A Recommender System For Improving Median Plane Sound Localization Performance Based on a Nonlinear Representation of HRTFs

Felipe Grijalva, *Student Member, IEEE,* Luiz César Martini, Bruno Masiero, *Member, IEEE,* and Siome Goldenstein, *Senior Member, IEEE*

***Abstract*—We propose a new method to improve median plane sound localization performance using a nonlinear representation of head-related transfer functions (HRTFs) and a recommender system. First, we reduce the dimensionality of an HRTF dataset with multiple subjects using manifold learning in conjunction with a customized intersubject graph (ISG) which takes into account relevant prior knowledge of HRTFs. Then, we use a sound localization model to estimate a subject's localization performance in terms of polar error (PE) and quadrant error rate (QE). These metrics are merged to form a single rating per HRTF pair that we feed into a recommender system. Finally, the recommender system takes the low-dimensional HRTF representation as well as the ratings obtained from the localization model to predict the best HRTF set, possibly constructed by mixing HRTFs from different individuals, that minimizes a subject's localization error. The simulation results show that our method is capable of choosing a set of HRTFs that improves the median plane localization performance with respect to the mean localization performance using non-individualized HRTFs. Moreover, the localization performance achieved by our HRTF recommender system shows no significant difference to the localization performance observed with the best matching non-individualized HRTFs but with the advantage of not having to perform listening tests with all individuals' HRTFs from the database.**

***Index Terms*—Spatial Audio, HRTF, Manifold Learning, Recommender Systems**

I. Introduction

As augmented reality applications become more relevant, there is an increasing effort in 3D audio research and specifically in head-related transfer functions (HRTFs) to obtain high quality spatial audio. HRTFs are the main component of binaurally rendered 3D audio and are used to simulate sound sources as if they were coming from arbitrary positions in space [1]. HRTFs are complex-valued frequency functions that model the relationships between human anatomy and a sound source before reaching the ears. These functions ideally should be measured for each subject individually to avoid poor localization performance due to mismatch of spatial cues contained in the HRTFs [2].

However, HRTF measurement [3], [4] is a complex procedure that requires an expensive apparatus (e.g. a (semi-)anechoic chamber, in-ear microphones, and a loudspeaker array). Moreover, it is usually time-consuming for high-spatial resolutions and tiring for the participants.

In order to avoid such measurements, several alternatives have been proposed, including theoretical [5], numerical [6], and inference methods [7], [8]. In contrast to the above mentioned physically-based techniques, in perceptual-based techniques the subjects have an active role during the personalization process by tuning some parameters (e.g. PCA weights [9]) for several target directions until they achieve an acceptable spatial accuracy. However, this procedure might also be time-consuming, depending on the ability of the human listener and the number of parameters and target directions. An alternative approach is to optimize these parameters through the use of a machine learning algorithm where the listener is required to localize a sound source with [10] or without [11] knowledge of the target directions. There are also database matching techniques [12] where the listener selects the best HRTFs among a set of HRTFs from other subjects. Although there is no need to tune any parameter, these methods still require the listener to perform listening tests. In order to speed up these techniques, it is desirable to find a way to reduce the number of listening trials while still minimizing the localization error.

In the light of facts exposed above, we propose the use of a recommender system to find the best HRTF set, with HRTFs for each direction selected from a larger HRTF dataset constituted by HRTFs from multiple subjects, in order to improve the listener's localization performance in the median plane. Our system recommends the best mixed HRTF set by estimating a subject's localization performance through a human sound-source localization model [13]. Moreover, with a small number of listening tests, our HRTF recommender system achieves a performance statistically similar to the best performance with non-individualized HRTFs but without having to perform listening trials for every individuals HRTF in the database.

Inspired by our previous works [7], [8], [14], the recommender's input feature vectors are low-dimensional HRTFs

Manuscript received March 12, 2018; revised April xx, 2018; accepted April 16, 2017. Date of publication May xx, 2018; date of current version May xx, 2018. This work was supported by São Paulo Research Foundation (FAPESP) under Grant 2012/50468-6, Grant 2013/21349-1 and Grant 2014/14630-9, National Counsel of Technological and Scientific Development (CNPq) under Grant 308882/2013-0 and Grant 454082/2014-2, and CAPES. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pasquale De Meo.

F. Grijalva, L. Martini and B. Masiero are with the School of Electrical and Computer Engineering, University of Campinas, Campinas, SP, Brazil, 13083-852 (email: {felipe84, martini}@decom.fee.unicamp.br, masiero@unicamp.br).

