome Goldenstein, *Senior Member, IEEE*

***Abstract*—We present a new anthropometry-based method to personalize head-related transfer functions (HRTFs) using manifold learning in both azimuth and elevation angles with a single nonlinear regression model. The core element of our approach is a domain-specific nonlinear dimensionality reduction technique, denominated Isomap, over the intraconic component of HRTFs resulting from a spectral decomposition. HRTF intraconic components encode the most important cues for HRTF individualization, leaving out subject-independent cues. First, we modify the graph construction procedure of Isomap to integrate relevant prior knowledge of spatial audio into a single manifold for all subjects by exploiting the existing correlations among HRTFs across individuals, directions, and ears. Then, with the aim of preserving the multifactor nature of HRTFs (i.e. subject, direction and frequency), we train a single artificial neural network to predict low-dimensional HRTFs from anthropometric features. Finally, we reconstruct the HRTF from its estimated low-dimensional version using a neighborhood-based reconstruction approach. Our findings show that introducing prior knowledge in Isomap's manifold is a powerful way to capture the underlying factors of spatial hearing. Our experiments show, with p-values less than 0.05, that our approach outperforms using, either a PCA linear reduction, or the full HTRF, in its intermediate stages.**

***Index Terms*—HRTF personalization, manifold learning, spatial audio, virtual auditory displays.**

I. Introduction

THE *Head-Related Transfer Functions* (HRTFs) encode audio localization cues such as *Interaural Time Difference* (ITD), *Interaural Level Difference* (ILD) and spectral coloring, caused by sound scattering around the head, pinna and torso before it reaches the eardrum [1].

Since HRTFs differ widely among individuals, it is necessary to personalize them to ensure high-quality spatial audio. Nonindividualized HRTFs hinder localization accuracy, often causing front-back and up-down confusions [2].

The unsolved problem of HRTF customization is increasingly subject of much research due to the growing importance of auditory augmented reality applications [3], [57]. The most accurate approach to personalizing HRTFs is through direct measurements. However, this is a complex, time-consuming, expensive, and not scalable procedure [4].

In light of this, several alternative methods aimed at avoiding measuring HRTFs have been proposed, including the estimation of HRTFs from a small set of measurements [5]. Furthermore, there are several theoretical models (e.g. spherical head model [6], snowman model [7], structural models [8], [9]) which attempt to approximate the complicated human anatomy. Additionally, several numerical methods (e.g. boundary element method [10], [11], finite-difference time-domain method [12]) have been proposed. However, they require expensive acquisition hardware and are computationally intensive. Pursuing a different direction, several authors have proposed perceptual-based methods, where subjects choose their HRTFs through listening tests by tuning some parameters until they achieve an acceptable spatial accuracy [13], [14]. Moreover, Sunder et al. proposed an individualization method in the horizontal plane that uses a frontal projection headphone to introduce idiosyncratic pinna cues [15].

Alongside the aforementioned methods, HRTFs can also be customized from anthropometric measurements. Anthropometry-based regression methods predict individualized HRTFs using a model derived from a baseline database. It is precisely this kind of individualization methods that this work focuses on.

This paper introduces a new customization method to personalize HRTFs using Isomap, a nonlinear dimensionality reduction technique. Here, we extend for all directions the ideas of our preliminary study in the horizontal plane [16]. Our main contribution is our graph construction procedure for learning a single Isomap manifold for all subjects that incorporates important prior knowledge of spatial audio to exploit the correlation existing among HRTFs across individuals, directions and ears. Besides, instead of personalizing the HRTFs directly, we customize the *intraconic* component of HRTFs resulting from a spectral decomposition [17]. The intraconic component of HRTFs aims at providing the most important cues for individualization, leaving out subject-independent cues. Finally, our approach constructs a single regression model using an artificial

Manuscript received August 10, 2015; revised November 30, 2015; accepted December 28, 2015. Date of publication January 12, 2016; date of current version February 23, 2016. This work was supported in part by Microsoft, São Paulo Research Foundation (FAPESP) under Grants 2012/50468-6, 2013/21349-1, and 2014/14630-9, in part by the National Counsel of Technological and Scientific Development (CNPq) under Grants 308882/2013-0 and 454082/2014-2, and in part by the Coordination for the Improvement of Higher Education Personnel (CAPES), and it is part of a project that was approved by Unicamp's IRB CAAE 15641313.7.0000.5404. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Augusto Sarti.

F. Grijalva and L. Martini are with the School of Electrical and Computer Engineering, University of Campinas, Campinas, SP 13083-852, Brazil (e-mail: felipe.l.grijalva@ieee.org; felipe84@decom.fee.unicamp.br; martini@decom.fee.unicamp.br).

S. Goldenstein is with the Institute of Computing, University of Campinas, Campinas, SP 13083-970, Brazil (e-mail: siome@ic.unicamp.br).

D. Florencio is with Microsoft Research, Redmond, WA 98052 USA (e-mail: dinei@microsoft.com).

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/TASLP.2016.2517565

neural network that does not break the inherent multifactor nature of HRTFs (i.e. frequency, direction and subject factors).

II. Related Work

Anthropometry matching is the most straightforward way to personalize HRTFs from anthropometric data. In this context, various approaches in the literature [18]–[20] customize HRTFs by finding the best match in a baseline database of anthropometric features.

Middlebrooks [21] introduced an anthropometry-based method that uses frequency scaling of HRTFs based on the assumption that inter-subjects difference in anatomy features produce a frequency shift in individualized HRTFs.

On the other hand, anthropometric regression methods predict the individualized HRTFs of a new subject using a model derived from a baseline database. Linear dimensionality reduction techniques such as *Principal Component Analysis* (PCA) [22] and *Independent Component Analysis* (ICA) [23] have been widely used prior to customization. There are several HRTF customization methods to map anthropometric features to low-dimensional HRTFs previously calculated with PCA [24]–[27].

Due to the inability of linear regression methods to predict the complex relationship between anthropometric features and low-dimensional HRTFs, various authors introduced nonlinear regression techniques such as *Artificial Neural Networks* (ANN) [28] and *Support Vector Regression* (SVR) [23], [29] in conjunction with PCA [28], [29] or ICA [23]. Moreover, because SVR is only capable of training a multiple-to-one regression model (i.e. SVR needs to train a separate model for each dimension of low-dimensional HRTFs), Wang et al. [30] proposed a *joint SVR* to exploit the correlation between components of low-dimensional HRTFs.

The anthropometry-based regression methods described so far construct a model for each direction, which in turn means that the inherent multi-factor nature of HRTFs (i.e. frequency, direction and subject) is broken [31]. To overcome this problem, Grindlay et al. [32] introduced a three-mode (i.e. frequency, direction and subject mode) multilinear tensor representation for HRTFs. A single linear regression model is used for mapping anthropometric features to a five -dimensional vector obtained by means of *N-mode Singular Value Decomposition* (N-mode SVD) and which represents the subject mode in the tensor. A similar tensor-based approach is used in [31] and [33], but to construct the regression model, they employed an ANN and high-order partial least squares, respectively. More recently, Bilinski et al. [34] used a HRTF tensor representation to learn a sparse vector of a subject's anthropometric features as a linear superposition of the anthropometric features of a training subset. They applied the same sparse vector to synthesize the HRTF of a subject.

In addition to linear representations, nonlinear dimensionality reduction techniques have been also applied to HRTFs. Duraiswami et al. [35] applied *Locally Linear Embedding* (LLE) [36] to learn the nonlinear manifold structure in median plane HRTFs of the same subject. They also proposed a new HRTF interpolation method that estimates an HRTF as a linear combination of its neighbors on the low-dimensional manifold.

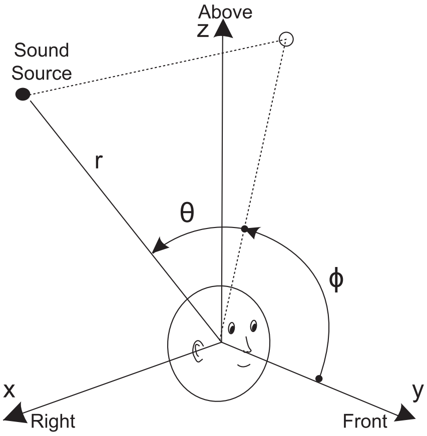


Fig. 1. The interaural coordinate system as described in [43], where azimuth is defined in the range $-90° \leq \theta \leq 90°$ and elevation in $-90° < \phi \leq 270°$.

Furthermore, Kapralos et al. compared PCA, Isomap [37] and LLE through correlation analysis [38] and subjective experiments [39], concluding that Isomap and LLE outperform PCA in finding the underlying factors of spatial hearing.

