Dimensionality Reduced HRTFs: A Comparative Study

Bill Kapralos
Faculty of Business and Information Technology, Health Education technology Research Unit, University of Ontario Institute of Technology. Oshawa, Ontario, Canada. L1H 7K4.
bill.kapralos@uoit.ca

Agnieszka Kopinska
Department of Psychology, York University. Toronto, Ontario, Canada. M3J 1P3.
kopinska@yorku.ca

Nathan Mekuz
mekuz@cse.yorku.ca

Saad Khattak
Business and Information Technology, University of Ontario Institute of Technology. Oshawa, Ontario, Canada. L1H 7K4.
saad.khattak@myspace.uoit.ca

ABSTRACT
Dimensionality reduction is a statistical tool commonly used to map high-dimensional data into lower a dimensionality. The transformed data is typically more suitable for regression analysis or classification than the original data. Being of high dimensionality, HRTF data is commonly reduced using Principal Components Analysis (PCA). While highly effective at compressing data that follows the assumed model, PCA compression performance suffers when data follows a nonlinear distribution or when outliers are present. More recent data reduction techniques such as Isomap and locally linear embedding (LLE) take advantage of local neighborhood information in order to learn a more suitable basis for the HRTF data at hand. Quantitative results from previous work indicate that the embeddings created by both LLE and Isomap are superior to those obtained with PCA. This paper presents a study that compares sound source localization accuracy by human observers when presented with a virtual sound synthesized using HRTFs whose dimensionality was reduced using either Isomap, LLE or PCA. Preliminary results indicate that good sound source localization judgement is obtainable using dimensionality reduced HRTFs and that Isomap and LLE produce superior results thus, confirming previous quantitative results.

Categories and Subject Descriptors
I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality

General Terms
Human Factors

Keywords
HRTF, 3D sound, dimensionality reduction, Principal Component Analysis (PCA), Locally Linear Embedding (LLE), Isomap.

1. INTRODUCTION
When incorporated into a virtual environment, spatial (3D) sound cues can add a better sense of “presence” or “immersion”, they can compensate for poor visual cues (graphics), lead to improved object localization and, at the very least, add a “pleasing quality” to the simulation [2, 14]. The foundation of spatial audio rests on the ability to control the auditory signals arriving at the listener’s ears such that these signals are perceptually equivalent to the signals that the listener would receive in the environment being simulated [16]. Collectively, the filtering effects of a sound by the listener’s head, shoulders, upper torso, and most notably the pinna, are modeled by a complex response function known as the head-related transfer function (HRTF) or the anatomical transfer function (ATF) [8]. HRTFs encompass various sound localization cues including interaural time differences (ITDs), interaural level differences (ILDs), and the changes in the spectral shape of the sound reaching a listener. The HRTF modifies the spectrum and timing of sound signals reaching each ear in a location-dependent manner [3]. The left \( H_L(\omega, \theta, \phi, d) \) and right \( H_R(\omega, \theta, \phi, d) \) ear HRTFs are functions of four variables: \( \omega \) is the angular frequency of the sound source, \( \theta \) and \( \phi \) are the sound source azimuth and elevation angles respectively, and \( d \) is the distance from the listener to the sound source (measured from the center of the listener’s head) [17].

A common technique used to measure an individual’s left and right ear HRTFs for a sound source at a position \( \vec{p} \) relative to the listener is to output an excitation signal \( s(t) \) with known spectral characteristics from a loudspeaker placed at position \( \vec{p} \), and to measure the resulting impulse response at the left and right ears, \( h_L \) and \( h_R \) respectively, using small probe microphones inserted into the individual’s left and right ear canals [3]. The responses \( h_L \) and \( h_R \) as measured at each ear in the time domain. The time domain representation of the HRTF is known as the head-related impulse response (HRIR). Applying the discrete Fourier transform (DFT) to the time domain impulse responses \( h_L \) and \( h_R \) results in the left \( H_L(\omega, \theta, \phi, d) \) and right \( H_R(\omega, \theta, \phi, d) \) ear HRTFs.