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

that we obtain using manifold learning in conjunction with a customized intersubject graph (ISG) aiming to capture relevant prior knowledge of HRTFs. The outputs of our recommender system are the ratings that we obtain through the sound-source localization model proposed by Baumgartner et al. [13] (henceforth called the Baumgartner model).

Note that our approach is not an individualization technique such as [10], [11]. Moreover, different from [10], in our approach the listener is not aware of the target direction as in [11]. We also use a human sound-source localization model [13] which is more suitable than the regression model used by [11] since it takes into account psychoacoustic factors.

The remaining of the manuscript is organized as follows. We describe our recommender system in Section II. We present the conditions of our simulations in Section III and we analyze the results in Section IV. Finally, we conclude in Section V.

II. Recommender System

Recommender systems are widely used to predict the preference that a user would give to an item (e.g. books, movies) [15]. Here, we are specifically interested in content-based recommender systems (see Figure 1) where a feature vector is available for each item (HRTF) and each rating made by the users (localization accuracy). In Section II-A, we describe how we obtain such feature vectors using a nonlinear representation of HRTFs. Next, in Section II-B, we show how the ratings were calculated through a sound-source localization model. Finally, in Section II-C, we describe mathematically the problem of content-based recommender systems and how spatial audio fits into it.

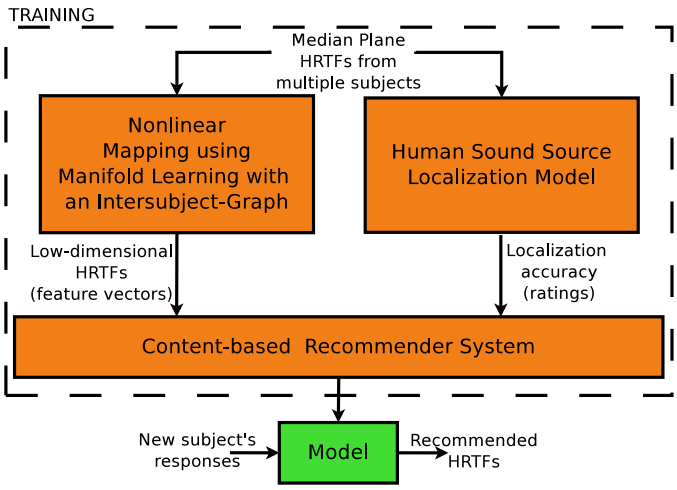

Fig. 1: We construct a model by training a recommender system using low-dimensional median plane HRTFs as feature vectors and localization accuracy as ratings, obtained through a nonlinear mapping and a human sound localization model, respectively. For a new subject's responses, our system recommends the best mixed HRTF set constituted by HRTFs from multiple subjects.

A. Nonlinear representation of HRTFs

Nonlinear dimensionality reduction techniques (i.e. manifold learning) reduce a high-dimensional dataset $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathbb{R}^D$ represented by a $D \times N$ matrix of N sample vectors $\mathbf{x}_i$ into a low-dimensional embedding $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \subset \mathbb{R}^d$ represented by a $d \times N$ matrix of N sample vectors $\mathbf{y}_i$, where $d < D$. Here, a datapoint $\mathbf{x}_i$ is the vector resulting from the concatenation of the left and right Directional Transfer Function (DTF) magnitudes, and $\mathbf{y}_i$ are the feature vectors used in the recommender system. A DTF is the component of an HRTF that is specific to sound source localization. It is obtained by dividing an HRTF by its direction-independent common component (i.e. the component including spectral features such as the ear canal resonance and microphone response [16]), which in turn is calculated by averaging all HRTFs from a specific individual [17].

A well-known manifold learning technique is Isomap [18] which attempts to preserve the pairwise geodesic distance (i.e. the distance over the manifold) in order to maintain the intrinsic geometry of the data unlike PCA that retains most variance and attempts to preserve pairwise Euclidean distances. For example, in nonlinear manifolds such as in the Swiss Roll dataset [19], PCA might map two datapoints as near points (measured by the Euclidean distance), while their geodesic distance is much larger.