Based on the results of Duraiswami et al. [35] and Kapralos et al. [38], [39] using LLE and Isomap for HRTF interpolation and dimensionality reduction, in our previous work [16], we proposed a novel technique for customizing horizontal plane HRTFs using Isomap.

All aforementioned manifold learning studies support the idea suggested by Seung et al. [40] that nonlinear manifold techniques are crucial for understanding how perception arises from the dynamics of neural networks in the brain.

In this paper, we extend the ideas of our previous customization method [16] for locations beyond the horizontal plane. In this line, we apply Isomap over HRTFs to construct a manifold structure and then we employ an artificial neural network to predict the HRTFs for a new subject based on his anthropometric parameters.

As in previous works [16], [25], [28], [32], we work with the minimum phase assumptions of HRTFs [41], i.e., a minimum-phase function cascaded with a pure delay. In practice, the pure delay is the ITD and it is commonly cascaded in either the left or right HRTF of each left-right HRTF pair. It is important to stress that the calculation of ITD is beyond the scope of this work. Various studies address the estimation of ITD, notably in [6], [42]. Besides, unlike previous works [16], [25], [28], [32], we do not personalize the HRTFs or the directional transfer functions (i.e. mean removed HRTFs [22]) directly. Instead, we customize the intraconic component of HRTFs resulting from a spectral decomposition of HRTFs magnitude as suggested by Romigh and Simpson [17]. Here, we focus only on the spectral features of the intraconic component of HRTFs magnitude and, unless otherwise stated, when we refer to HRTF we are referring to its intraconic portion. Finally, in this work, we only use the interaural coordinate system [43] depicted in Fig. 1.

III. Methodology

Fig. 2. summarizes the pipeline of our HRTF customization approach. First, the extraction of the intraconic portion from full

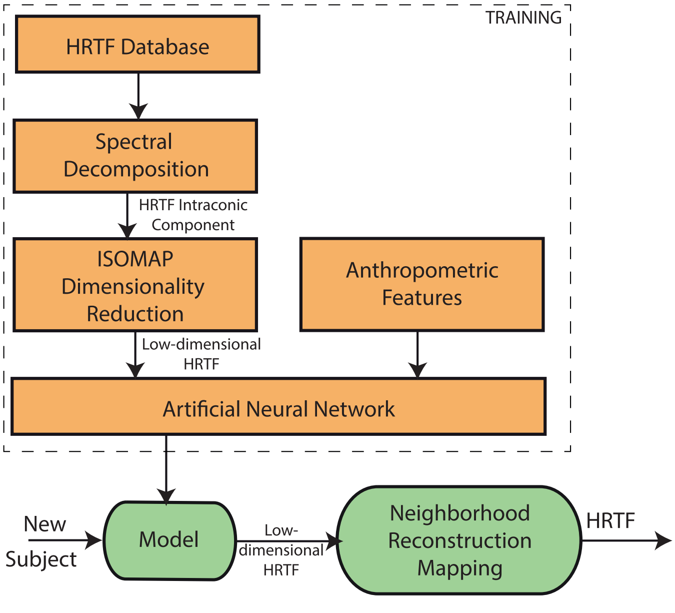


Fig. 2. Pipeline of our HRTF customization approach.

HRTFs aims at providing the most important cues for individualization, leaving out subject-independent cues. Then, Isomap with a custom graph construction procedure performs a nonlinear mapping of the intraconic component of HRTFs to a low-dimensional space. Subsequently, an ANN learns a regression model from a training dataset to relate anthropometric features to low-dimensional HRTFs. Finally, for a new subject with known anthropometric parameters, the model predicts his low-dimensional HRTFs which in turn are mapped back to the high-dimensional space by means of a neighborhood-based reconstruction approach.

We performed simulations of our approach on the CIPIC HRTF database [43]. We estimated the performance of such simulations using k-fold cross-validation and spectral distortion as metric. For comparison, we implemented PCA instead of Isomap for dimensionality reduction, and we tested the full HRTFs instead of their intraconic component. In summary, four conditions were tested: Isomap over full HRTFs, Isomap over the intraconic portion of HRTFs, PCA over full HRTFs and PCA over the intraconic portion of HRTFs. We also performed paired t-tests between the aforementioned conditions.

We chose only PCA for comparison because in this work we aim at exploring, first, whether it is worth using more complex techniques in the dimensionality reduction stage of anthropometry-based methods. Therefore, here we preferred to focus on how to construct and interpret a single manifold for all subjects, which had also not been addressed by prior works using manifold learning. Finally, we use a spectral distortion metric, as widely used in similar studies.

IV. HRTF Personalization

A. Spectral Decomposition

In a recent study, Romigh and Simpson [17] decomposed the HRTF at each location as the sum of average, lateral and intraconic spectral components. First, they obtained directional spectra by subtracting the mean across all locations (i.e. the average component) from each HRTF. Then, they calculated the lateral component for each azimuth angle as the median spectrum of all directional spectra measured at that azimuth angle. Lastly, they computed the intraconic component by subtracting the corresponding lateral component from the directional spectra at each location [17]. In order to recover the original HRTF spectrum at each location, they added together the corresponding average, lateral and intraconic components. Finally, the complex-valued HRTF were recovered using minimum phase assumptions.

After a series of psychoacoustic experiments where a listener's component were swapped out for the corresponding KEMAR's component, Romigh and Simpson found that the intraconic component encodes the most important cues for HRTF individualization and localization is only minimally affected by introducing non-individualized cues into the other HRTF components [17].

Based on these results, we used the intraconic spectral components as ground-truth HRTFs instead of the full ones. For simplicity, we will use the term intraconic HRTF when referring to its intraconic component.

B. Dimensionality Reduction using Isomap

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^D$ be a high-dimensional dataset in a $D \times N$ matrix of N sample vectors $\mathbf{x}_i$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^d$ be a corresponding low-dimensional representation in a $d \times N$ matrix of N sample vectors $\mathbf{y}_i$, where $d < D$.

Isomap is a nonlinear dimensionality reduction technique first introduced in [37] that provides a method for reducing $\mathbf{X}$ into a low-dimensional embedding $\mathbf{Y}$. Linear dimensionality reduction methods such as PCA attempt to preserve pairwise Euclidean distances by retaining most variance as possible [37]. However, such techniques does not take into account the datapoint neighborhood [44].

On the other hand, Isomap aims to maintain the intrinsic geometry of data (i.e. the datapoint neighborhood relationships) by preserving the pairwise geodesic distances (i.e. the distance over the manifold) [37]. For example, in nonlinear manifolds such as in the Swiss Roll dataset [44], PCA might map two datapoints as near points as measured by the Euclidean distance, while their geodesic distance is much larger.

Isomap can be summarized in three steps. The first step is to construct a graph $G(V, E)$ on the high-dimensional dataset $\mathbf{X}$. Each sample $\mathbf{x}_i \in \mathbf{X}$ is represented by a node $v_i \in V$, and two nodes v_i and v_j are connected by an edge $(v_i, v_j) \in E$ with length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ is one of the K nearest neighbor of $\mathbf{x}_j$. The edge length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ is given by some distance metric between $\mathbf{x}_i$ and $\mathbf{x}_j$ [37]. A common metric and the one used in this paper is the Euclidean distance.

In the second step, we calculate the geodesic distance between each pair of points by computing the shortest path between these two nodes in G. Then, after calculating the geodesic distances between all datapoints in $\mathbf{X}$, they are stored pairwise in a matrix $\mathbf{D}_G$. The pairwise geodesic distance matrix $\mathbf{D}_G$ represents the geodesic distances between all samples on the manifold [45].

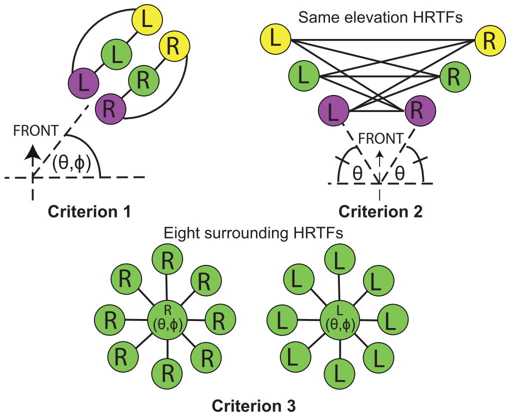


Fig. 3. Illustrative example of the criteria to construct the Isomap's graph for $P = 3$ subjects. Color represents HRTFs of the same subject, and (θ, ϕ) represents the azimuth and elevation in the interaural coordinate system. L = Left ear and R = Right ear.

