Study and Comparison of Efficient Methods for 3D Audio Spatialization Based on Linear Decomposition of HRTF Data

Veronique Larcher (1), Jean-Marc Jot (2), J. Guyard (2) and Olivier Warusfel (1),
(1) IRCAM, 1 place Igor Stravinsky, 75004 Paris, France.
(2) Creative/E-mu Technology Center, Scotts Valley, CA 95067, USA.

Presented at
the 108th Convention
2000 February 19-22
Paris, France

AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

Study and comparison of efficient methods for 3D audio spatialization based on linear decomposition of HRTF data

Véronique Larcher
Ircam. 1 place Igor Stravinsky. Paris, F 75004 France.

Jean-Marc Jot, Jérôme Guyard
Creative Advanced Technology Center. 1600 Green Hills Rd. Scotts Valley, CA 95067, USA

Olivier Warusfel
Ircam. 1 place Igor Stravinsky. Paris, F 75004 France.

Abstract

Several computationally efficient multi-source spatialization methods for headphone reproduction are reviewed. They rely on the linear decomposition of HRTFs into Spatial Functions and Reconstruction Filters. Decomposition methods based on statistical analysis, eigenvalue decomposition, and projection techniques are presented. Specifically, methods yielding non-individualized, "discrete" Spatial Functions, while minimizing the reconstruction error, are investigated.

0- Introduction

0.1- Principles

The design of binaural synthesis encoders is traditionally based on the 2-channel filtering of a monaural input sound by a pair of HRTFs. Each HRTF is decomposed into a pure delay cascaded with a minimum phase filter, so that time differences (ITD) and spectral information are modeled and rendered independently [1]. This leads to the implementation represented in Figure 1.
In order to simulate a moving sound source, these filters must be made variable. Updating filter coefficients lead to commutation artefacts, which can be overcome by duplicating the filtering stage to cross-fade between two sets of filters. Each additional source will require 4 new minimum-phase filters, which leads to costly implementations for typical applications requiring multiple spatialized sources.
A multichannel implementation of binaural synthesis was proposed in [9]. It relies on the linear decomposition of HRTF data into a finite number of Spatial Functions, denoted $C_i(\theta)$, and associated Basis Filters, denoted $L_i(f)$. HRTFs are thereby approximated by:

$$H(\theta, f) = \sum_i C_i(\theta) \cdot L_i(f) \qquad (0)$$

A multichannel binaural encoder will consist in the weighting of a monaural input signal by the set of Spatial Functions, while the decoder is a parallel bank of Reconstruction Filters. This implementation is illustrated in Figure 2. All sources are mixed at the output of the encoders, and therefore share the same Reconstruction Filters. Hence, each additional source will only require a few gain factors, while the filtering cost will remain independent of the number of sources.

The advantages of this decomposition are:
- Reduced computational cost in the context of several simultaneous sound sources,
- Accurate spatial interpolation in real-time: Spatial Functions can be extrapolated for unmeasured positions,
- Compact representation of the HRTF data for the analysis of inter-individual variations: a number of individual Spatial Functions and Reconstruction Filters instead of hundreds of measured HRTFs.

Following the decomposition used in the 2-channel implementation of Fig. 1, we can take advantage of HRTF properties and decompose the minimum-phase HRTF according to Eq. (0), while the ITD is synthesized explicitly in the encoder for each sound source, as shown in Figure 3. Removing the delays from the data allows to reduce the number of channels required to achieve a close approximation of the original HRTF.

In the following sections, we compare several methods to derive Spatial Functions C_i and Reconstruction Filters L_i, for minimum-phase HRTF decomposition. The results are illustrated for a 7-channel decomposition of HRTFs in the horizontal plane.

0.2- Objective design criteria

The comparison between the decomposition methods will be considered according to 3 objective criteria:

- **Accuracy of the reconstruction**
  Least-squares minimization of the difference between original and reconstructed HRTFs leads to tractable mathematical resolution with linear algebra tools. As the focus of this paper is on the decomposition methods, results will be presented by referring to the error that is minimized (L_2 error). An alternative error measure, possibly more relevant from a perceptual point of view, would be the difference between the log-magnitude spectra of the original and reconstructed HRTFs.

- **Interindividual differences**
  As each head yields a different (or individual) set of HRTFs, the decomposition will a priori provide individual Spatial Functions and Reconstruction Filters. From an application standpoint, it is advantageous to exhibit "universal" Spatial Functions, so that the resulting multichannel encoding format is suitable for any listener, and the reproduction may be optimized for a given listener simply by adapting the Reconstruction Filters in the decoder [13]. From a research standpoint, this simplification of the model would also make the

investigation of interindividual differences in HRTFs more tractable, since the problem reduces to modeling the ITD and the set of Reconstruction Filters.

- **Discreteness of the Spatial Functions**
  The computational cost of the encoder is reduced if only a few non-zero gain factors are needed to encode the position of a given sound source. We will refer to "discrete" Spatial Functions when these have only three non-zero Spatial Functions for any position in three-dimensional space. Discrete panning laws used for multichannel spatialization over loudspeakers are examples of discrete Spatial Functions.

An ideal decomposition method should then provide a minimum reconstruction error, while exhibiting "universal" and "discrete" Spatial Functions.

1- Mathematical Formulation

1.1- Notations

We will use the following notations:

- $\boldsymbol{A}^\dagger$ denotes the conjugate transpose of matrix (or vector) $\boldsymbol{A}$,
- [.|.] denotes the vector dot product defined by:

$$[\boldsymbol{a}|\boldsymbol{b}] = \boldsymbol{b}^\dagger.\boldsymbol{a},$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are column vectors.

This dot product defines the vector norm, known as the L_2 norm:

$$|\boldsymbol{a}|^2 = [\boldsymbol{a}|\boldsymbol{a}]$$

By extension, we will use the notation $[\boldsymbol{A}|\boldsymbol{B}]$ for the matrix product $\boldsymbol{B}^\dagger.\boldsymbol{A}$, which stores all possible dot products between columns of matrix $\boldsymbol{A}$ and columns of matrix $\boldsymbol{B}$.

- [.‖.] denotes the matrix dot product defined by:

$$[\boldsymbol{A}\|\boldsymbol{B}] = \sum \mathrm{diag}[\boldsymbol{A}|\boldsymbol{B}] = \sum \mathrm{diag}(\boldsymbol{B}^\dagger.\boldsymbol{A}).$$

This dot product defines the matrix norm known as the Frobenius norm:

$$\|\boldsymbol{A}\|^2 = \sum \mathrm{diag}[\boldsymbol{A}|\boldsymbol{A}] = \sum \mathrm{diag}(\boldsymbol{A}^\dagger.\boldsymbol{A}) = \sum_i |a_i|^2$$

where the a_i, $i = 1...M$, are the M columns of matrix $\boldsymbol{A}$.

1.2- Application to the decomposition of HRTF data

Let $\boldsymbol{H}$ be a $N \times M$ matrix containing the complex HRTFs to be projected, for N positions in space and M frequency points. It is equivalent to describe $\boldsymbol{H}$ as a matrix of N filters, or as a matrix of M Spatial Functions. We look for an $N \times r$ matrix of Spatial Functions $\boldsymbol{C}$ and for an $r \times M$ matrix of Reconstruction Filters $\boldsymbol{L}$ so that $\boldsymbol{H}$ is approximated by:

$$\boldsymbol{H} = \boldsymbol{C}.\boldsymbol{L}$$

An efficient representation of $\boldsymbol{H}$ should involve a number r of Reconstruction Filters and Spatial Functions as small as possible. Furthermore, the Spatial Functions (resp. Reconstruction Filters) should define a set of linearly independent vectors, so that $\boldsymbol{C}$ and $\boldsymbol{L}$ are matrices of rank r. Lastly, we want the Spatial Functions to be real in order to minimize the computing cost of the encoder.