Isomap has three steps. First, it takes into account the datapoint neighborhood relationships by constructing a graph $G(V, E)$ from $\mathbf{X}$, where each sample $\mathbf{x}_i \in \mathbf{X}$ represents a node $v_i \in V$. 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 neighbors 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 $\mathbf{x}_j$ (e.g. Euclidean distance). Then, we estimate the geodesic distances on the manifold between each pair of points in $\mathbf{X}$ by computing the shortest path between each corresponding pair of nodes in G. We store these distances in the pairwise geodesic distance matrix $\mathbf{D}_G$. Finally, we construct the d-dimensional embedding by applying multidimensional scaling [20] (MDS) on $\mathbf{D}_G$ to find the d-dimensional coordinate vectors $\mathbf{y}_i$.

Since neighborhood selection presents an opportunity to incorporate prior knowledge [21], instead of using common approaches (e.g. K nearest neighbors) and inspired by our previous works on HRTF personalization [7], [8] and interpolation [14], we construct the graph G by exploiting the correlations among the HRTFs across directions and subjects (we named it the Intersubject Graph, ISG), according to the following criteria:

Criterion 1: if $\mathbf{x}_i$ and $\mathbf{x}_j$ represent datapoints of the same location but different subject, then connect them. Instead of applying Isomap separately for each subject as in [22], 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.

Criterion 2: Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be datapoints of the same subject. If $\mathbf{x}_j$ is one of the k_s datapoints spatially closest to $\mathbf{x}_i$, then

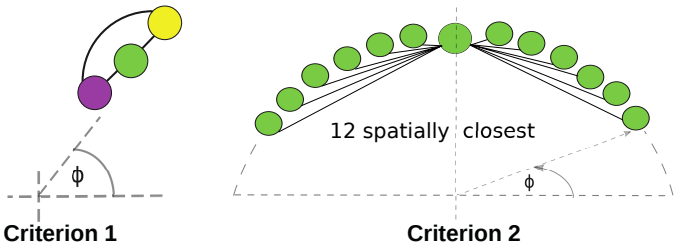

Fig. 2: Illustrative example of the ISG for $P = 3$ subjects and $k_s = 12$. Color represents same subject HRTFs, and ϕ represents elevation.

connect them. The k_s neighbors obtained from this criterion emphasize the similarities between spatially close HRTFs of the same subject.

It is straightforward to prove that the ISG is always connected (see Fig. 2 for an illustrative example). Isomap takes as parameters the number of neighbors, K, and the intrinsic dimensionality, d. Due to our ISG, the number of neighbors is fixed to $K = P + k_s - 1$, i.e., $P - 1$ from Criterion 1 and k_s from Criterion 2. To estimate the intrinsic dimensionality we use the maximum likelihood intrinsic dimensionality estimator [23]. This estimator has been previously employed in other manifold learning problems [19], [24] and tries to reveal the intrinsic geometric structure of the observed data.

Note that there are other manifold learning methods that could be used with our ISG procedure. For example Laplacian Eigenmaps [25] which is similar to Isomap in that both construct a graph representation of the datapoints. In contrast to Isomap, Laplacian Eigenmaps attempts to preserve only local properties of the manifolds based on the pairwise distances between near neighbors [19].

B. Ratings from localization model

We use the model for sound-source localization in sagittal planes proposed by Baumgartner et al. [13]. Although the model is applicable to several sagittal planes within the lateral range $\pm 30^\circ$, we only focus on the median plane responses. In this model, it is possible to predict the listener performance in terms of localization error, which, in turn, can be interpreted as the rating a subject would give to certain HRTF, i.e., the localization accuracy obtained with that HRTF. Specifically, we use the model to predict the localization error of listening through non-individualized HRTFs in the median plane. Therefore, we run a series of virtual psychoacoustic experiments to measure a subject localization performance using others' instead of their own ears [13].

The model, that requires a listener-specific calibration, is based on the comparison of an internal sound representation with a template obtained from human listeners' HRTFs. Since it returns a probabilistic prediction of a polar angle response, we are able to predict the localization performance through local polar error (PE) and quadrant error rate (QE). For both, we follow Middlebrooks [17] and define the PE as the RMS average of polar errors that were less than 90° in magnitude, and the QE as polar errors expressed in percentage form that were larger than 90° .

We normalize the QE and PE to $[0, 1]$ interval, where 1 represents the lowest error (i.e. better localization performance). In order to obtain a single rating to use on the recommender system, we calculate the rating $z^{(i,j)}$ by subject j using HRTF i as the minimum between the normalized PE and QE. We decided to use the minimum because if one of the normalized metrics is low, the overall localization performance is degraded, which in turn means that the corresponding low-rated HRTF is not suitable for the listener.