The third and final step is to construct the d-dimensional embedding by applying multidimensional scaling (MDS) on $\mathbf{D}_G$ [44]. Formally, the eigenvectors of the double-centered matrix $\tau(\mathbf{D}_G)$ are calculated, where $\tau(\mathbf{D}_G) = -\mathbf{H}\mathbf{S}_G\mathbf{H}/2$, $\{S_G\}_{ij} = (\{D_G\}_{ij})^2$ (i.e. $\mathbf{S}$ is the matrix of squared distances) and $H_{ij} = \delta_{ij} - 1/N$ (i.e. $\mathbf{H}$ is the centering matrix). Recall that N is the number of sample points and δ is the Kronecker delta function. Finally, let λ_p be the pth eigenvalue (in decreasing order) of the matrix $\tau(\mathbf{D}_G)$, and v_p^i be the ith component of the pth eigenvector. Then set the pth component of the d-dimensional coordinate vector $\mathbf{y}_i$ equal to $\sqrt{\lambda_p} v_p^i$ [37].

In the first step of Isomap, we need to construct a graph, i.e., we need to select a number of neighbors for each high-dimensional point. Common approaches construct the graph by finding the K nearest neighbors or all neighbors within a specified radius r of each data point. In general, neighborhood selection in Isomap presents an opportunity to incorporate a priori knowledge from data [46]. With this in mind, we aim at constructing the graph G by taking advantage of the existing correlations among the HRTFs at different directions, frequencies, and individuals. One of our contributions is our graph G construction procedure:

Criterion 1. if $\mathbf{x}_i$ and $\mathbf{x}_j$ represent HRTFs of the same location and ear but different subject, then connect them.

In previous studies [23], [25], [26], [28], [30], they performed dimensionality reduction separately for each direction. Here, instead of applying Isomap separately for each location and ear, with this criterion, we tried to exploit the correlation of HRTFs among subjects across same directions. Using this criterion, $P - 1$ neighbors were obtained, where P is the number of subjects in the dataset $\mathbf{X}$.

Criterion 2. Let (θ_i, ϕ_i) and (θ_j, ϕ_j) be interaural coordinates (azimuth θ and elevation ϕ) of HRTFs represented by $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. Regardless of the subject, if $\mathbf{x}_i$ and $\mathbf{x}_j$ represent HRTFs of same elevation (i.e. $\phi_i = \phi_j$) and opposite ears, and θ_i is the mirror horizontal azimuth of θ_j (i.e. $\theta_i = -\theta_j$), then connect $\mathbf{x}_i$ and $\mathbf{x}_j$.

The intuition behind this criterion was to take advantage of the correlation existing due to left-right symmetry of HRTFs at frequencies below 5.5 kHz [47]. Applying this criterion, P neighbors were obtained.

Criterion 3. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be HRTFs of the same subject and ear. If $\mathbf{x}_j$ is one of the eight HRTFs surrounding $\mathbf{x}_i$, then connect them.

The intuition behind this criterion was to emphasize the similarities between spatially close HRTFs of the same subject and ear. Using this criterion, eight neighbors were obtained.

With the aim of clarifying how the above mentioned criteria were applied, Fig. 3. shows an illustrative example. Note that with our criteria, it is straightforward to prove that the constructed graph G is always connected.

Before applying Isomap, we first need to select the number of neighbors, K, and the intrinsic dimensionality, d. Due to the criteria proposed for the graph construction explained earlier, the number of neighbors was set to $K = 2P + 7$, i.e., $P - 1$ from Criterion 1, P from Criterion 2 and eight from Criterion 3. We determined the intrinsic dimensionality by means of the maximum likelihood intrinsic dimensionality estimator [48]. This dimensionality estimator attempts to reveal the intrinsic geometric structure of the observed data and it has demonstrated to be a good choice in manifold learning problems [44], [49].

Finally, note that, unlike previous works [23], [25]–[28], [30], we applied dimensionality reduction only once, over the entire dataset, for HRTFs of all subjects, directions and ears. This way, as tensor-based approaches [31]–[34] do, we tried to preserve the multi-factor (i.e. frequency, direction and subject) nature of HRTFs.

C. *Regression using an Artificial Neural Network*

Artificial Neural Networks (ANN) are systems capable of approximating nonlinear functions of their inputs. Since the relationship between HRTFs and anthropometric parameters is very complex, a nonlinear predictor is suitable for this task. Here, we used a back propagation ANN whose inputs are s anthropometric parameters, the azimuth angle, the elevation angle, and the ear (Left $= 1$, Right $= -1$). The outputs of the ANN are the low-dimensional HRTFs obtained by Isomap. Besides, the ANN uses sigmoid activation functions in the hidden layer and a linear activation function in the output layer.

We trained the ANN using Levenberg-Marquardt optimization and an early stop approach for improving generalization and to avoid overfitting. This way, we used a training subset for updating the network parameters. We also monitored a validation subset during the training process. When the validation error increased for 10 iterations, the training was stopped and the network parameters at the minimum of the validation error were returned.

We varied the number of hidden layer units and selected 35 hidden nodes that produced the lowest mean squared validation error. With this network topology, we achieved a mean squared validation error of 0.0078 that corresponds to a 0.91 coefficient of determination (R^2-value).

After the regression model is learned, the individual HRTF on the low-dimensional space for a new subject can be predicted by his anthropometric parameter measurements.

Finally, since our approach trains only one ANN for all HRTF data, the ANN exploits the relationships between low-dimensional components of HRTFs across directions and ears.

D. Neighborhood Reconstruction Mapping

Unlike linear reduction techniques, Isomap produce a low-dimensional embedding $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \in \mathbb{R}^d$ from the samples in $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^D$ without generating an explicit map [44]. As we are interested in high-dimensional HRTFs, we need to project a low-dimensional HRTF predicted by the ANN back into the original space. Since Isomap assumes that a sample and its neighbors are locally linear, we can perform the mapping using a linear combination of a sample's K neighbors. Thus, the reconstructed HRTF, $\hat{H}$, is

$$\hat{H} = \sum_{i=1}^{K} w_i \mathbf{x}_i. \tag{1}$$

To calculate the weights w_i, we followed Brown et al. [45], and chose w_i to be the inverse Euclidean distance between the sample and the neighbor i in the low-dimensional embedding.

V. Experiments

We implemented the proposed personalization method according to the block diagram in Fig. 2. Next, we define the elements and conditions of the simulations.

1) HRTF Database: We used the publicly available CIPIC database [43] which contains Head-Related Impulse Responses (HRIRs) of both ears measured for 45 subjects at 25 azimuths and 50 elevations (i.e. $M = 1250$ locations per subject and ear) in the interaural coordinate system. We selected only the subjects whose anthropometric features were complete (i.e. 35 subjects). Because not all anatomical features of CIPIC database are relevant for HRTF individualization, we selected $s = 8$ anthropometric parameters according to [50]: head depth, pinna offset back, cavum concha width, fossa height, pinna height, pinna width, pinna rotation angle and pinna flare angle. For selecting those parameters, the authors in [50] performed a statistical analysis in the entire virtual auditory space (i.e. azimuth and elevation) based on PCA, Pearson's product-moment correlation coefficient analysis and multiple linear regression analysis. Note that the only parameter related to head dimensions is head depth. Although both head depth and head width are important for ITD estimation [6] (head height is less relevant), keep in mind that our work does not deal with ITD but with the spectral features of minimum phase's magnitude of HRTFs as stated in Section II. Thus, the fact that we use head depth as the only parameter related to head dimensions in our study does not pose a critical problem.

2) HRIR Pre-processing: Each HRIR from CIPIC database has roughly 4.5 ms (i.e. 200 samples long) for a frequency sampling of 44.1 kHz and 16 bit resolution. First, we transformed each HRIR into a HRTF by means of a 512-point FFT. In order to reduce the effects of the limitations in the frequency response of the equipments utilized for HRIR measurement, we filtered the HRTFs to retain frequencies between 200 Hz and 15 kHz, leaving 172 magnitude coefficients (i.e. each HRTF is a point in a $D = 172$ dimensional space). Finally, we applied the spectral decomposition described in Section IV-A to the filtered HRTFs, preserving the intraconic component. In order to analyze the effects of personalizing the intraconic HRTFs instead of the full ones, we performed the same simulation for both conditions.

3) k-fold Cross-Validation: We divided the dataset into five folds of seven subject each. Then, we applied k-fold cross-validation, using four folds for training (i.e. $P = 28$ subjects) and one for testing. So, we estimated the model from $N = 2 \cdot P \cdot M = 70,000$ HRTFs.