- Deriving $\boldsymbol{L}$ from $\boldsymbol{C}$ and $\boldsymbol{H}$

Knowing $\boldsymbol{C}$, we can define the matrix $\boldsymbol{L}$ minimizing the error norm $\|\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L}\|^2$. $\boldsymbol{L}$ must then be such that the columns of $\boldsymbol{C}.\boldsymbol{L}$ are the orthogonal projections of the columns of $\boldsymbol{H}$ on the space spanned by the columns of $\boldsymbol{C}$. We will demonstrate this point in two steps:

1. $\boldsymbol{L}$ is taken so that the columns of $\boldsymbol{C}.\boldsymbol{L}$ are the orthogonal projection of the columns of $\boldsymbol{H}$ over the space spanned by the columns of $\boldsymbol{C}$, i. e. the columns of the reconstruction error matrix $\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L}$ are orthogonal to the columns of C:

$$[(\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L})\,|\boldsymbol{C}] = 0$$

This equation leads to the solution:

$$\boldsymbol{L} = \boldsymbol{G}_{\boldsymbol{C}}^{-1}.[\boldsymbol{H}|\boldsymbol{C}]$$

where $\boldsymbol{G}_{\boldsymbol{C}}$ denotes the symmetrical $r \times r$ Gram Matrix related to $\boldsymbol{C}$, defined by:

$$\boldsymbol{G}_{\boldsymbol{C}} = [\boldsymbol{C}|\boldsymbol{C}]$$

As $\boldsymbol{C}$ is of rank r by definition, $\boldsymbol{G}_{\boldsymbol{C}}$ is invertible, and, in the case of orthogonal Spatial Functions, $\boldsymbol{G}_{\boldsymbol{C}}$ is a diagonal matrix.

2. The so-defined $\boldsymbol{L}$ minimizes the Frobenius norm of the reconstruction error.

Let $\boldsymbol{L}'$ be another matrix of r Reconstruction Filters. We have:

$$|\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L}'|^2 = |\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L}|^2 + |\boldsymbol{C}.(\boldsymbol{L}-\boldsymbol{L}')|^2 + [(\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L})|(\boldsymbol{C}.(\boldsymbol{L}-\boldsymbol{L}'))] + [(\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L})|(\boldsymbol{C}.(\boldsymbol{L}-\boldsymbol{L}'))]^{\dagger}$$

Since the columns of $\boldsymbol{L}$ are orthogonal to the columns of the reconstruction error, the last two terms vanish, and we can write:

$$\|\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L}'\|^2 = \|\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L}\|^2 + \|\boldsymbol{C}.(\boldsymbol{L}-\boldsymbol{L}')\|^2$$

which is minimized only by making $\boldsymbol{C}.\boldsymbol{L} = \boldsymbol{C}.\boldsymbol{L}'$. As $\boldsymbol{C}$ is of rank r, this equality is achieved if and only if $\boldsymbol{L} = \boldsymbol{L}'$.

The error norm thereby evaluated can easily be linked to the more familiar L_2 norm. We can in a first step compute the L_2 norm of the error for each frequency. This frequency-dependent error norm rates how close the reconstructed HRTFs (columns of $\boldsymbol{C}.\boldsymbol{L}$) are to the target HRTFs (columns of $\boldsymbol{H}$), in a L_2 sense. The Frobenius norm then consists of the sum over all frequencies of these L_2 errors.

- Deriving $\boldsymbol{C}$ from $\boldsymbol{L}$ and $\boldsymbol{H}$

In a similar way, $\boldsymbol{C}$ can be obtained from the orthogonal projection of the columns of $\boldsymbol{H}$ over the Reconstruction filters stored in $\boldsymbol{L}$. This time, we have:

$$\boldsymbol{C} = [\boldsymbol{L}^\dagger|\boldsymbol{H}^\dagger].\boldsymbol{G}_{\boldsymbol{L}}^{\dagger\text{-}1}$$

where $\boldsymbol{G}_{\boldsymbol{L}}^\dagger$ denotes the inverse of the $r\times r$ Gram Matrix related to $\boldsymbol{L}^\dagger$, defined by:

$$\boldsymbol{G}_{\boldsymbol{L}}^\dagger = [\boldsymbol{L}^\dagger|\boldsymbol{L}^\dagger] = [\boldsymbol{L}|\boldsymbol{L}]^\dagger$$

It can be verified that, when the filters are stored in matrices $\boldsymbol{L}$ and $\boldsymbol{H}$ as full-length discrete Fourier spectra (including the negative frequencies), the Hermitian symmetry in these spectra around the Nyquist frequency causes all the coefficients in the solution $\boldsymbol{C}$ to be real.

The matrix $\boldsymbol{C}$ thereby defined minimizes the Frobenius norm of $(\boldsymbol{H}-\boldsymbol{C}.\boldsymbol{L})^\dagger$. This time, the L_2 norm of the error has been computed for each position. This position-dependent error norm rates how close the reconstructed filters (lines of $\boldsymbol{C}.\boldsymbol{L}$) are to the target filters (lines of $\boldsymbol{H}$), in a L_2 sense. The Frobenius norm consists of the sum over all positions of these L_2 errors.
In the following we will therefore either consider the L_2-norm of the reconstruction error as a function of frequency (respectively position), which is obtained by averaging the reconstruction errors across all positions (respectively frequency), or consider the Frobenius norm of the global reconstruction error (averaged across both positions and frequencies).

1.3- Taking elevation into account

We wish to optimize the reconstruction of HRTFs in the horizontal plane, and propose an optimization scheme making use of the mechanism of orthogonal projection reviewed above.
We choose to devote a single channel to the synthesis of HRTF in elevation, while the 7 other channels optimize the reconstruction for horizontal positions. Given the seven Reconstruction Filters which optimize the reconstruction for horizontal positions, the previously described relations can be used to extrapolate the Spatial Functions associated to these "horizontal Reconstruction Filters" for elevated positions, so that they optimally contribute to the synthesis of

elevated positions, and are complemented by an "elevation Reconstruction Filter". The decomposition thereby handled takes place in 2 steps:

1. Extrapolation of the Spatial Functions for elevated positions

This step allows to make use of the Reconstruction Filters obtained by a given decomposition method for HRTFs in the horizontal plane, or "horizontal Reconstruction Filters", in order to reconstruct the HRTFs for elevated positions.

$\boldsymbol{C}$ can be obtained by: $\boldsymbol{C} = [\boldsymbol{L}^{\dagger}|\boldsymbol{H}^{\dagger}].\boldsymbol{G}_{\boldsymbol{L}}{}^{\dagger\text{-}1}$

With the decomposition methods described below, this relation is verified in the horizontal plane already, so that the extrapolation process does not alter the Spatial Functions obtained for these privileged positions. This way, the "horizontal Reconstruction Filters" are involved into the reconstruction of HRTF in elevation, and their weighting gains are optimized to minimize the Frobenius norm of the reconstruction error.

2. Design of the "elevation Reconstruction Filter"

We want to find a filter which is the optimal complement to the "horizontal Reconstruction Filters" in order to yield an accurate synthesis of HRTF in elevation, and which is not involved into the reconstruction of HRTF in the horizontal plane. We obtain this filter L_e and the associated Spatial Function C_e by decomposing the reconstruction error provided by the "horizontal Reconstruction Filters" and the extrapolated Spatial Functions.