C. Content-based recommender system

In content-based recommender systems, we have a d-dimensional feature vector $\mathbf{y}^{(i)} \in \mathbb{R}^{d+1}$ (i.e. including the intercept or bias term) for each item i (e.g. movie, book). We also have a set of ratings on certain scale (e.g. 5-star rating scale) given by user j over a part of the i items we want to recommend.

The goal is to predict user j ratings using a separate linear regression model per user $\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)}$, where $\boldsymbol{\theta}^{(j)}$ is a parameter vector for user j. More formally, we want to learn $\boldsymbol{\theta}^{(j)}$ by minimizing the following linear regression problem per user

$$J = \min_{\boldsymbol{\theta}^{(j)}} \frac{1}{2} \sum_{i:r(i,j)=1} \left(\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)} - z^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{k=1}^{d} {\theta_k}^{(j)}, \tag{1}$$

where $r\,(i, j) = 1$ if user j has rated item i (0 otherwise), $z^{(i,j)}$ is the rating by user j on item i (if defined), and ${\theta_k}^{(j)}$ is the k-th parameter from the parameter vector $\boldsymbol{\theta}^{(j)}$. The last term from Eq. 1 is an L1 regularizer which reduce overfitting and encourage sparsity.

Since we are interested in more than one user, we can reformulate Eq. 1 to include n_u users as follows

$$J_m = \min_{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left(\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)} - z^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^{d} {\theta_k}^{(j)} \tag{2}$$

The cost function J_m from Eq. 2 can be optimized by means of Gradient Descent or similar algorithms. Prior to the optimization, we perform a mean normalization operation on the ratings by subtracting the corresponding average item rating μ_i. Hence, to make predictions of ratings, we need to add back the corresponding mean, i.e.,

$$\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)} + \mu_i \tag{3}$$

In the context of binaural spatial audio, the items i that we want to recommend are HRTFs. The feature vector $\mathbf{y}^{(i)} \in \mathbb{R}^{d+1}$ is the low-dimensional HRTF representation as explained in Section II-A. The rating $z^{(i,j)}$ by subject j on HRTF i is the predicted localization performance obtained by the localization model as described in Section II-B.

Since in practice it is unfeasible that a user has rated all the HRTFs in a database, we randomly selected only

a limited number of ratings per direction from the the test subject's ratings. In real psychoacoustic experiments, it would be equivalent to a user that performs a limited number of trials per direction. Note that 0 trials per direction means that the listener has not rated any HRTF. In this case, the recommender system just returns the average rating for each HRTF as stated in Eq. 3. Finally, our method recommends the HRTFs with the highest predicted rating (i.e. lowest error) per direction. Note that this implies that our method might recommend HRTFs from different subjects, to be combined to form a new HRTF set.

III. Simulations

Database and Localization model: Since the Baumgartner localization model (implemented in the Auditory Modeling Toolbox [26] as `baumgartner2014`) requires a listener-specific calibration [13], the model is only available for 23 subjects, which are included in the 97 subjects from the ARI database[1]. We used the localization model to calculate the ratings for the 23 subjects. So, each of the 23 subjects has ratings for all HRTFs from the whole ARI database, including ratings for the listener's own HRTFs. We only selected 44 directions corresponding to median plane HRTFs.

Pre-processing and Dimensionality reduction: We filtered the HRTFs to preserve frequencies between 200 Hz and 18 kHz and calculated the DTFs. We then concatenated the left and right ear DTFs into a single feature vector per direction, as explained in Section II-A. Finally, we reduced the dimensionality of the z-score scaled feature vectors (i.e. normalized to have zero mean and unit variance). These low-dimensional vectors serve as HRTF feature vectors for the recommender system. We compared several linear and nonlinear methods implemented in the *Matlab Dimensionality Reduction* Toolbox [19]. With respect to linear methods, we used PCA with 95% of variance retained. For nonlinear methods, we implemented Isomap and Laplacian Eigenmaps with our ISG (labeled as Isomap-ISG and LEM-ISG) and without it (labeled as Isomap and LEM). For all nonlinear methods, the maximum likelihood estimator [23] established the intrinsic dimensionality to $d = 13$. With respect to the ISG parameter k_s, it should be chosen according to the spatial resolution (5° for the ARI database in almost the entire median plane) and the localization blur in the median plane which varies from $\pm 9^\circ$ to $\pm 22^\circ$ [27]. For instance, we chose $k_s = 12$ since the 12 spatially closest sampling points in ARI database cover a $\pm 30^\circ$ region, i.e, it covers the entire region of the maximum localization blur in the median plane at the 5° spatial resolution of the ARI database.