4) Dimensionality Reduction: For Isomap, the intrinsic dimensionality was estimated by means of the maximum likelihood intrinsic dimensionality estimator [48]. So, we reduced the N HRTFs of dimension $D = 172$ to $d = 5$ dimensions. Since the number of subjects is $P = 30$, for the Isomap graph construction, each HRTF is connected to $K = 2P + 7$ neighbors. On the other hand, instead of Isomap, we implemented also PCA for comparing both methods. For PCA, we used $d = 5$ components that correspond to 88% of variance retained, which is in line with previous studies showing that five PCA components (approximately 90% of variance) capture the most perceptually relevant properties of HRTFs [22], [51]. Finally, k-fold cross-validation was also applied to evaluate the PCA performance. Both Isomap and PCA were implemented using Matlab Dimensionality Reduction Toolbox [44].

5) Neural Network: As explained in Section IV-C, the inputs of the artificial neural network are the $s = 8$ anthropometric parameters, the ear $(\mathrm{Left} = -1,\ \mathrm{Right} = 1)$, the azimuth and the elevation. The outputs are the low-dimensional HRTFs. We used Matlab Neural Network Toolbox to implement the ANN.

6) Performance Metric: As an error metric, we chose the mean spectral distortion in dB defined by

$$SD_M = \sqrt{\frac{1}{N_f} \sum_{k=1}^{N_f} \left(20\log_{10} \frac{|H(k)|}{\left|\hat{H}(k)\right|} \right)^2}, \tag{2}$$

where H and $\hat{H}$ represent the measured and reconstructed HRTF respectively and N_f is the number of frequency points. The reconstructed HRTF, $\hat{H}$, was calculated using Equation (1).

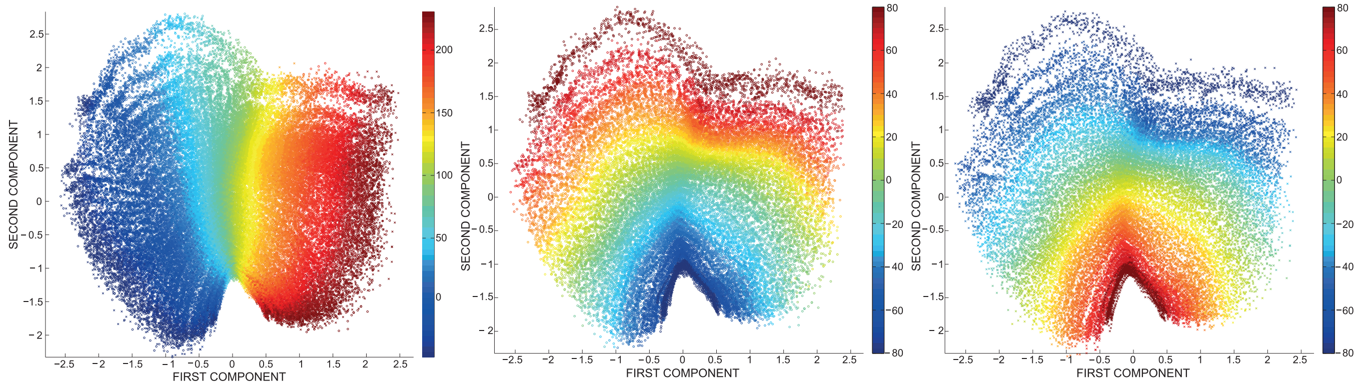
In summary, we tested the following four conditions:

1. Intraconic HRTFs as ground-truth and Isomap as dimensionality reduction method, labeled as Intra-ISO.
2. Full HRTFs as ground-truth and Isomap as dimensionality reduction method, labeled as Full-ISO.
3. Intraconic HRTFs as ground-truth and PCA as dimensionality reduction method, labeled as Intra-PCA
4. Full HRTFs as ground-truth and PCA as dimensionality reduction method, labeled as Full-PCA

VI. Analysis and Results

A. Isomap Manifold Analysis

It is important to analyze how the Isomap embedded components relate to source azimuth and elevation. Fig. 4(a). shows



a) Both Ears. Color represents Elevation.

b) Left Ear. Color represents Azimuth.

c) Right Ear. Color represents Azimuth.

Fig. 4. Two-dimensional Manifold. All Isomap components are normalized to have zero mean and unit variance. (a) Both Ears. Color represents Elevation. (b) Left Ear. Color represents Azimuth. (c) Right Ear. Color represents Azimuth.

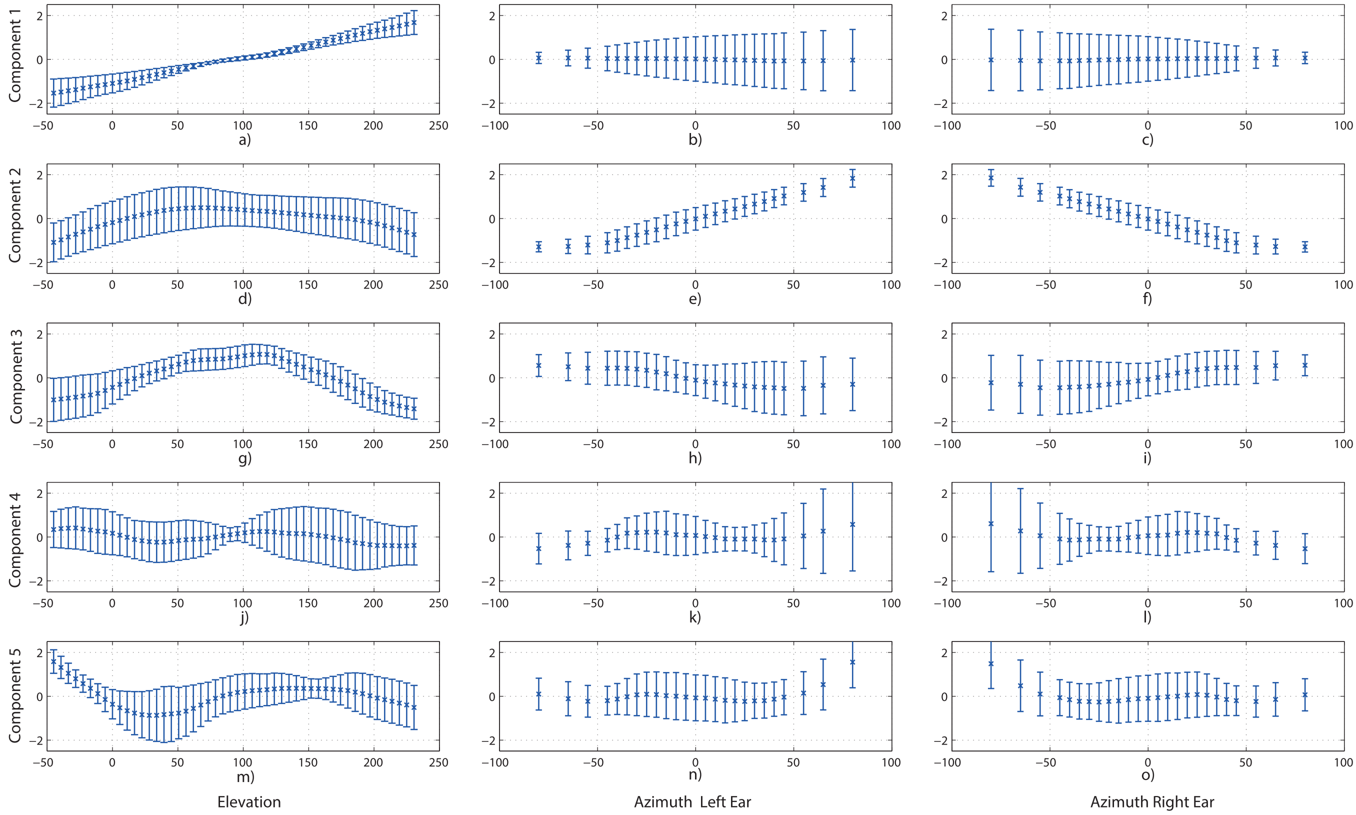


Fig. 5. Isomap components as a function of location. In all plots, datapoints represent the mean across all subjects for a specific location and error bars correspond to a ± 1 standard deviation interval. For azimuth data (second and third column plots), we separated left and right ear plots to put in evidence ipsilateral and contralateral variability across individuals. All Isomap components are normalized to have zero mean and unit variance.

the two-dimensional manifold (i.e. first embedded dimension vs second one) where the color represents the elevation angle. We observe that the first component of Isomap embedding roughly increases with elevation. That tendency is confirmed by the correlation coefficient between elevation angle and the first component value, which is 0.94.