The measured response forms the basis of a filter that can
be used to modulate source sound material (e.g., anechoic or synthesized sound) via a convolution operation. When the filtered sounds are presented to the listener, they create the impression of a sound source located at the corresponding HRTF measurement position. Convolution is unfortunately an extremely computationally expensive technique that greatly limits the performance of any real-time virtual audio system. In addition, the HRTF filters are typically high dimensional, containing as many as 512 coefficients. Further complicating matters, when the modeling system accounts for room acoustics (e.g., reflected sound), relative to the receiver, each reflection has its own direction of arrival. Given the direction (position) dependent filtering of the HRTF, each reflection should be filtered with an HRTF pair corresponding to that particular direction of reflection in order to recreate “binaural listening” [3]. This HRTF filtering accounts for most of the computational complexity of a spatial sound system and is clearly impractical for any interactive (real-time) virtual audio system [7] (see [9] for an overview of virtual audio).

Dimensionality reduction techniques are commonly used to map high-dimensional data such as images and speech signals into lower dimensionality. The transformed data is typically more suitable for regression analysis or classification than the original data. The underlying assumption is that observed high-dimensional samples lie on or near a lower-dimensional manifold embedded within the original high-dimensional space. The purpose of the reduction is to project the high-dimensional data into a more compact representation while preserving certain properties of the data. Depending on the application, compression of the HRTF is possible. Common methods used to reduce the dimensionality of HRTFs include principal components analysis (PCA) [11] and selection of a basis of HRTF spectra using genetic algorithms [4]. By eliminating the features of lesser importance, the size of an HRTF filter is greatly reduced thus reducing the computational time required for an HRTF filtering operation. In addition to compression, reducing data to a more compact representation can potentially lead to better generalization. PCA is commonly used to summarize data with an assumed multivariate normal distribution for which PCA is optimal. However, for data that does not conform to this model (e.g., nonlinear data, or data with a significant number of outliers with respect to the Gaussian model) another, more recently introduced locally linear dimensionality reduction techniques, such as locally linear embedding (LLE) [12], and Isomap [15] have been demonstrated to be more effective. PCA compression performance suffers when the data follows a nonlinear distribution or when outliers are present. Several preliminary studies have suggested that LLE, and Isomap could overcome the problems and limitations associated with PCA techniques [12, 15].

This paper presents a comparison of sound source localization accuracy by human observers when presented with a virtual sound synthesized using HRTFs that have been reduced in dimensionality using Isomap, LLE and PCA. Results indicate that good sound source localization judgment is obtainable using dimensionality reduced HRTFs and Isomap and LLE perform better than PCA on HRTF data.

1.1 Paper Organization

The remainder of this paper is organized as follows. Section 2 provides background information regarding the LLE, Isomap, and PCA techniques. The experimental procedure is described in Section 3 while experimental results and a discussion of the results is provided in Section 4. Finally, concluding remarks and plans for future research are presented in Section 5.

2. BACKGROUND AND PREVIOUS WORK

The PCA model assumes that the data lies on a linear subspace (a hyperplane) of the input space, contaminated by Gaussian noise. PCA computes a maximum likelihood estimate of this subspace by maximizing the variance in that subspace (i.e., minimizing the variance of error). It seeks to maximize the retained variance, implicitly making a Gaussian assumption on the data through the least-squares formulation. As previously described, PCA compression performance suffers when data follows a nonlinear distribution or when outliers are present. In practice, this is usually overcome by increasing the dimensionality or the number of retained components. For instance, consider 2D data of the form \([x, x^2]\). This data is intrinsically one dimensional as it can be explained by only one latent variable \(x\). However, linear correlation alone (in the statistical sense) is insufficient to fully explain the relationship between the observed dimensions and PCA will fail to learn a subspace that characterizes this data well at any target dimensionality.