Recommender system predictions: We used a leave-one-out cross validation scheme [28] to test the performance of our method. To do so, We use the localization model of 23 listeners obtained from the ARI database. For each test subject we select its HRTF feature vectors and train our model with the HRTF feature vectors from the remaining 22 subjects. This is done independently for each one of the 23 subjects and the results are latter combined. The recommender was constructed using the ratings from all these 23 subjects on the HRTF feature vectors of the remaining $P = 96$ subjects (i.e. including the test subject's ratings but excluding its HRTF feature vectors). Since in practice a subject performs a limited number of psychoacoustic experiments per direction, we randomly selected only 0 to 8 ratings per direction from the the test subject's ratings. Finally, the regularization term λ was selected using a grid search.

[1] http://www.kfs.oeaw.ac.at/

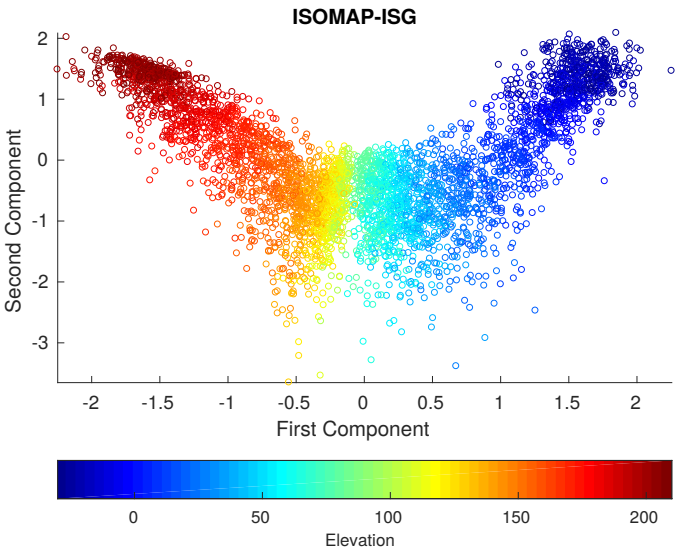

Fig. 3: Two-dimensional manifold recovered with Isomap using our ISG. All components are normalized to have zero mean and unit variance. Color represents elevation.

Metrics: Once we have the HRTFs with the highest predicted rating per direction, we can evaluate them for a specific listener using its Baumgartner model to estimate the listener's localization performance in terms of QE and PE.

IV. Results and Discussion

Figure 3 shows the two-dimensional manifold (i.e. first embedded dimension vs second one) recovered with Isomap using our ISG. Observe that there is a strong correlation (the correlation coefficient is 0.98) between the first component and the elevation angle. A similar correlation coefficient (0.99) is found for Laplacian Eigenmaps with our ISG whereas for PCA the correlation coefficient is much lower (0.82).

Before analyzing the localization performance using our recommender, we first analyze the localization performance without it. For example, Fig. 4 shows the performance predicted by the Baumgartner model for subject NH12 when using different HRTF sets from ARI database. As expected, the best performance is obtained using the subject's own HRTFs. Observe also that the best performance with someone else's HRTFs is attained using the HRTFs of subject NH93. Ideally, we expect that our recommender system achieves this performance (i.e. the best performance with others' HRTFs) which is better than the mean performance with others' HRTFs (doted line in Fig. 4). It is worth mentioning that for subject NH12 to achieve the best performance with others' HRTFs (i.e. with NH93's HRTFs) without our recommender, the listener

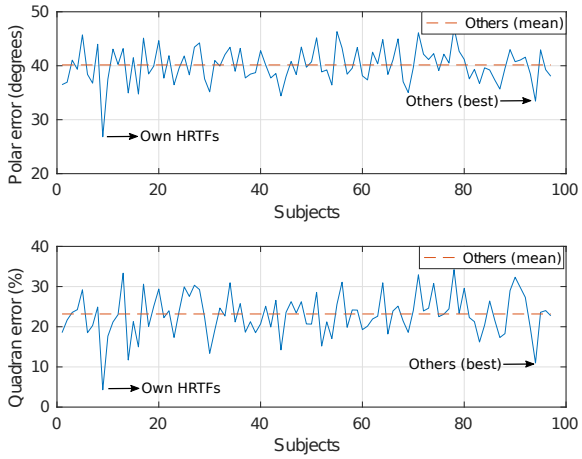

Fig. 4: Localization performance for subject NH12 when using different HRTFs sets from ARI database.