2- Review of linear decomposition techniques

2.1. Joint optimization of Spatial Functions and Reconstruction Filters

Statistical analysis methods such as Principal Components Analysis and Independent Components Analysis yield a compact representation of a set of variables by a reduced number of new variables. This "data compression" process works by getting rid of the mutual information shared by the initial variables and concentrate the remaining information into a reduced number of statistically independent new variables. The two approaches differ in the level of independence they achieve: PCA provides variables independent up to the second order, i.e. decorrelated variables, while ICA achieves independence up to a higher-order. Choosing decorrelated variables to span our set of HRTF is of interest since it can be shown that it minimizes the L_2-norm of the reconstruction error of each variable. Our interest into higher-order independence relies in the link we wish to make with discreteness as defined in Section 0.2. The theory behind both methods has the same starting point, and is exhaustively developed in [3].

a) Formal relationship between PCA and ICA

Suppose we have a collection of M 0-mean variables $\{x_1, x_2,..., x_M\}$, observed N times. Let p_{xi} be the probability density of variable x_i, and p_x the joint probability density of N-dimensional variable $\boldsymbol{x} = [x_1\ x_2,...x_M]$. From a practical point of view, we can see each x_i variable as a column vector of length N, while variable x is a $N \times M$ matrix. Variables x_i are independent if and only if:

$$p_x(\boldsymbol{u}) = \prod_{i=1}^{M} p_{x_i}(u_i)$$

This equality can be reformulated as setting to zero the distance between the two terms. This distance between probability densities can be defined using Kullback divergence metric. It will be called "mutual information" of x, and will be denoted $I(p_x)$. Having independent variables is then equivalent to having a zero mutual information. Some simple manipulations lead to a very informative expression of $I(p_x)$:

$$I(p_x) = J(p_x) - \sum_{i=1}^{M} J(p_{x_i}) + I(\phi_x)$$

where $I(\phi_x)$ denotes the mutual information of a gaussian probability density with same mean and variances as p_x,

$J(p_x)$ denotes the negentropy of p_x,

$J(p_{xi})$ denotes the negentropy of p_{xi}.

The negentropy of a probability density is an always positive term which measures how close this density is to a Gaussian distribution.

Minimizing mutual information $I(p_x)$can then be achieved in two steps:

1. find the linear transform $\boldsymbol{A}$ such that:

$$I(\phi_{A.x})=0 \qquad (1)$$

2. find the linear transform $\boldsymbol{B}$ such that:

$$I(\phi_{B.A.x}) \text{ remains } 0 \qquad (2)$$

$$J(p_{B.A.x}) \text{ is constant} \qquad (3)$$

$$\sum_{i=1}^{N} J(p_{B.A.x_i}) \text{ is maximum} \qquad (4)$$

It can be shown that equality (1) is achieved if and only if the new set of variables $y=\boldsymbol{A}.x$ has a diagonal covariance matrix $\frac{1}{N}.[y][y]$, i.e. if variables y_i are decorrelated. This is actually what PCA

performs. We can therefore say that PCA reduces the mutual information shared by our variables up to the second order, since only second order statistics are involved in this first step.
Achieving (1) leads to find the basis in which matrix $[x|x]$, is diagonal, which is simply achieved by an eigenvalue decomposition. If k denotes the rank of $[x|x]$, we know that matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{Q}$ exist so that:

$$\frac{1}{N}.[x|x] = Q.\Sigma.Q^{\dagger}$$

with matrix $\boldsymbol{\Sigma}$ being a diagonal matrix with k non-zero elements, all positive,
matrix $\boldsymbol{Q}$ being an orthogonal matrix.

If we choose: $$\boldsymbol{A} = \boldsymbol{Q}$$

then we have: $$\frac{1}{N}.[y|y] = \boldsymbol{\Sigma}$$

PCA outputs k decorrelated new variables, associated to the k non-zero eigenvalues of $[x|x]$. However, one will usually discard the variables less involved into the reconstruction of the variance of the original variables, so that only the r ($<k$) most "representative" variables remain. These r variables are the ones associated to the r largest eigenvalues of $[x|x]$, which also minimizes the effect of this data reduction over the L_2-norm of the reconstruction error of each variable [2]. From a practical point of view, variable y and transform $\boldsymbol{A}$ can easily be obtained by use of a Singular Value Decomposition of x, as will be described in section 2.1.c).

Although ICA also starts with the previously described step, it will further extend mutual information reduction up to a higher order, by achieving step 2. It can be shown that (2) and (3) are obtained if $\boldsymbol{B}$ is an orthogonal transform, i.e. $\boldsymbol{B}^{\dagger}\boldsymbol{B} = \mathbf{Id}$. As the analytical expression of probability density $p_{A \cdot x}$, is most often unknown, (4) can achieved by maximizing an approximate of $\sum_{i=1}^{N} J(p_{B.A.x_i})$. This approximate can be derived from the Edgeworth Expansion of $p_{A \cdot x}$, which uses higher-order cumulants. Most often however, maximizing $\sum_{i=1}^{N} J(p_{B.A.x_i})$ will be done by maximizing a "contrast function", which has the same evolution but has a simpler expression than its Edgeworth's expansion.

ICA algorithms differ in their choice for such a contrast function. We used J.-F. Cardoso's matlab routines which are freely available at ftp://sig.enst.fr/pub/jfc/Algo/Jade/jade.m. The contrast used for minimization is [4], [5]:

$$\kappa = \sum_{i,k,l=1,N} | \, Cum(y_i, y_i^*, y_k, y_l^*) \, |^2$$

where *Cum(.)* denotes the fourth order cumulant of variable y.

b) Application to HRTF data

These statistical analysis methods have already been applied to HRTF decomposition in the past. The first studies with PCA decomposed HRTF magnitude spectra [6], [7], [8]. Chen was the first to apply the Karhunen-Loeve expansion (another name for PCA used in the field of communication theory), to mixed-phase complex spectra [9], [10]. Such a decomposition leads straightforwardly to the multi-channel implementation of binaural synthesis according to Eq. (0) and Fig. 2 or 3.
The interest for applying ICA to HRTF modeling was raised by the possible link between statistical independence and non-overlapping support, or discreteness. It was applied on magnitude spectra by Emerit and yielded "independent filters", showing a main boost around given frequencies and near-unity gain elsewhere [11]. In this study, we investigate an extension of Chen's approach by decomposing complex HRTF spectra using ICA.

In order to apply the aforementioned concepts, the first task is to define what our variables are. As our data depend upon two dimensions, position and frequency, we have the following alternative:

(a) HRTF are seen as filters, and each of them is a variable observed M times (at each frequency sample). In other words, $\boldsymbol{H}$ stores the variables linewise.

(b) HRTF are seen as spatial functions, and each of them is a variable observed N times (at each position). This time, matrix $\boldsymbol{H}$ stores the variables columnwise.

The choice of which "dimension" defines the variables (position for ***(a)***, frequency for ***(b)***) can be made according to two implementation issues:

1. These two options lead to a different centering process.
In case ***(a)***, 0-mean will be achieved for each line of $\boldsymbol{H}$, and the mean is a position-dependent data, in other words a Spatial Function. The decomposition of the so-centered data therefore leads to:

$$HRTF(\theta, f) = \sum_{i=1}^{r} C_i(\theta).L_i(f) + \sum_{k=1}^{M} HRTF(\theta, f_k)$$

$$HRTF(\theta, f) = \sum_{i=1}^{r} C_i(\theta).L_i(f) + g_0(\theta)$$

The related implementation is shown in Figure 4, where the additional Spatial Function, denoted g_0, is added for each ear, without any associated filter.
In case ***(b)***, the variable mean is computed and subtracted columnwise, which leads to an "average filter" denoted $HRTF_0(f)$:

$$HRTF(\theta, f) = \sum_{i=1}^{r-1} C_i(\theta).L_i(f) + \sum_{k=1}^{N} HRTF(\theta_k, f)$$

$$HRTF(\theta, f) = \sum_{i=1}^{r-1} C_i(\theta).L_i(f) + HRTF_0(f)$$

We therefore obtain an additional filter which is associated to a constant Spatial Function, equal to 1. The corresponding implementation is shown in Figure 5.

2. The choice of variables also determines the dimension across which discreteness (or independence) will be achieved by ICA. The previous studies applying ICA to HRTF decomposition sought non-overlapping filters [11]. In order to obtain discrete Spatial Functions, as discussed in section 0.2, we will adopt option ***(b)***.

c) Relation between PCA and eigenvalue decomposition

As mentioned above, PCA relies on the eigenvalue decomposition of the data covariance matrix $\frac{1}{N}.[\boldsymbol{H}|\boldsymbol{H}]$. If the variables are not centered, this matrix no longer is the covariance matrix, but its eigenvalues can still be used to select a reduced number r of variables which minimize the L_2 reconstruction error. These new variables no longer are statistically decorrelated, but remain orthogonal. In order to estimate the advantage of decorrelated variables instead of only orthogonal variables, we apply the same eigenvalue decomposition algorithm to centered or non centered data.