A recent surge in interest in locally linear manifold learning techniques has resulted in the introduction of several new techniques including LLE and Isomap. LLE and Isomap take advantage of local neighborhood information in order to learn a more suitable nonlinear basis for the HRTF data at hand. These techniques view the manifold as a patchwork of connected linear surfaces, and seek to preserve certain properties in the projection. If the manifold is continuous and sufficiently well sampled, then using Taylor’s theorem, small patches can be approximated as linear. If parts of the manifold are linear, globally nonlinear methods may be overly complex and difficult to train due to the large number of parameters. On the other hand, locally linear techniques may model the manifold effectively by fitting parts of it separately if they are able to decompose it into linear components. However, local modeling is sensitive to noise and the modeling of noise remains a challenge. Consequently, most locally linear techniques do not address the issue of noise. LLE introduced by Roweis and Saul [12], is a locally linear dimensionality reduction technique that identifies local neighborhood distance relationships and computes a lower dimensional mapping that preserves them. At each input point, a set of coefficients (weights) is determined which approximate it, in the least-squares sense, from its neighbors. Then a global mapping to a low-dimensional embedding is computed which optimizes the preservation of these points. Isomap [15] is a popular locally linear technique that works by assuming isometry of geodesic distances in the manifold. The geodesic distance is defined as the distance of the shortest path between two points that passes on the embedded manifold [1]. Isomap estimates geodesic distances by constructing a graph with Euclidean distances between neighboring points as edge weights and computing shortest paths in the graph.

Seung and Lee [13] argue that nonlinear learning techniques are crucial to understanding how our perception of sound source direction arises from the dynamics of neural networks in our brain and could also lead to a low-
dimensional representation of the HRTF. Another study has demonstrated that LLE is capable of encoding perceptual information of the HRTF and proposes an HRTF interpolation method based on these findings [5]. Finally, a quantitative comparison between PCA, LLE, and Isomap applied to HRTFs has confirmed that LLE and particularly Isomap are superior to PCA, with Isomap exhibiting the largest correlation especially with respect to azimuth [10]. Despite these promising results, the application of these newer data reduction techniques to the HRTF is sparse. Furthermore, although quantitative results are promising, the ultimate user of any spatial sound system is a human and therefore, perceptual effects must be examined and accounted for.

3. EXPERIMENTAL METHOD

3.1 Participants

Ten paid participants took part in the experiment (average age 27 years; range 21 to 40 years). Participants had no reported history of auditory disease/disorders. None of the participants reported any difficulties in hearing the stimuli or in completing any of the tasks. Average time to complete the experiment was approximately 30 minutes and all participants completed the experiment in one session with a five minute break at approximately 15 minutes into the experiment.

3.2 Auditory Stimuli

The unprocessed auditory stimulus for each experiment consisted of a broadband, uniformly distributed, white-noise signal sampled at a rate of 44.1 kHz. The noise was band-pass filtered using a 256-point Hamming windowed FIR filter with low and high frequency cut-offs of 200 Hz and 10 kHz respectively. The duration of each stimulus was 250 ms and was conveyed through a pair of AKG Acoustics K240 headphones. Sound level in the absence of headphones was set to be 69 dB, measured with a Radio Shack sound level meter (model 33-2055) with an A-weighting averaged over a 15 s period.

The HRTFs were obtained from the MIT HRTF dataset [6]. This dataset of “raw” HRTFs was measured using the anthropomorphic dummy KEMAR. In total, 710 measurements were sampled, one elevation at a time, by moving the loudspeaker to some particular elevation, from $-25^\circ$ to $90^\circ$ (in $10^\circ$ increments) and rotating the KEMAR a total of $360^\circ$, in equal increments for each elevation. The increment size was chosen to maintain approximately $5^\circ$ great-circle increments [6]. Each HRTF measurement contains 512 coefficients and was sampled at a rate of 44.1 kHz.