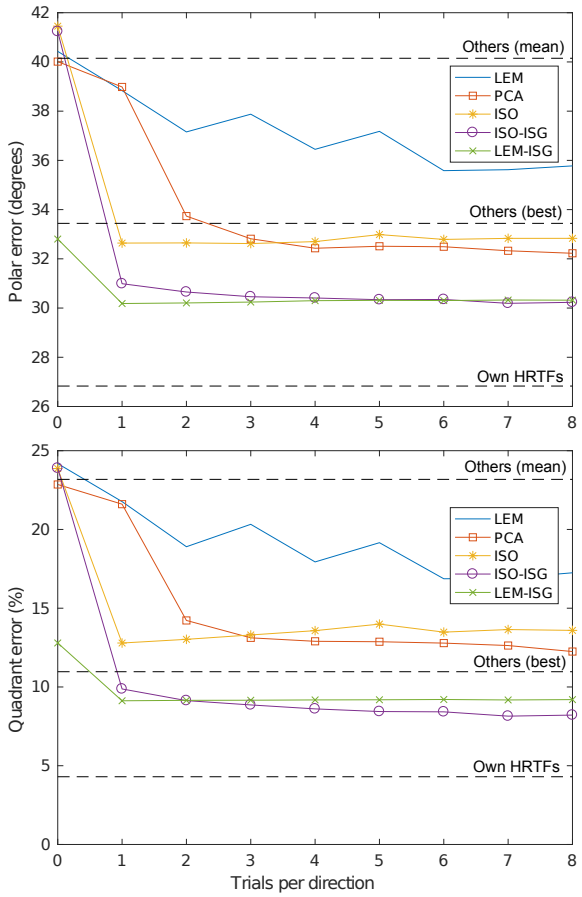

Fig. 5: Localization performance using our recommender as a function of the number of trials per direction for subject NH12.

should perform listening tests with the HRTFs from all 96 subjects, which is unfeasible in practice. For instance, to find the best performance with others' HRTFs by carrying out an exhaustive search with three trials for each of the 44 median plane directions on every ARI's subject, a listener should perform $44 \times 3 \times 96 = 12672$ listening tests.

In contrast, Fig. 5 presents the localization error using our recommender as a function of the number of trials per direction for subject NH12. Note that ISO-ISG and LEM-ISG outperform the other methods even with only one trial per direction, reaching a better performance than the mean and best performance with others' HRTFs. Although there is some minor improvement when increasing the number of trials beyond three for both ISG conditions, the largest improvement occurs during the first three trials per direction.

Since the ISG conditions have outperformed the others, the remaining analysis will focus on LEM-ISG with three trials per direction. In Fig. 6, for LEM-ISG, we show the localization performance relative to the listener-specific performance with its own HRTFs (i.e. the PE and QE variation). For example, the 8° PE variation for NH16 means that the recommended HRTFs provide localization performance that is 8° worse than its own HRTFs. In general, the proposed method has a tendency to reduce the localization error with respect to the mean performance with others' HRTFs and in many cases the error reduction is better than that achieved with the best HRTFs from others. Note also that there are a few negative variations. For instance, the negative PE variation for NH39 means that the recommended HRTFs provide localization performance that is roughly 1° better than its own HRTFs. On the other hand, for NH42 the performance using our recommender is worst when compared to the other subjects. This might be due to the fact that even the best performance variation with others' HRTFs is relatively high with respect to the other individuals.

Finally, Fig. 7 shows the localization error for LEM-ISG averaged across all subjects. The bars represent 95% confidence intervals. Paired t-tests confirm that the recommended HRTFs reduce the PE and QE errors with respect to the mean performance with others' HRTFs. Moreover, there is no statistical significance between the performance with the recommended and the best performance with others' HRTFs, which confirms that our recommender system actually improves the localization performance without having to subject the user to perform listening tests on HRTFs from all other subjects on the ARI database aiming at finding the best performance with HRTFs from someone else. For instance, to achieve a similar performance to the best performance with others' HRTFs, our recommender would only need 44 directions $\times$ 3 trials/direction $=$ 132 listening tests, in contrast to the 12672 required by an exhaustive search. On the other hand, the performance with the subject's own HRTFs is still better than the performance with the recommended HRTFs.