Figs. 5(a) through 5(c) present the first component as a function of source location. In Fig. 5(a). is evident the strong correlation between elevation and first component. Besides, first component's value is negative for front locations (i.e. $\phi < 90°$) while it is positive for rear positions. This pattern suggests that the first component can distinguish front locations from back ones. Notice also in the same figure that there is a tendency for the first component to increase in magnitude as the source moves from the frontal plane (i.e. $\phi = 90°$). On the other hand, in Figs. 5(b) and 5(c), there is no clear pattern between first component and azimuth as in the case of elevation. In fact, error bars tend to increase as source moves towards contralateral locations and their mean value (i.e. datapoints in plots) keep roughly constant with respect to azimuth angle.

Figs. 4(b) and 4(c) show the same two-dimensional manifold but this time the color represents azimuth angles. We

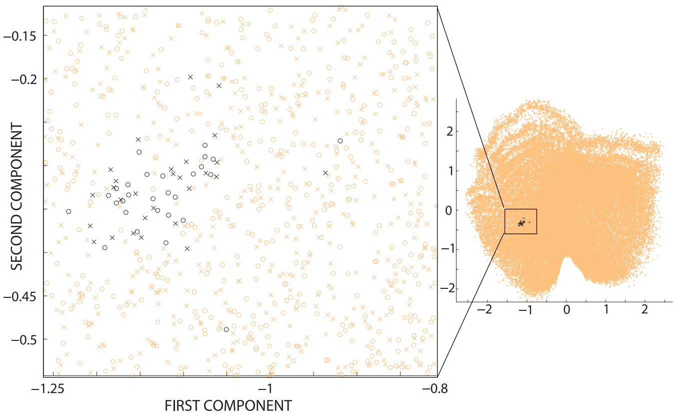


Fig. 6. Two-dimensional manifold. Black datapoints are low-dimensional HRTFs of same direction but different subject. Left and right ear are represented by circle and cross markers, respectively.

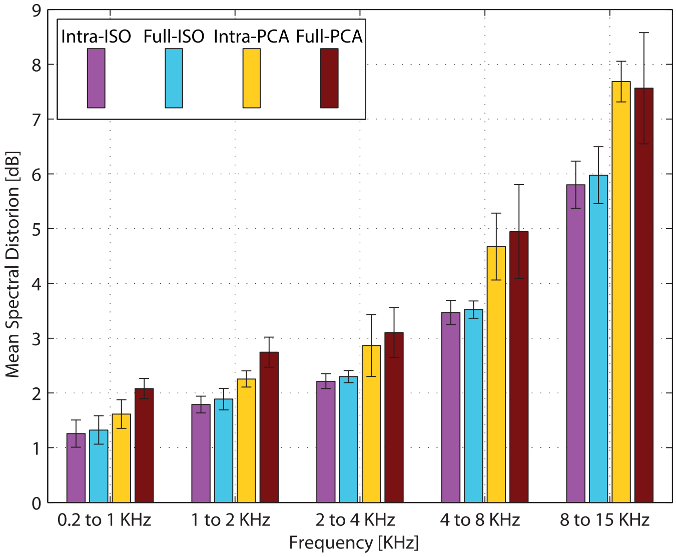


Fig. 7. Mean Spectral Distortion for different frequency bands. Error bars represent 95% confidence intervals ($\pm 2\sigma$). MSD values are tabulated in Table II.

plotted separately low-dimensional HRTFs of each ear to put in evidence the symmetry introduced in the graph construction procedure proposed in Section IV-B. Observe that the second component of Isomap embedding roughly increases with azimuth for left ear. This trend, as expected because the left/right symmetry, is inverted (i.e. Isomap second component decreases with azimuth) for low-dimensional HRTFs of the right ear. Moreover, the correlation coefficient between azimuth angle and Isomap second component is 0.879 for left ear and -0.880 for right ear.

Figs. 5(e) and 5(f) present the second component as a function of source azimuth where it is evident the strong correlation between azimuth and Isomap second component. Observe also that the component's value tends to be roughly positive for contralateral locations and negative for ipsilateral directions. On the other hand, Fig. 5(d) shows that the second component encodes also some elevation cues, although this relationship is not as strong as in the case of first component. Notice that the second component only tends to decrease when the source moves from frontal plane (i.e. $\phi = 90^\circ$). However, it is not capable of distinguishing front locations from back ones since it exists front and back elevation angles that produce the same component's value. Furthermore, error bars are larger when compared to Fig. 5(a) which suggests that second component varies widely across subjects at a specific elevation angle.

Fig. 5(g) through 5(i) show the third Isomap component as a function of direction. In Fig. 5(g), the third component tends to decrease as the source move from frontal plane, which revels that the third component is capable of encode some elevation information. However, this component can't resolve front/back ambiguities because back elevation angles can produce the same component's value as front ones. Thus far, the third component behavior is similar to the second component. Nonetheless, Figs. 5(h) and 5(i) show no clear pattern between azimuth and the third component as in the case of the second component.

So far, we have mainly analyzed directional relationships of Isomap components. With respect to inter-subject differences captured by Isomap, they are far more complex to visualize than directional ones due to its non-linear nature. Still, observe the variability of the black points in Fig. 6. These are low-dimensional HRTFs of same direction but different subject in the two-dimensional manifold where most of variance is captured (i.e. first and second dimensions). Although they are relatively close to each other, as expected because of our graph construction procedure, their high variability is due to inter-subject differences.

Moreover, in general, error bars in Fig. 5. for all Isomap components increase for contralateral locations and the same pattern is observed when the source moves from the frontal plane. This trend confirms that the head shadowing effect and vertical cues (mostly introduced by the pinna) causes wide variations in HRTFs.

In summary, we found that the Isomap first component is strongly correlated with elevation and the second one with azimuth. Although the second and third components also encode some elevation cues in a lesser degree, they are not able to distinguish front locations from back ones. The pattern of the remaining two dimensions shown in Fig. 5 is considerably more complicated. Still, in general, for all components inter-subject variability increases for contralateral locations and when sources move away from frontal plane.

B. Spectral Distortion Analysis

As stated in Section V, we performed simulations for four conditions: Intra-ISO, Full-ISO, Intra-PCA and Full-PCA.

Fig. 7 shows the mean spectral distortion (MSD) for four frequency bands. In the same figure, error bars represent 95% confidence intervals ($\pm 2\sigma$). Observe that, as expected, the MSD increases with frequency but Isomap performed better than PCA, specially in frequencies above 4 kHz that normally are harder to predict because of their high inter-subject variability. Still, in Intra-ISO and Full-ISO conditions, Isomap manage s to keep MSD roughly below four dB and six dB for 4–8 kHz and 8–15 kHz bands, respectively. Moreover, note that the

TABLE I
PAIRED T-TEST FOR DIFFERENT FREQUENCY BANDS. ALL BOLD ENTRIES REFER TO A STATISTICALLY SIGNIFICANT DIFFERENCE AT A 95% CONFIDENCE LEVEL. THE LOWER THE P-VALUE THE BETTER. P-VALUES ARE SHOWN UP TO THE THIRD DECIMAL PLACE

Paired t-test	0.2 - 1.0 kHz	1.0 - 2.0 kHz	2.0 - 4.0 kHz	4.0 - 8.0 kHz	8.0 - 15.0 kHz
Intra-ISO with respect to Full-ISO	**0.001**	**0.007**	**0.008**	**0.046**	**0.003**
Intra-ISO with respect to Intra-PCA	**0.001**	**0.000**	**0.004**	**0.002**	**0.000**
Full-ISO with respect to Full-PCA	**0.000**	**0.000**	**0.000**	**0.002**	**0.003**
Intra-PCA with respect to Full-PCA	**0.000**	**0.001**	**0.029**	0.175	0.365

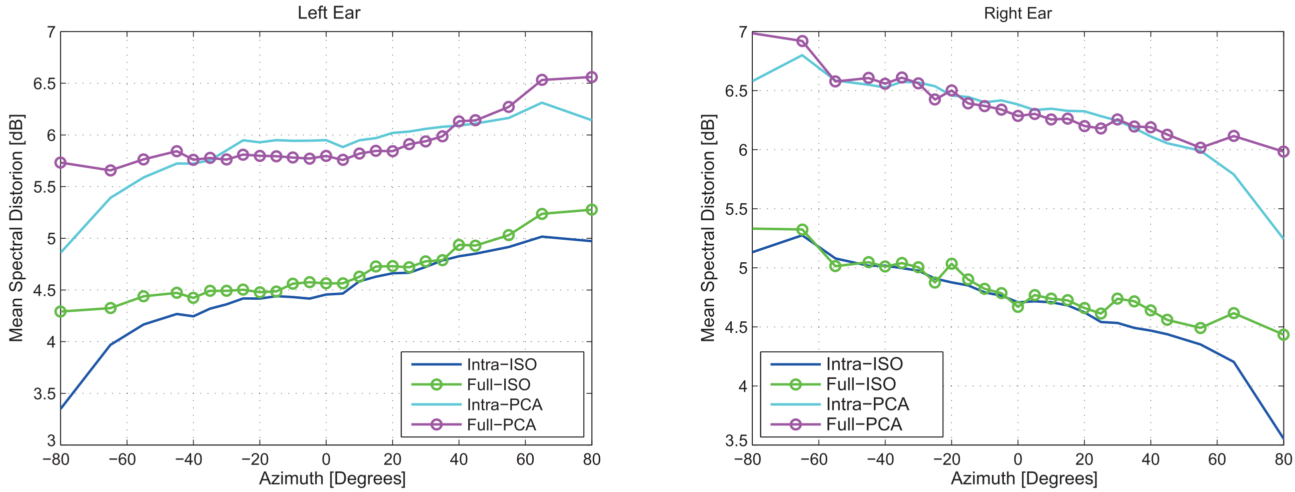


Fig. 8. Vertical Mean Spectral Distortion.