The practical implementation of this decomposition may use a Singular Value Decomposition (SVD) of $\boldsymbol{H}$:

$$\boldsymbol{H} = \boldsymbol{U}.\boldsymbol{\Sigma}\boldsymbol{D}^{\dagger}$$

with $\boldsymbol{U}$ and $\boldsymbol{D}$ being respectively of dimension $N\times r$ and $M\times r$, and so that $\boldsymbol{U}^{\dagger}\boldsymbol{U} = \mathbf{Id}$ and $\boldsymbol{D}^{\dagger}\boldsymbol{D} = \mathbf{Id}$,

and $\boldsymbol{\Sigma}$ being a $r\times r$ diagonal matrix.

We will then choose: $\boldsymbol{C} = \boldsymbol{U}$, $\boldsymbol{L} = \boldsymbol{\Sigma}\boldsymbol{D}^{\dagger}$

We can check that $[\boldsymbol{C}|\boldsymbol{C}] = \mathbf{Id}$

In other words, the resulting Spatial Functions are orthonormal. In the case of centered variables, they are actually decorrelated, and we can also check that $\boldsymbol{C}$ is 0-mean columnwise: the new variables are centered and SVD yields the same result as PCA.

Contrast maximization can then be performed on the above matrix C resulting from SVD, and will yield a new matrix $\boldsymbol{C'}$ containing the r "independent" variables columnwise, and an $r\times r$ orthogonal matrix $\boldsymbol{B}$. We derive the new set of independent Spatial Functions $\boldsymbol{C'}$ and the corresponding set of Reconstruction Filters by:

$$\boldsymbol{C'} = \boldsymbol{C}.\boldsymbol{B}$$

$$L' = B^{\dagger}.L$$

Strictly speaking, this procedure can be called ICA only if it is applied to a matrix of HRTF data H that is centered (by subtracting its average across positions). However, this will yield independent Spatial Functions which are 0-mean themselves, a constraint that is not favorable to obtaining discrete Spatial Functions. This suggests applying the contrast maximization algorithm of ICA to a matrix C obtained by SVD of non-centered HRTF data.

2.2 Predetermined Spatial Functions

In this second decomposition approach, we first fix the set of Spatial Functions. Reconstruction Filters are derived by orthogonal projection of the HRTFs over these Spatial Functions, by applying the method recalled in section 1.2. By imposing this constraint, we can expect the accuracy of HRTF reconstruction to be altered in comparison to the Statistical Analysis methods (PCA or ICA). However, it may allow practical advantages.
The choice of Spherical Harmonics as Spatial Functions, for instance, is reasonable since they constitute a basis spanning all position-dependent functions [12]. Moreover, they are obviously non-individual dependent, and therefore provide a "universal" encoder. When this method is applied to minimum-phase HRTF data, the corresponding multichannel audio encoding format has been called "Binaural B format" [13]. Another practical advantage, especially when only the four 1rst order harmonics are retained, is that directional encoding can then be realized by a recording using a pair of Soundfield microphones. The binaural B format is also well suited to transaural decoding over multichannel loudspeaker, overcoming some limitations of 2-channel transaural reproduction [13], [14].

As our study focuses on the horizontal plane, C contains 7 spatial functions to cover all spherical harmonics up to order 3. As the first of these components is omnidirectional, the associated Reconstruction Filter corresponds to the mean filter exhibited in the statistical approach. A Binaural B encoder will therefore be of the kind illustrated in Figure 5: we can consider projection over Spherical Harmonics as a decomposition over centered data, the variables being chosen along the frequency dimension.
The elevation Reconstruction Filter is obtained as described in section 1.3, and the decomposition of the reconstruction error is projected over the vertical first-order Spherical Harmonic.

2.3 Predetermined Reconstruction Filters

In this approach, the Reconstruction Filters are taken from a fixed set of filters. Bill Gardner recently proposed and evaluated a method in which the Reconstruction Filters are minimum-phase HRTFs associated to specific positions, chosen by minimizing the Frobenius norm of the reconstruction error [15]. In the following, we will refer to his approach as "subset selection method".
The first step of this method is an SVD applied to non centered HRTFs, as described in section 2.1.c. The Spatial Functions thereby derived are decomposed using a QR decomposition with column pivoting [16]. This decomposition provides a permutation matrix which allows to sort the lines of C, i.e. the positions, according to their norm $\|[c_{i1}\ c_{i2}\ \ldots\ c_{i7}]^t\|^2$, where c_i denotes the ith

Spatial Function in $\boldsymbol{C}$. This ordering of position is used to select the Reconstruction Filters among the original HRTF. The corresponding Spatial Functions are derived by orthogonal projection of the HRTF over the Reconstruction filters, as described in section 1.2. As observed by Gardner, each function reaches its maximum value for the position corresponding to the HRTF selected, for which all the others equal zero. By construction, each Spatial Function tends to concentrate its energy around a given position, and the set of spatial functions shows good discreteness properties, as illustrated by the results presented in [15].

We applied this decomposition method to both non-centered and centered HRTF. In the centered case, the centering of $\boldsymbol{H}$ is achieved columnwise, as for the two preceding approaches, and the Reconstructed Filters are no longer selected among HRTFs, but among centered HRTFs.

3- Compared performance of different encoding/decoding schemes

The accuracy of the reconstruction is evaluated by computing the L_2 norm of the error for each frequency point, normalized by the energy of the original data. We are therefore plotting:

$$e(f) = \sqrt{\frac{|H - L.C|^2}{|H|^2}}$$

We will refer to eigenvalue decomposition as the generic name for PCA or Singular Value Decomposition applied on non-centered data. We will refer to contrast maximization as the generic name for ICA or contrast maximization applied to non-centered data. These four different methods will be globally (and improperly) called "statistical analysis methods".

3.1- L2- reconstruction error for the three approaches

The reconstruction error for one head is shown in Figure 6, for both centered and non centered data. Several observations can be made:

- Eigenvalue decomposition and contrast maximization lead to the same error, which can be expected from their theoretical derivation.
- All methods yield a similar error from 1 to 6kHz.
- The binaural B format scheme is less accurate than the other methods at higher frequencies.
- Centering the data before decomposition yields a more exact reconstruction at low frequencies. This is because the "mean filter" contains all the information common to all positions, and therefore all of the low-frequency information. However, the low-frequency error always remains very low for the "non-centered" methods.

Subset selection leads to a slightly less accurate reconstruction than statistical analysis, because of the constraint imposed in the determination of the Reconstruction Filters.

3.2- Deriving individual-independent Spatial Functions

While the Binaural B format provides "universal" Spatial Functions, statistical analysis and subset selection lead to individual-dependent Reconstruction Filters and Spatial Functions. However, provided that the Spatial Functions obtained show a sufficient consistency across subjects, a common set of spatial functions can be derived by averaging across subjects.
The first step of this procedure consists in the re-ordering of the Spatial Functions to obtain the best possible match for all of the 15 heads that we decomposed. The reordered Spatial Functions are then averaged. The optimal permutation is searched so that the total energy of the averaged Spatial Functions is maximal. The result of our permutation algorithm is illustrated in Figure 7, in the case of contrast maximization for non-centered data. The averaged Spatial Functions are shown in Figure 8 for contrast maximization and subset selection. The two yield similar spatial functions, both from centered and non-centered data. Contrast maximization seems to yield less overlapping functions than subset selection. This observation is confirmed by contrast values given in Table 1.

In order to evaluate the relevance of this averaging method, we check two criteria: the percentage of the total energy remaining after averaging, and contrast of the averaged spatial functions. If individual Spatial Functions cannot be matched properly, one would expect the total energy of the averaged Spatial Functions to be drastically inferior to the total energy of the individual ones. Furthermore we want to verify that averaging doesn't spoil the contrast reached for individual Spatial Functions. In Table 1, we report the values of these criteria for the analysis of centered data. For contrast computation, the "omni" Spatial Function associated to the average filter is removed from *C*.

	subset selection	**PCA**	**ICA**
% of total energy	65	66	60
mean contrast of individual S.F.	55	9	83
contrast of averaged S.F.	25	8	64

Table 1: Criteria for the relevance of Spatial Function averaging, decomposition of centered HRTF for 15 heads.