V. Conclusion

We show that although the performance with the subject's own HRTFs is still better than the performance with the recommended HRTF set constructed by combining HRTFs from different individuals, our HRTF recommender can actually reduce the localization error with respect to the mean performance with others' HRTFs. Moreover, our technique achieves a performance statistically similar to the best performance with others' HRTFs but with the advantage of not having to perform long and tiring listening test on multiple subjects'

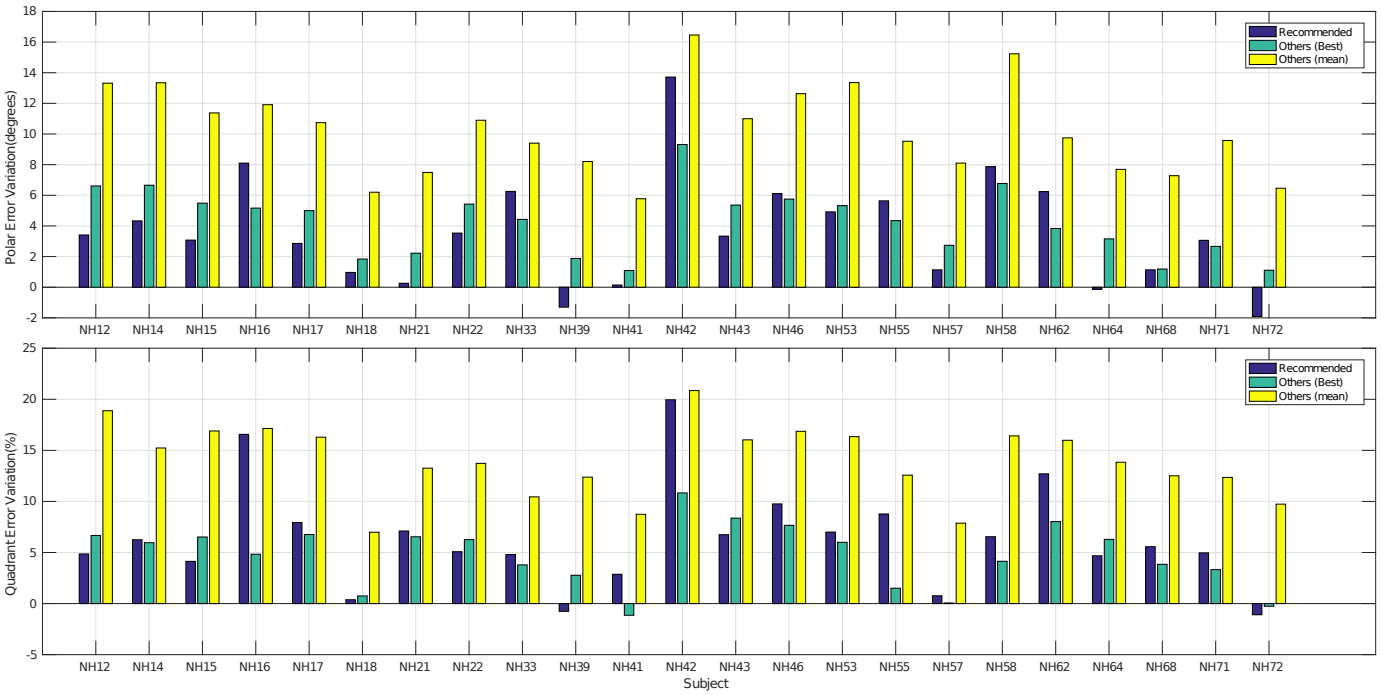

Fig. 6: Localization performance relative to the listener-specific performance with its own HRTFs (i.e. PE and QE variation) for three trials per direction using LEM-ISG.

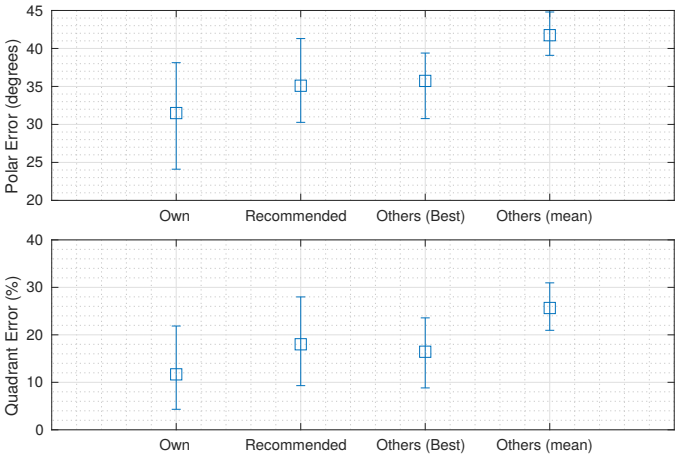

Fig. 7: Localization performance averaged across subjects for three trials per direction using LEM-ISG. Bars represent 95% confidence intervals.