TABLE II
MEAN SPECTRAL DISTORTION IN dB FOR DIFFERENT FREQUENCY BANDS. BOLD ENTRIES REFER TO THE LOWER MSD IN A SPECIFIC FREQUENCY BAND

Band [kHz]	Intra-ISO	Full-ISO	Intra-PCA	Full-PCA
0.2 - 1.0	**1.2572**	1.3231	1.6134	2.0795
1.0 - 2.0	**1.7882**	1.8873	2.2548	2.7443
2.0 - 4.0	**2.2142**	2.2972	2.8642	3.1020
4.0 - 8.0	**3.4673**	3.5218	4.6730	4.9446
8.0 - 15.0	**5.8007**	5.9760	7.6836	7.5635

TABLE III
PAIRED T-TEST FOR VERTICAL MEAN SPECTRAL DISTORTION. ALL BOLD ENTRIES REFER TO A STATISTICALLY SIGNIFICANT DIFFERENCE AT A 95% CONFIDENCE LEVEL. THE LOWER THE P-VALUE THE BETTER. P-VALUES ARE SHOWN UP TO THE THIRD DECIMAL PLACE

Paired t-test	Left	Right
Intra-ISO with respect to Full-ISO	**0.000**	**0.000**
Intra-PCA with respect to Full-PCA	0.351	0.279

confidence interval shows that in the frequency ranges 2 to 4 kHz and 4 to 8 kHz the Isomap conditions have much less variability than their PCA counterparts.

Table I summarizes the results of a series of paired t-tests along different frequency bands, where bold entries indicate a statistically significant difference at a 95% confidence level ($p < 0.05$). We found that for all frequency bands, both Isomap conditions data (i.e. Intra-ISO and Full-ISO) come from a population with a mean less than its corresponding PCA condition (i.e. Intra-PCA and Full-PCA), confirming that Isomap performs better than PCA. Moreover, for all frequency bands, Intra-ISO shows a small but statistically significant improvement over Full-ISO. On the other hand, for PCA, although Intra-PCA presents a more evident improvement over Full-PCA, this improvement is only statistically significant in low frequency bands. Still, observe that the error bars, particularly for high frequency bands, are in general smaller for both intraconic conditions.

Observe that for the intraconic conditions in Table I, although the relatively high p-value of 0.046 at the 4–8 kHz band corresponds to a statistically significant difference at a 95% confidence level, they do not differ in a statistically significant way at a 99% confidence level. However, observe also that MSD sub-band analysis has a strong propensity to hide some important causes of distortion, e.g., MSD at ipsilateral and contralateral locations tend to cancel each other out due to the averaging across ears and directions. For the aforesaid reasons, it is convenient to analyze the MSD of each ear separately as a function of sound source position.

Fig. 8. shows the MSD as a function of azimuth. Because we calculate this MSD across all elevations for a specific azimuth (i.e. the MSD in the cone of confusion), we refer it as vertical MSD. As before, both Isomap conditions present less vertical MSD than the PCA ones. On the other hand, in a general way, Intra-ISO condition performs better than Full-ISO. This is especially notable for ipsilateral locations where Intra-ISO reaches up to 1 dB improvement over Full-ISO. For contralateral locations, due to the head shadowing effect, the vertical MSD tends to increase for both Full-ISO and Intra-ISO conditions.

Furthermore, in a paired t-test performed for each ear separately (refer to Table III), the Intra-ISO condition showed a statistically significant improvement ($p < 0.05$) over the Full-ISO condition. Nonetheless, in a similar paired t-test

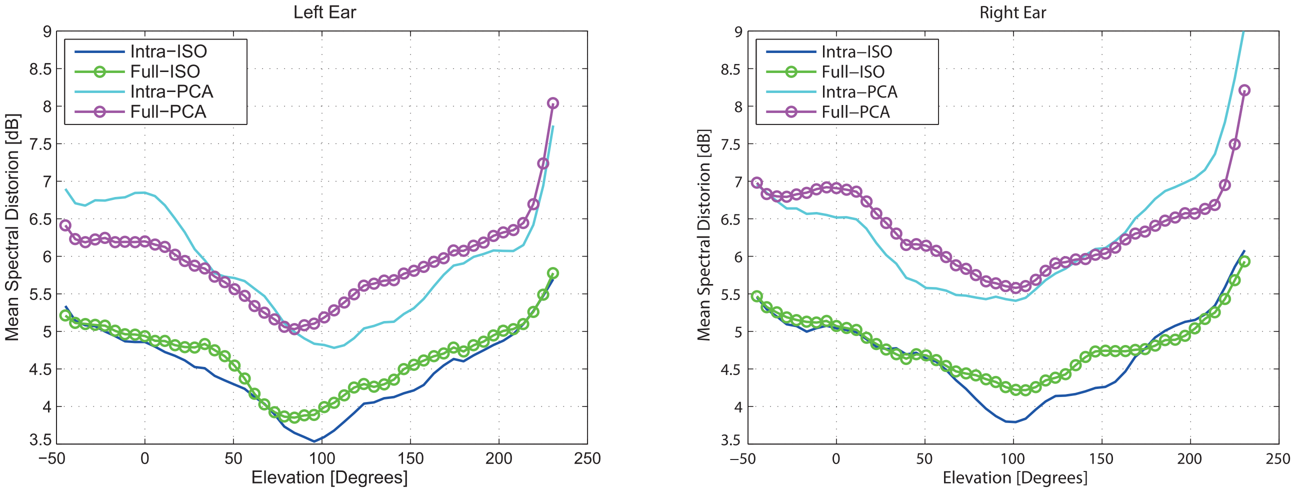


Fig. 9. Lateral Mean Spectral Distortion.

TABLE IV

PAIRED T-TEST FOR LATERAL MEAN SPECTRAL DISTORTION. ALL BOLD ENTRIES REFER TO A STATISTICALLY SIGNIFICANT DIFFERENCE AT A 95% CONFIDENCE LEVEL. THE LOWER THE P-VALUE THE BETTER. P-VALUES ARE SHOWN UP TO THE THIRD DECIMAL PLACE

Paired t-test	Left	Right
Intra-ISO with respect to Full-ISO	**0.000**	**0.000**
Intra-PCA with respect to Full-PCA	0.368	0.271

between Intra-PCA and Full-PCA, no statistically significant improvement was found. This last result is not surprising taking into account that we found no statistically significant difference (refer to Table I) between PCA conditions for high-frequency bands which in turn are the major contributors to elevation perception. Moreover, we expected some improvement of Intra-ISO over Full-ISO because the information lost after the spectral decomposition (i.e. the lateral and average components) is less perceptually relevant for HRTF personalization [17].

Fig. 9. shows the MSD as a function of elevation. Because we calculate this MSD across all azimuths for a specific elevation, we refer it as lateral MSD. In this figure, it is clear that Isomap performs better than PCA in all conditions. Observe that, for all conditions, the lateral MSD decreases as the sound source moves toward the frontal plane (i.e. $\phi = 90°$), reaching a minimum around top directions. Note also that the Full-ISO and Intra-ISO lateral MSD stays roughly below 5.5 dB, except for very low elevations at back locations where the lateral MSD reaches up to 6 dB. This increase of lateral MSD confirms that complex scattering of sound waves coming from low elevations are harder to predict. Besides, the Intra-ISO lateral MSD shows a modest improvement over Full-ISO that is more prominent for back locations closer to the frontal plane.

Again, we performed paired t-tests for each ear separately between Intra-ISO and Full-ISO, and between Intra-PCA and Full-PCA conditions (refer to Table IV). We found that Intra-ISO data comes from a population with a MSD less than Full-ISO condition ($p < 0.05$). However, we did not found statistically significance difference between PCA conditions. Taking into account that high-frequency cues are needed for front/back discrimination, this last result is in accordance with the lack of statistically significance difference found in high-frequency bands for PCA conditions in Table I.