As we would have expected it, PCA provides a much lower contrast than subset selection, and than ICA which systematically yields the highest contrast. According to these figures, all three decomposition techniques show satisfying and similar "universality properties", since the energy of the resulting averaged Spatial Functions holds a large part of the initial energy.
We obtain effectively discrete Spatial Functions by forcing small gain factors to zero outside of the main lobe of each function. After the averaging stage and the discretization stage, we derive the reconstruction filters by an orthogonal projection of the HRTFs over the new Spatial Functions. The resulting L_2 reconstruction errors are shown in Figures 9 and 10.

As all three decomposition approaches show similar "universality properties", they also manage to reconstruct HRTF with a similar accuracy when using averaged Spatial functions. As expected, the error for eigenvalue decomposition and contrast maximization methods now differ. After "universalization" of the spatial functions, the accuracy of all methods is comparable to that of the binaural B format, which exhibits universal but non-discrete spatial functions. Compared to the performance obtained with individual spatial functions (Fig. 6), the deterioration is observed mostly for frequencies above 9 kHz. The performance of subset selection improves compared to Figure 6. This is actually explained by the re-design of the Reconstruction Filters corresponding to the averaged Spatial Functions, by use of orthogonal projection: the filters are no longer chosen among the original HRTF, which removes a constraint in optimizing the L2 error.

After "discretization" of the universal spatial functions (Fig. 10), the reconstruction error further increases. The performance of the subset selection and the contrast maximization methods are comparable, and now inferior to that of the binaural B format decomposition.

3.3- Consistency of the Spatial Functions vs. order of decomposition

We already noticed from Figure 8 that subset selection and contrast maximization yield spatial functions pointing to the same set of directions. In this section, we try to check the consistency of these "privileged" directions with the order of the decomposition, i.e. with the number of channels used. Figures 11, 12 and 13 show the polar diagrams of Spatial Functions for subset selection, eigenvalue decomposition and contrast maximization, for both centered and non-centered data. Several results can be underlined:

- All methods yield spatial functions biased towards the ipsilateral side (angles ranging from 180 to 360 degrees). This is consistent with Gardner's results for subset selection [15].

- By construction, the Reconstruction Filters given by subset selection at a given order remain selected as the order is increased. The associated Spatial Functions change somewhat from 3 to 7 channels because of the Reconstruction Filters are not mutually orthogonal and the projection involves the multiplication with the non-diagonal Gram matrix formed by these filters. However, the directions exhibited are inherently consistent.

- When applied to non-centered HRTF data, the subset selection and contrast maximization methods yield similar spatial functions for 3, 5 or 7 components. When applied to non centered data, the two methods yield less similar results for 5 and especially 3 components.

- The discreteness achieved with contrast maximization or subset selection seems to improve as the decomposition order is increased: the smaller the number of channels, the fewer possibilities exist to re-combine the Spatial Functions given by eigenvalue decomposition to achieve optimal discreteness. For centered data, the 3 channel case actually corresponds to 2 independent Spatial Functions (if we discard the omnidirectional component) and show little difference between PCA and ICA, as if the contrast maximization had not been applied.

3.4- Residual inter-individual differences

The modeling techniques described above introduce approximation errors at each step: (a) linear decomposition of HRTFs for each subject, (b) averaging across subjects, and (c) zeroing out spurious lobes in the spatial functions. It is now necessary to verify that enough information is preserved to allow modeling interindividual differences by modifying only the ITD values and the bank of Reconstruction Filters (or decoder). The errors introduced by the modeling techniques and the subsequent simplifications must remain small compared to the differences between heads.

Figure 14 shows the L_2-distance between the HRTFs of a given head and 14 other heads. On the same figure, we show the L_2 reconstruction error for ICA decomposition of this same head, after each step of the modeling procedure. The modeling error is systematically less than the interindividual differences. Up to 2 kHz, both the interindividual differences and the modeling errors remain low. As interindividual differences increase at higher frequencies, they clearly surpass the reconstruction error, which remains roughly unchanged. This indicates that a significant amount of individual information is preserved after reconstruction.

Conclusion

This paper has presented and compared several approaches to achieve a linear decomposition of minimum-phase HRTFs into a set of Spatial Functions and a set of Reconstruction Filters:

- Statistical analysis techniques, Principal Components Analysis and Independent Component Analysis, applied to centered or non-centered HRTF data
- Subset selection, applied to centered or non-centered HRTF data
- Projection of HRTFs over Spherical Harmonics (or Binaural B format decomposition).

For a single head, Principal or Independent Component Analysis minimizes the L_2 reconstruction error. All methods but the Binaural B format decomposition provide individual (or "non-universal") Spatial Functions. After averaging the Spatial Functions across subjects, the modeling accuracy is no longer significantly improved compared to the Binaural B format.

Independent Component Analysis and Subset Selection yield a similar set of Spatial Functions pointing towards consistent directions. After zeroing out the spurious lobes in these spatial functions to obtain "discrete" functions, a moderate deterioration of overall accuracy is observed. Centering the HRTF data prior to the decomposition improves the reconstruction performance to a limited extent, but is less advantageous from an implementation standpoint.

When 7 reconstruction filters are used to model the horizontal-plane HRTFs, the overall reconstruction error remains inferior to the overall interindividual differences. This indicates that significant inter-individual differences are preserved in the modeled HRTFs.

References

[1] Jot, J.-M., Larcher V. and Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. In Proc. 98th Conv. of the Audio Eng. Soc. Preprint 3980.

[2] Delchamps, D. (1988). State Space and Input-Output Linear Systems. Springer Verlag.

[3] Comon, P. (1994). Independent component analysis, a new concept ? Signal Processing, Elsevier, vol. 36, pp. 287-314.

[4] Cardoso, J. F. and Comon, P. (1996). Independent Component Analysis, a survey of algebraic methods. In Proc. ISCAS'96, vol. 2, pp.93-96.

[5] Cardoso, J.-F. (1998). Blind signal separation: statistical principles. In Proc. of the IEEE, vol. 9, n°10, pp. 2009-2025.

[6] Kistler, D. J. and Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. J. Acoust. Soc. Am., vol. 91, pp. 1637-1647.

[7] Middlebrooks, J. C. and Green, D. M. (1992). Observations on a principal components analysis of head-related transfer functions. J. Acoust. Soc. Am., vol. 92(1), pp.597-599.

[8] Wightman, F. and Kistler, D. (1993). Multidimensional scaling analysis of head-related transfer functions. In Proc. of IEEE Workshop on Application of Sig. Proc. to Audio and Acoustics.

[9] Chen, J. and Van Veen, B. D. and Hecox, K. E. (1992). External ear transfer function modeling: a beamforming approach. J. Acoust. Soc. Am, vol. 91, pp. 1333-1344.

[10] Chen, J. and Van Veen, B. D. and Hecox, K. E. (1995). A spatial feature extraction and regularization model for the head-related transfer function. J. Acoust. Soc. Am. vol. 97(1), pp.439-452.

[11] Emerit, M. (1995). Simulation binaurale de l'acoustique des salles de concert. PhD, INP Grenoble, France.

[12] Evans, J. E., Angus, J. A. S. and Tew, A. I. (1998). Analysing head-related transfer function measurements using surface spherical harmonics. J. Acoust. Soc. Am., vol. 104(4), pp. 2400-2411.

[13] Jot, J.-M., Wardle, S. and Larcher, V. (1998). Approaches to binaural synthesis. In Proc. 105th Conv. Audio Eng. Soc., preprint 4861.

[14] Jot, J.-M., Larcher, V. and Pernaux, J.-M. (1999). A comparative study of 3D audio encoding and rendering techniques. In Proc. 16th Conf. Audio Eng. Soc.

[15] Gardner, W. G. (1999). Reduced-rank modeling of head-related impulse responses using subset selection. In Proc. IEEE Workshop on Application of Sig. Proc. to Audio and Acoustics.