datasets looking for the best performance with HRTFs from other listener. We also demonstrate that our ISG on manifold learning techniques such as Isomap and LEM can reduce the error with a small number of trials, outperforming PCA, Isomap and LEM.

Although three trials per direction seems to be too much, note that in practice the number of directions can be reduced if the recommender system is used in conjunction with some interpolation technique [14], [16]. For instance, if the listener performs three trials per direction every 20° instead of every 5°, the number of total trials would reduce drastically. Then, an interpolation method might be used to increase the spatial resolution.

Future works might try different criteria to construct the manifold. For example, instead of taking only neighbors from the same location in Criterion 1, we can select more HRTFs from the vicinity in sagittal planes adjacent to the median plane since it might occur that two subjects are not perfectly aligned during measurement. Furthermore, in a future work, it would be interesting to use more complex recommender algorithms such as [29] to try to obtain a larger improvement when increasing the number of trials beyond three.

REFERENCES

[1] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. Plantation, FL, USA.: J Ross, 2013.

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

[3] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 55, no. 7/8, pp. 623–637, 2007.

[4] M. Pollow, B. Masiero, P. Dietrich, J. Fels, and M. Vorländer, "Fast measurement system for spatially continuous individual HRTFs," in *Audio Engineering Society Conference: UK 25th Conference: Spatial Audio in Today's 3D World*. Audio Engineering Society, 2012.

[5] M. Geronazzo, S. Spagnol, and F. Avanzini, "Mixed structural modeling of head-related transfer functions for customized binaural audio delivery," in *Digital Signal Processing (DSP), 2013 18th International Conference On*. IEEE, 2013, pp. 1–8.

[6] M. Otani and S. Ise, "Fast calculation system specialized for head-related transfer function based on boundary element method," *J. Acoust. Soc. Am.*, vol. 119, no. 5, pp. 2589–2598, 2006.

[7] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features." *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 559–570, 2016.

[8] F. Grijalva, L. C. Martini, S. Goldenstein, and D. Florencio, "Anthropometric-based customization of head-related transfer functions using Isomap in the horizontal plane," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2014, pp. 4473–4477.

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

[10] K. Yamamoto and T. Igarashi, "Fully Perceptual-Based 3D Spatial Sound Individualization with an Adaptive Variational AutoEncoder," *ACM Trans Graph*, 2017.

[11] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Virtual autoencoder based recommendation system for individualizing head-related transfer functions," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[12] B. F. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. EL99–EL105, 2012.

[13] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 791–802, Aug. 2014.

[14] F. Grijalva, L. C. Martini, D. Florencio, and S. Goldenstein, "Interpolation of Head-Related Transfer Functions Using Manifold Learning," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 221–225, Feb. 2017.

[15] F. Ricci, L. Rokach, and B. Shapira, "Introduction to Recommender Systems Handbook," in *Recommender Systems Handbook*. Springer, Boston, MA, 2011, pp. 1–35.

[16] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurementsa)," *J. Acoust. Soc. Am.*, vol. 132, no. 1, pp. 282–294, 2012.

[17] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1493–1510, Aug. 1999.

[18] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[19] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative," *J Mach Learn Res*, vol. 10, pp. 66–71, 2009.

[20] W. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[21] L. K. Saul and S. T. Roweis, "Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds," *J Mach Learn Res*, vol. 4, pp. 119–155, Dec. 2003.

[22] R. Duraiswami and V. C. Raykar, "The manifolds of spatial hearing," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 3. IEEE, 2005, pp. iii–285.

[23] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in Neural Information Processing Systems*, 2004, pp. 777–784.

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

[25] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering." in *NIPS*, vol. 14, 2001, pp. 585–591.

[26] P. L. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening*, ser. Modern Acoustics and Signal Processing. Springer, Berlin, Heidelberg, 2013, pp. 33–56.

[27] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[28] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[29] P. Chiliguano and G. Fazekas, "Hybrid music recommender using content-based and social information," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2618–2622.