Although prior works have performed experiments on different baseline datasets, anthropometric features, frequency bands and spatial locations, we would like to make a reasonable comparison with the approaches used in those works in terms of the spectral distortion reported by them. However, it should be kept in mind that most works do not report standard deviation values, which makes a fair comparison harder. Personalization methods based on linear dimensionality reduction techniques in conjunction with linear regressors [25], [26] report MSD scores across all frequencies near 6 dB, which is higher than our MSD (4.6 dB, $\sigma = 0.15$). On the other hand, MSD across all frequencies on customization techniques using linear dimensionality reduction together with nonlinear regressors ranges from roughly 3 [28] to 5 dB [30]. Although our results are slightly better than [30], they are lower than [28], which is–according to our research–the smallest score reported among studies using MSD. Finally, in general, tensor-based approaches [31]–[34] perform better than PCA-based methods, reaching their best performance at 3.5 dB in the frequency band 50 Hz–8 kHz [34], which is comparable to our results in the frequency band 0.2–8 kHz (2.9 dB, $\sigma = 0.0735$).

So far, we have restricted our comparison to methods using MSD as metric. We considered relevant to analyze our data using the variance metric proposed by Middlebrooks [21], which produced very similar results to those using MSD in the sense that Intra-ISO presented the best performance, while both PCA conditions performed the worst. The error of Intra-ISO for frequency bands up to 4 kHz is less than 3.58 dB^2, grows in the 4 to 8 kHz band (8.9 dB^2, $\sigma = 0.7$) and reaches its maximum for frequencies above 8 kHz (24.96 dB^2, $\sigma = 1.13$). As a reference, using a frequency scaling approach, Middlebrooks [21] found that the 95 percentile of inter-subject spectral difference (measured by the variance metric) across 990 pairs of subjects was 9.3 dB^2. This confirms what we found using MSD with respect to the weaker performance of our method in frequency bands above 8 kHz.

VII. Conclusion and Future Work

The findings of this paper show that Isomap has proven to be a powerful technique to discover the manifolds of spatial hearing. By incorporating important prior knowledge, Isomap was capable of explaining the directional factor (i.e. azimuth and elevation) of spatial audio. Even though no Isomap component alone explains inter-subject differences, the wide inter-dimension variability observed confirms its nonlinear behavior. Hence the importance of nonlinear regressors such as Artificial Neural Networks (ANN) to map anthropometric features into low-dimensional HRTFs. Unlike regression techniques such as Support Vector Regression, ANN is a multiple output predictor that permits to exploit the correlations between Isomap components (i.e. inter-dimension correlations). Moreover, instead of constructing one regression model per direction, our approach lets to construct a single model that does not break the inherent multifactor nature of HRTFs (i.e. frequency, direction and subject factors).

In all simulations performed, the results show that Isomap has a better performance and less variability than PCA as measured by the mean spectral distortion (MSD) with 95% confidence intervals. Furthermore, our results put in evidence that Isomap can capture high-frequency cues from intraconic HRTFs where PCA does not. Thus, we confirmed that the intraconic representation effectively encodes the most important cues for individualization of HRTFs.

On the other hand, the main weakness of Isomap is the lack of an explicit mapping function [44] to project new high-dimensional datapoints into an existing low-dimensional embedding (i.e. out-of-sample extension), and to reconstruct a low-dimensional datapoint into a high-dimensional representation(i.e. back-projection). Out-of-sample extension might be performed by means of the Nystrom approximation [52], [53], so that, for new datapoints, there is no need to recalculate the entire manifold.

The back-projection is a more challenging problem to overcome. Here, we have reconstructed high-dimensional HRTFs using a linear combination of its neighbors (i.e. neighborhood-based reconstruction). It should be observed that the main weakness of this reconstruction is that its accuracy depends on how dense the initial database is. This problem might be addressed using some spatial HRTF interpolation before HRTF personalization to guarantee a more populated manifold. However, note that if the initial database is not sampled adequately in space, the resultant interpolated HRTFs will not be suitable to reconstruct the personalized HRTFs. In this sense, although the CIPIC dataset is one the most complete publicly available HRTF datasets including anatomical measurements, we expect that our method could perform better given a more suitable input dataset (i.e. higher spatial resolution, more subjects and better quality anatomical measurements). Lastly, although we chose the reconstruction weights to be the corresponding neighbor's euclidean inverse distances, there is an alternative approach where the weights are determined in a least-squares optimization. However, this approach proved to produce larger spectral distortion.

One question that arises for practical use is whether our method will produce an acceptable perceptual result. Since low-frequency distortion is low, we expect that cues acting on those bands will not be affected in listening tests. On the other hand, the error in high-frequency bands is relatively high, which will affect elevation perception. However, it should be kept in mind that we demonstrated that an important part of this distortion is due to contralateral and low elevation errors. Moreover, previous studies concluded that the spectral detail of HRTFs at high frequency is inaudible [54], which in turn, implies that the high contralateral error is likely to be, to some extent, perceptually irrelevant. Thus, in listening tests, we expect the localization accuracy to be good at lateral locations, reasonable at vertical directions but poor in low elevations.

Another problem that might arise is how to apply Isomap when the graph has two or more connected components. In such case, the resultant components would lie on different manifolds. Further studies might address this problem using techniques to merge multiple manifolds as proposed, e.g., in [55]. It should be noted that, since our approach guarantees a connected graph, we do not address the non-connected graph case.

In future work, we plan to explore a subband representation for HRTFs in conjunction with manifold learning. Since different localization cues act in different frequency bands, a subband representation would permit a more flexible way to construct the manifold structure. In this context, it would be possible to incorporate prior knowledge in the subband where this prior is effectively valid. For example, we could introduce left/right symmetry only in subbands where symmetry is more prominent.

We also plan to explore a multi-task learning (MTL) approach to learn the regression model. MTL learns multiple related tasks simultaneously using a shared representation aimed at improving generalization [56]. In the HRTF personalization context, a task could be, e.g., learning a regression model per direction. The MTL approach might preserve the multi-factor nature of HRTFs by using a shared representation instead of learning a single regression model as we do in this paper.

Acknowledgment

The authors would like to thank the four anonymous reviewers for their valuable time and useful contributions. We also wish to thank Ms. Estefania Grijalva who contributed to proofreading this manuscript as well as Senescyt (Ecuador) for all the support.

References

[1] D. R. Begault, *3D Sound for Virtual Reality and Multimedia*, Cambridge, U.K.: AP Professional, 1994.

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 111–123, Jul. 1993.

[3] V. Valimaki, A. Franck, J. Ramo, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 92–99, Mar. 2015.

[4] H. Möller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, 3–4, pp. 171–218, 1992.

[5] B. Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *J. Acoust. Soc. Amer.*, vol. 132, no. 1, pp. 282–94, Jul. 2012.

[6] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479, Jun. 2001.

[7] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Amer.*, vol. 112, no. 5, 2053, Oct. 2002.

[8] C. Brown and R. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, Sep. 1998.

[9] M. Geronazzo, S. Spagnol, and F. Avanzini, "Mixed structural modeling of head-related transfer functions for customized binaural audio delivery," *Proc. 18th Int. Conf. Digital Signal Process.*, 2013, pp. 1–8.

[10] M. Otani and S. Ise, "Fast calculation system specialized for head-related transfer function based on boundary element method," *J. Acoust. Soc. Amer.*, vol. 119, no. 5, pp. 2589, May 2006.

[11] Y. Kahana and P. Nelson, "Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models," *J. Sound Vibr.*, vol. 300, no. 3, pp. 552–579, 2007.

[12] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," *J. Acoust. Soc. Amer.*, vol. 132, no. 6, pp. 3832–3841, Dec. 2012.

[13] S. Hwang, Y. Park, and Y. Park, "Modeling and customization of head-related impulse responses based on general basis functions in time domain," *Acta Acust. united with Acust.*, vol. 94, no. 6, pp. 965–980, Nov. 2008.

[14] K. J. Fink and L. Ray, "Individualization of head related transfer functions using principal component analysis," *Appl. Acoust.*, vol. 87, pp. 162–173, Jan. 2015.

[15] K. Sunder, E.-L. Tan, and W.-S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 989–1000, Dec. 2013.

[16] F. Grijalva, L. Martini, S. Goldenstein, and D. Florencio, "Anthropometric-based customization of head-related transfer functions using Isomap in the horizontal plane," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, 2014, pp. 4473–4477.