[16] Golub, G. H. and Van Loan, C. F. (1987). Matrix computations, Johns Hopkins Univ. Press, Baltimore.

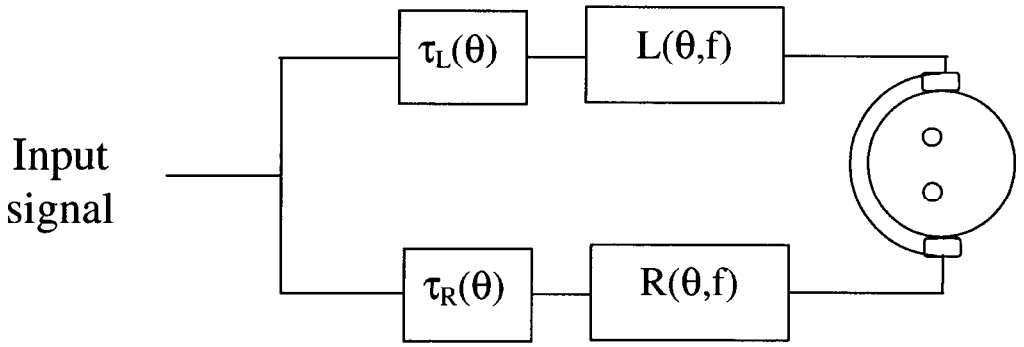

Figure 1: Traditionnal 2-channel implementation of binaural synthesis: τ_L and τ_R = pure delay filters ; L and R = minimum-phase filters.

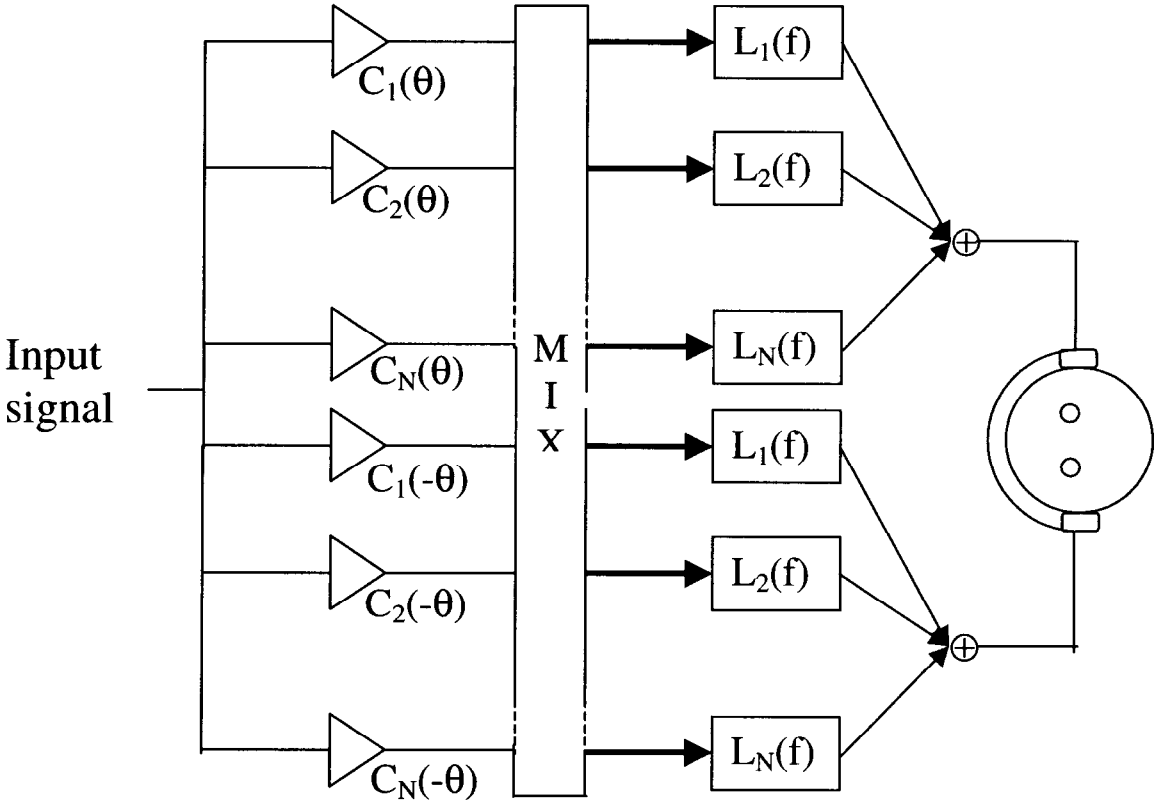

Figure 2: Multi-channel implementation of binaural synthesis using a decomposition of mixed-phase HRTF.

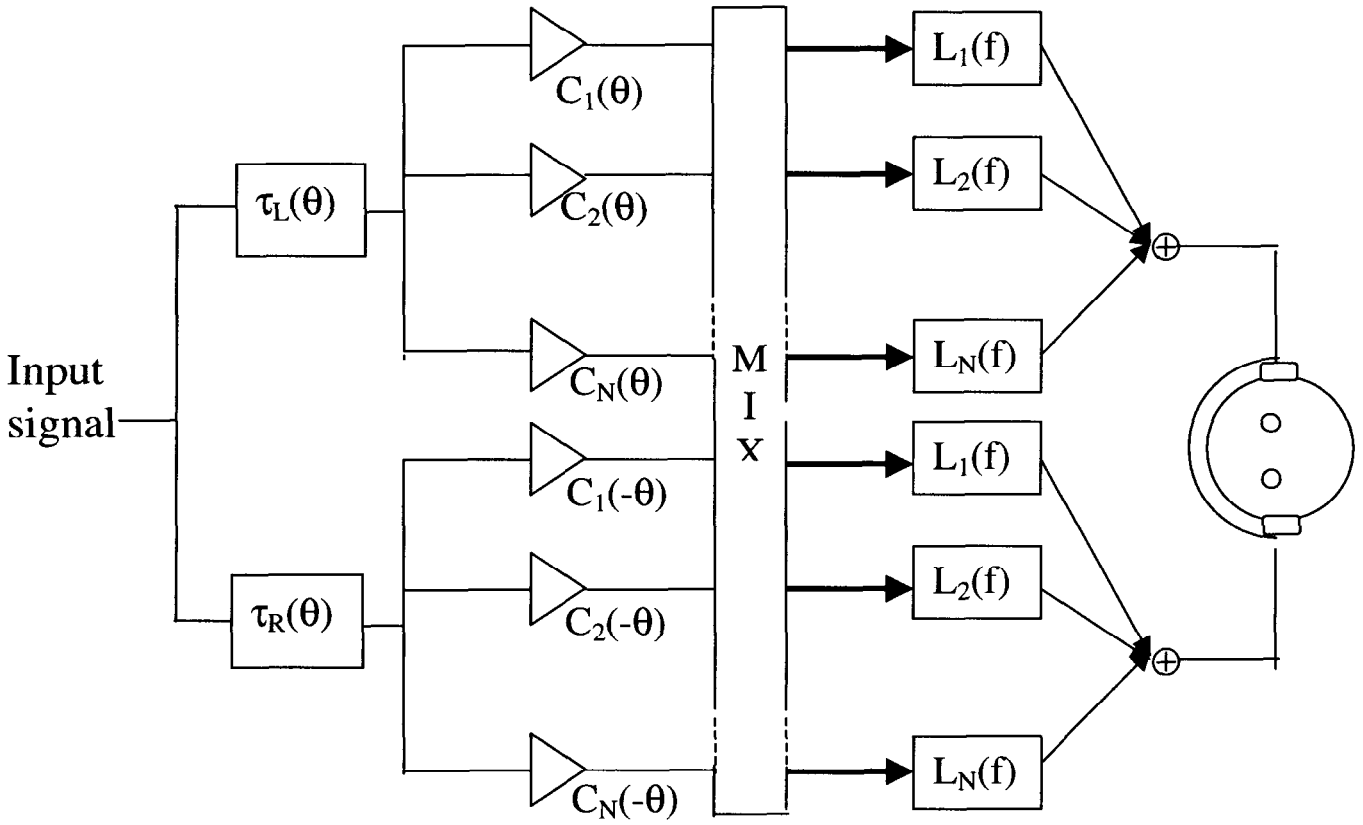

Figure 3: Multi-channel implementation of binaural synthesis using a decomposition of minimum-phase HRTF.