3.3 Experimental Procedure

In each trial, participants were stationary and presented with two auditory stimuli: $s_1(t)$ and $s_2(t)$ - broadband white noise with a duration of 250 ms as described in Section 3.2 above. Both stimuli were HRTF processed with HRTFs whose elevation was $0^\circ$ (azimuth was varied). All processing was performed once off-line. Each data-reduced HRTF measurement contained 50 coefficients (a reduction of approximately 90% with respect to the original HRTFs). The azimuth angle for $s_1(t)$ was selected from $s_1(t) \in \{-25^\circ, 0^\circ, 25^\circ\}$. The azimuth angle for $s_2(t)$ was set to $s_2(t) \in \{s_1(t) - 25^\circ, s_1(t) + 25^\circ\}$.

In addition to varying HRTF location, the HRTFs were also reduced in dimensionality. As a result, the HRTFs were chosen from one of the three following subsets: data reduced HRTFs using (i) PCA, (ii) LLE, and (iii) Isomap.

In each trial, $s_1(t)$ served as a “reference” and was presented first, followed by $s_2(t)$, 250 ms later. Both stimuli were selected from the same HRTF subset as described above. In each trial, the participant’s task was to judge whether they perceived $s_2(t)$ to the left or to the right of $s_1(t)$. Participants indicated their choice of either left or right by pressing a “left” or “right” button on a video-game control pad. Pressing of either button indicated the completion of the trial. Trials were spaced five seconds apart. Each setting of $[s_1(t), s_2(t)]^T$ was presented 20 times, for a total of 360 trials. The trials were chosen in random order.

4. EXPERIMENTAL RESULTS

In this section, the suitability of the Isomap, LLE, and PCA techniques to the reduction of HRTF data is evaluated. This is accomplished by comparing sound source localization accuracy by human observers who were presented with a virtual sound synthesized with each of the three sets of dimensionality reduced HRTFs. In each trial, the participant’s choice was either correct or incorrect. Here, results are presented in the form of the number of trials for each reference angle and for each HRTF data reduction category, which were incorrect averaged across all participants.

A graphical summary of the results for PCA, LLE, and Isomap is provided in Figure 1, where, for each of the three reference angles, the number of incorrect responses for each of the corresponding test stimuli (averaged across all participants), is provided. The mean number of incorrect responses is also provided.

Given the large separation between the reference and test stimuli angles ($25^\circ$), one would expect that the total number of incorrect responses to be minimal if not zero. However, this is not the case, particularly for the HRTFs whose dimensionality was reduced with PCA. The total number of incorrect responses across each of the three references angles for PCA, LLE, and Isomap are 14 (11.7%), 2 (1.7%), and 4 (3.3%) respectively. These results are in agreement with earlier quantitative results which found that the embeddings created with LLE and Isomap are superior to those created with PCA [10]. Although the difference is small, the LLE-reduced HRTFs resulted in a smaller number of errors (two in total) than those created with the Isomap-reduced HRTFs (four in total) in contrast to the previous quantitative results where the embeddings created with Isomap resulted in the largest correlation [10].

5. SUMMARY

In this paper we have presented the results of an experiment that examined the applicability of data-reduced HRTFs for sound synthesis. Preliminary results indicate that dimensionality-reduced HRTFs can be used effectively to synthesize a sound source. Furthermore, these results indicate that LLE and Isomap provide better sound localization capabilities than PCA, confirming previous quantitative results. Future work includes further, more extensive human-based experiments that will examine sound source localization with data-reduced HRTFs over a greater range of azimuth and elevations that are also more finely sampled.
Figure 1: Summary of results: number of incorrect responses obtained with the different dimensionality reduction techniques at different azimuth angles (three leftmost subplots) and the mean (right subplot).

Future work will also examine sound source localization capabilities using HRTFs that are reduced to different target dimensionalities, as well as using other dimensionality reduction techniques.

6. REFERENCES