[17] G. D. Romigh and B. D. Simpson, "Do you hear where I hear?: Isolating the individualized sound localization cues," *Frontiers Neurosci.*, vol. 8, no. 370, Dec. 2014.

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

[19] E. Torres, F. Orduña, and F. Arámbula, "Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database," *Appli. Acoust.*, vol. 97, pp. 84–95, Apr. 2015.

[20] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," *J. Acoust. Soc. Amer.*, vol. 136, no. 1, pp. 317–33, Jul. 2014.

[21] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1480–1492, 1999.

[22] D. Kistler and F. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1637–47, Mar. 1992.

[23] Q. Huang and Q. Zhuang, "HRIR personalisation using support vector regression in independent feature space," *Electron. Lett.*, vol. 45, no. 19, pp. 1002, 2009.

[24] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, "Enabling individualized virtual auditory space using morphological measurements," *Proc. 1st Pacific-Rim Conf. Multimedia (2000 Int. Symp. Multimedia Inf. Process.)*, 2000, pp. 235–238.

[25] T. Nishino, K. Iida, N. Inoue, K. Takeda, and F. Itakura, "Estimation of HRTFs on the horizontal plane using physical features," *Appl. Acoust.*, vol. 68, no. 8, pp. 897–908, 2007.

[26] S. Xu, Z. Li, and G. Salvendy, "Improved method to individualize head-related transfer function using anthropometric measurements," *Acoust. Sci. Technol.*, vol. 29, no. 6, pp. 388–390, 2008.

[27] D. Schönstein and B. Katz, "HRTF selection for binaural synthesis from a database using morphological parameters," *Proc. 20th Intl. Congr. Acoust.(ICA)*, Aug. 2010, pp. 1–6.

[28] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Appl. Acoust.*, vol. 69, no. 2, pp. 163–172, Feb. 2008.

[29] Q. Huang and Y. Fang, "Modeling personalized head-related impulse response using support vector regression," *J. Shanghai Univ.*, vol. 13, no. 6, pp. 428–432, 2009.

[30] Z. Wang and C. F. Chan, "HRIR customization using common factor decomposition and joint support vector regression," *Eur. Signal Process. Conf.*, 2013, pp. 1–5.

[31] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 3707–3710.

[32] G. Grindlay, and M. A. O. Vasilescu, "A multilinear (tensor) framework for HRTF analysis and synthesis," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 1, pp. 161–164, IEEE.

[33] Q. Huang and L. Li, "Modeling individual HRTF tensor using high-order partial least squares," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 58, 2014.

[34] P. Bilinski, J. Ahrens, M. Thomas, I. Tashev, and J. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4468–4472.

[35] R. Duraiswami and V. Raykar, "The manifolds of spatial hearing," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 285–288.

[36] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–6, Dec. 2000.

[37] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–23, Dec. 2000.

[38] B. Kapralos and N. Mekuz, "Application of dimensionality reduction techniques to HRTFs for interactive virtual environments," *Proc. Int. Conf. Advances Comput. Entertain. Technol.*, 2007, pp. 256–257, ACM.

[39] B. Kapralos, N. Mekuz, A. Kopinska, and S. Khattak, "Dimensionality reduced HRTFs: A comparative study," *Proc. Int. Conf. Adv. Comput. Entertain. Technol.*, Dec. 2008, pp. 59, ACM.

[40] H. Seung and D. Lee, "The manifold ways of perception," *Science*, vol. 290, pp. 2268–2269, 2000.

[41] A. Kulkarni, S. Isabelle, and H. Colburn, "On the minimum-phase approximation of head-related transfer functions," *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 1995, pp. 84–87.

[42] K. Watanabe, K. Ozawa, Y. Iwaya, Y. I. Suzuki, and K. Aso, "Estimation of interaural level difference based on anthropometry and its effect on sound localization," *J. Acoust. Soc. Amer.*, vol. 122, no. 5, pp. 2832–41, Nov. 2007.

[43] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2001, pp. 99–102.

[44] L. Van Der Maaten, E. Postma, and J. Van Den Herik, "Dimensionality reduction: A comparative review," *J. Mach. Learn. Res.*, vol. 10, pp. 1–41, 2009.

[45] W. M. Brown, S. Martin, S. N. Pollock, E. A. Coutsias, and J.-P. Watson, "Algorithmic dimensionality reduction for molecular structure analysis," *J. Chem. Phys.*, vol. 129, no. 6, Aug. 2008.

[46] K. S. Lawrence, and T. R. Sam, "Think globally, fit locally: Unsupervised learning of nonlinear manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, 2002.

[47] B. Xie, X. Zhong, D. Rao, and Z. Liang, "Head-related transfer function database and its analyses," *Sci. China Ser. G: Phys., Mech. Astron.*, vol. 50, no. 3, pp. 267–280, Jun. 2007.

[48] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," *Adv. Neural Inf. Process. Syst.*, 2005, NIPS Foundation.

[49] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 796–809, May 2008.

[50] M. Zhang, R. Kennedy, T. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in HRTF individualization," *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, 2011, pp. 213–218.

[51] J. Breebaart, "Effect of perceptually irrelevant variance in head-related transfer functions on principal component analysis," *J. Acoust. Soc. Amer.*, vol. 133, no. 1, pp. EL1–EL6, Jan. 2013.

[52] Y. Bengio, J. F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering," *Adv. Neural Inf. Process. Syst.*, 2004, NIPS Foundation.

[53] J. Platt, "Fastmap, metricmap, and landmark MDS are all Nystrom algorithms," *Proc. Int. Workshop Artif. Intell. Statist.*, 2005, pp. 261–268.

[54] B. Xie and T. Zhang, "The audibility of spectral detail of head-related transfer functions at high frequency," *Acta Acust. united with Acust.*, vol. 96, no. 2, pp. 328–339, Mar. 2010.

[55] H. Choi, S. Choi, A. Katake, Y. Kang, and Y. Choe, "Manifold alpha-integration," *Proc. PRICAI 2010: Trends Artif. Intell.*, 2010, pp. 397–408, Springer.

[56] J. Baxter, “A model of inductive bias learning,” *J. Artif. Intell. Res.*, vol. 12, pp. 149–198, 2000.

[57] K. Sunder, J. He, E. L. Tan, and W-S. Gan, “Natural sound rendering for headphones: integration of signal processing techniques,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 100–113, Mar. 2015.

Felipe Grijalva received a B.S. degree in electrical engineering and telecommunications from the Army Polytechnic School, Quito, Ecuador, in 2010, and a M.Sc. degree in electrical engineering with major in computing engineering from the University of Campinas, Campinas, Brazil, in July 2014. Currently, he is a Ph.D. candidate at the University of Campinas. His research focus areas include spatial audio, machine learning, computer vision applications and assistive technologies aimed at visually impaired people.

Luiz Martini received the B.S. (1976), M.S. (1981), and Ph.D. (1989) degrees from State University of Campinas, Campinas, Brazil, all in electrical engineering. He is a Professor and head of the Telecommunications Department at the School of Electrical and Computer Engineering, University of Campinas, Campinas, Brazil. His research interests lie in assistive technologies aimed at visually impaired people and education for engineering students with blindness and visual impairments.

Dinei Florencio received the B.S. and M.S. degrees from the University of Brasilia, Brasilia, Brazil, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA, all in electrical engineering. He has been a Researcher with Microsoft Research, Redmond, WA, USA, since 1999. Before joining Microsoft, he was a Member of the Research Staff at the David Sarnoff Research Center from 1996 to 1999. He was also a Co-Op Student with the AT&T Human Interface Laboratory (now part of NCR) from 1994 to 1996, and a Summer Intern at Interval Research Corporation, Palo Alto, CA, USA, in 1994. He has authored over 70 papers, and holds 55 granted U.S. patents. Dr. Florencio was General Co-Chair of MMSP 2009, WIFS 2011, Hot3D 2010 and 2013, and Technical Co-Chair of WIFS 2010, ICME 2011, and MMSP 2013. He is the Chair of the IEEE SPS Technical Committee on Multimedia Signal Processing (from 2014 to 2015), and a member of the IEEE SPS Technical Directions Board.

Siome Goldenstein received a Ph.D. in computer and information science from University of Pennsylvania in 2002, an M.Sc. in computer science from Pontifícia Universidade Catolica do Rio de Janeiro in 1997, and an Electronic Engineering degree from the Federal University of Rio de Janeiro in 1995. He is an Associate Professor at the Institute of Computing, University of Campinas, Unicamp, Brazil and a senior IEEE member. He is an Area Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (T-IFS), *Computer Vision and Image Understanding (CVIU)* and *Graphical Models (GMOD)*, has been in the program committee of multiple conferences and workshops. His interests lie in computational forensics, computer vision, computer graphics, and machine learning.