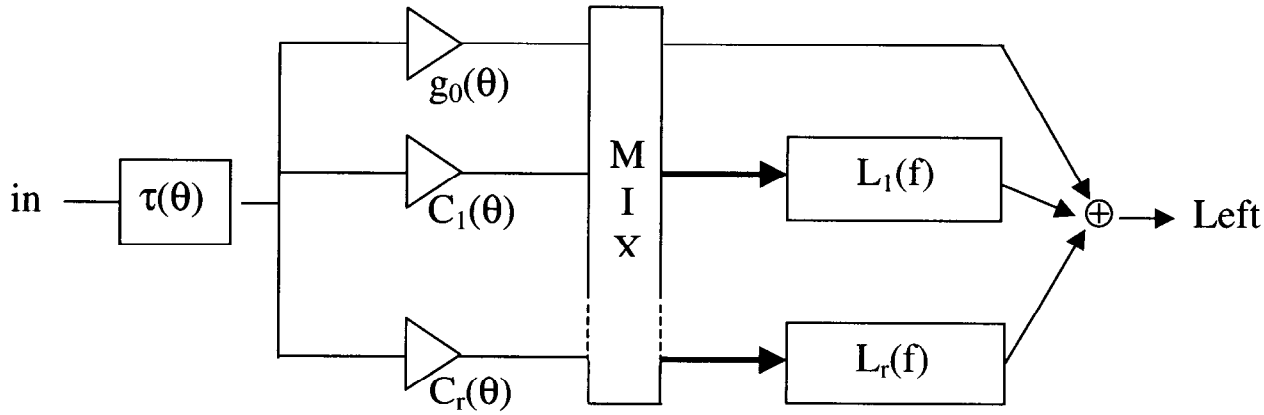

Figure 4: Decomposition of HRTF centered across frequency: implementation for one ear.

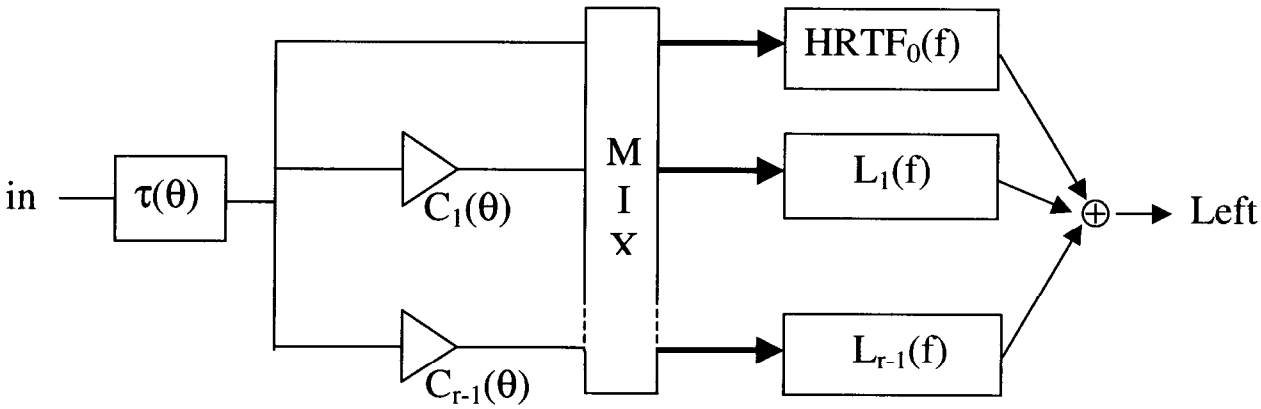

Figure 5: Decomposition of HRTF centered across position: implementation for one ear.

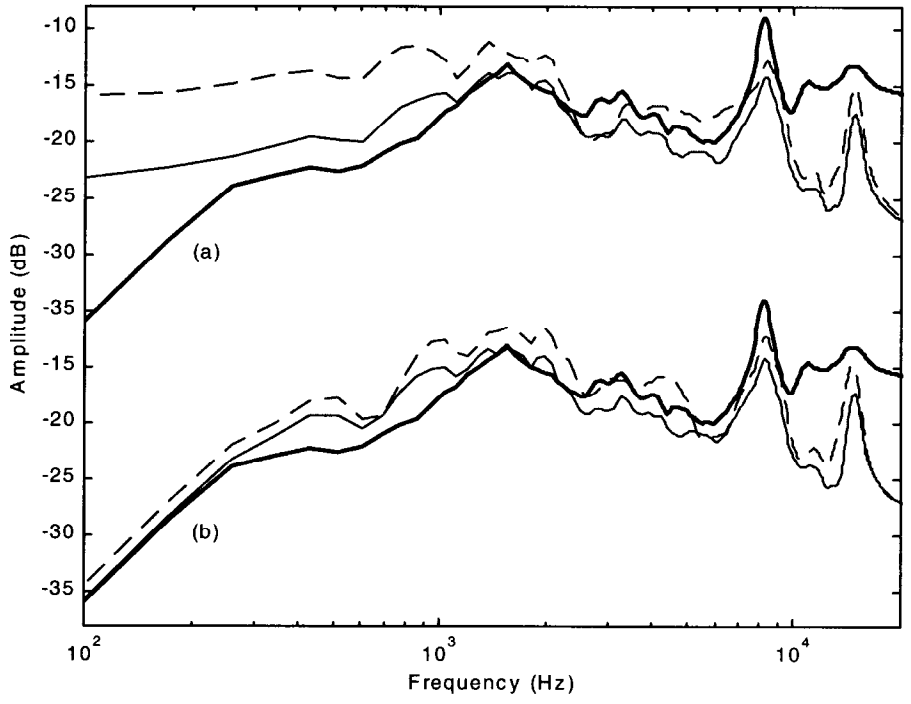

Figure 6: L2 reconstruction error in the horizontal plane, for several decompositions of non-centered data (a) and centered data (b). Thick continuous line: Binaural B, thin continuous line: Principal Component Analysis/ Independent Component Analysis, dashed line: subset selection.

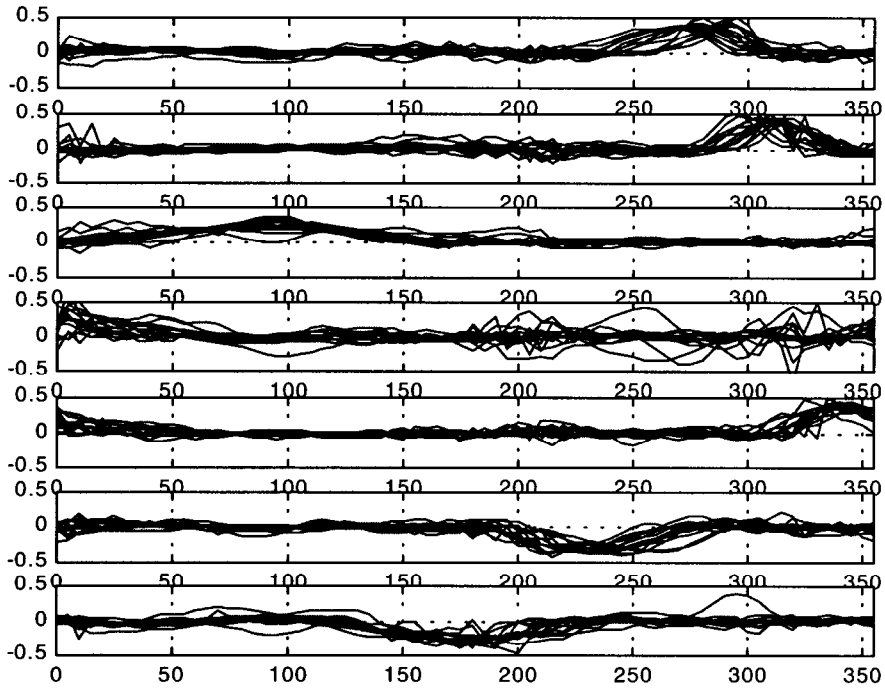

Figure 7: Spatial Functions for ICA decomposition of non-centered data: 15 heads, horizontal plane.

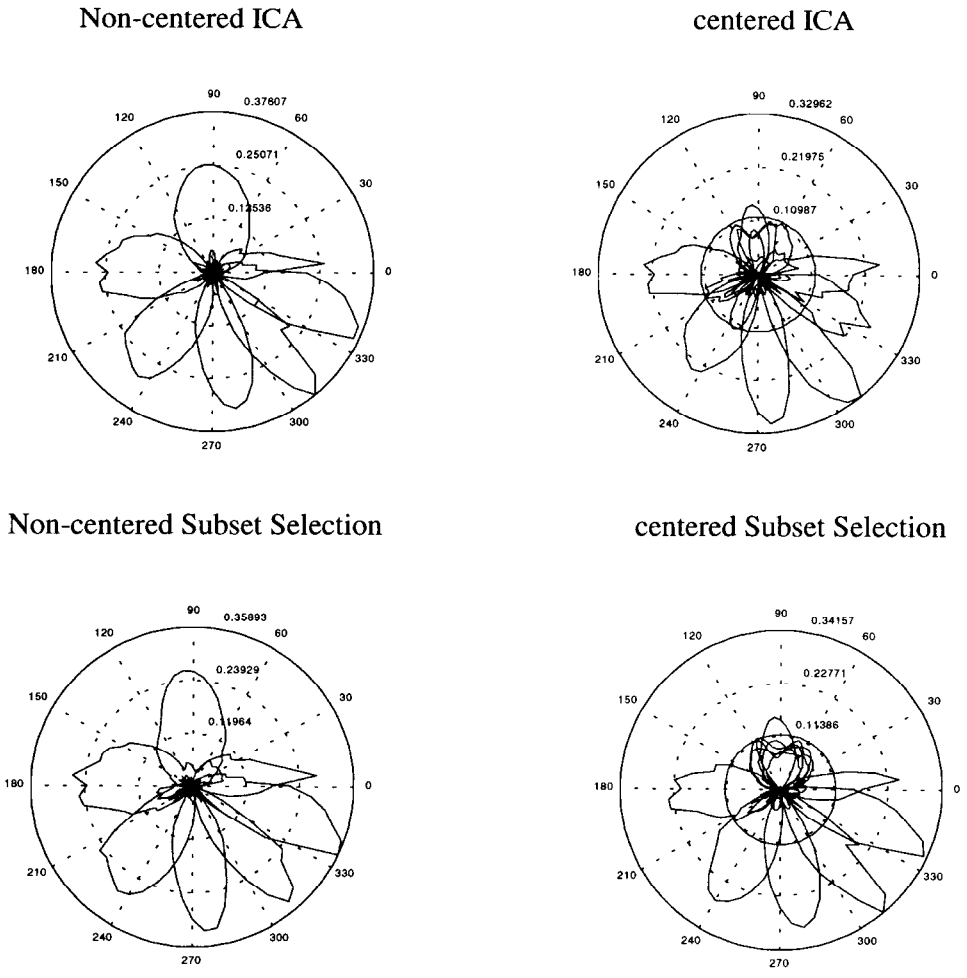

Figure 8: Averaged Spatial Functions in the horizontal plane.

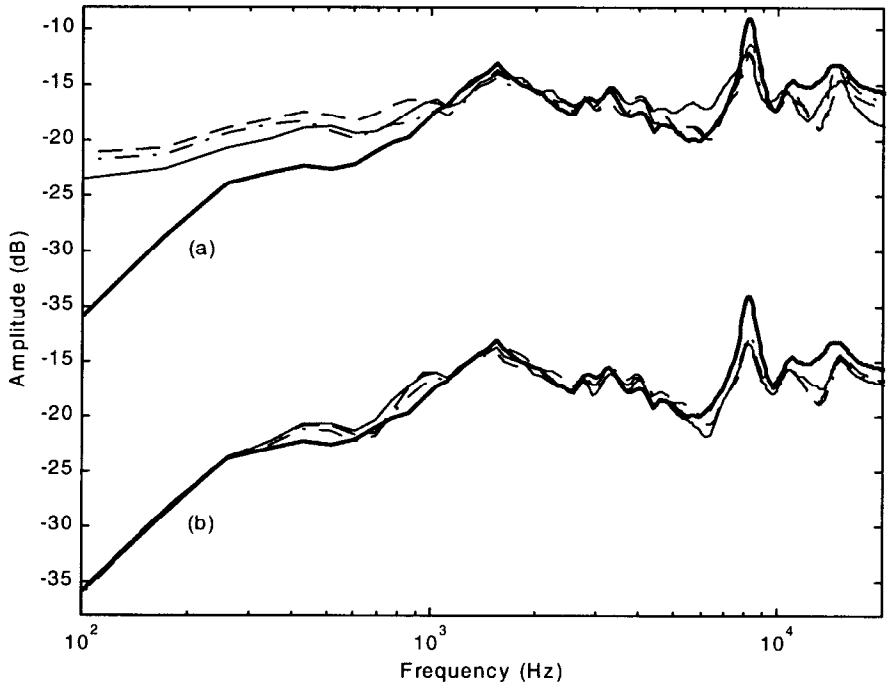

Figure 9: L2 reconstruction error in the horizontal plane, for reconstruction with averaged Spatial Functions of non-centered data (a) and centered data (b). Thick continuous line: Binaural B, thin continuous line: Independent Component Analysis, dash-dotted line: Principal Component Analysis, dashed line: subset selection.

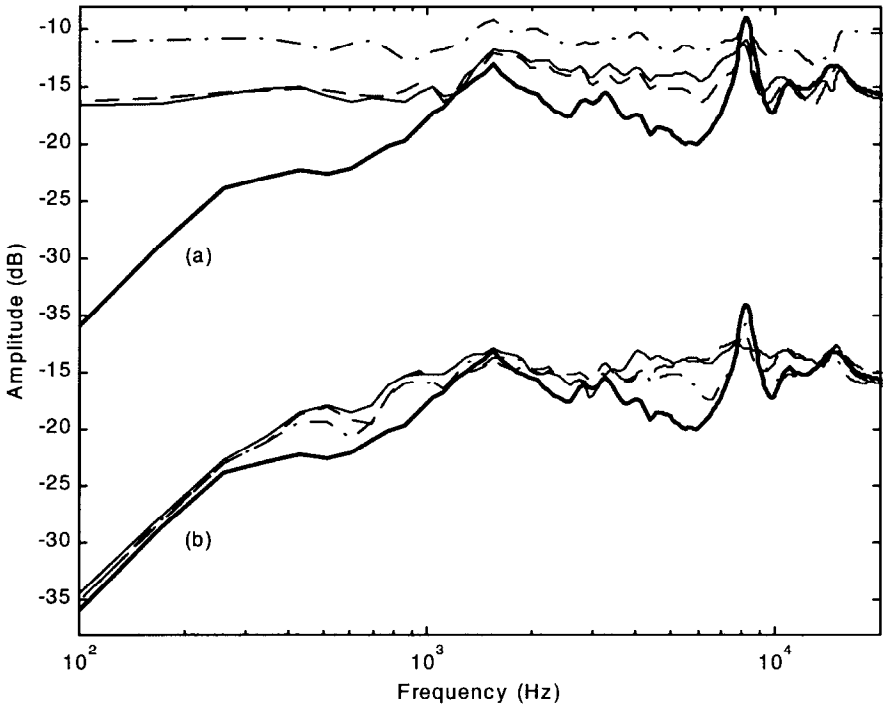

Figure 10: L2 reconstruction error in the horizontal plane, for reconstruction with discrete Spatial Functions of non-centered data (a) and centered data (b). Same line type conventions as above.

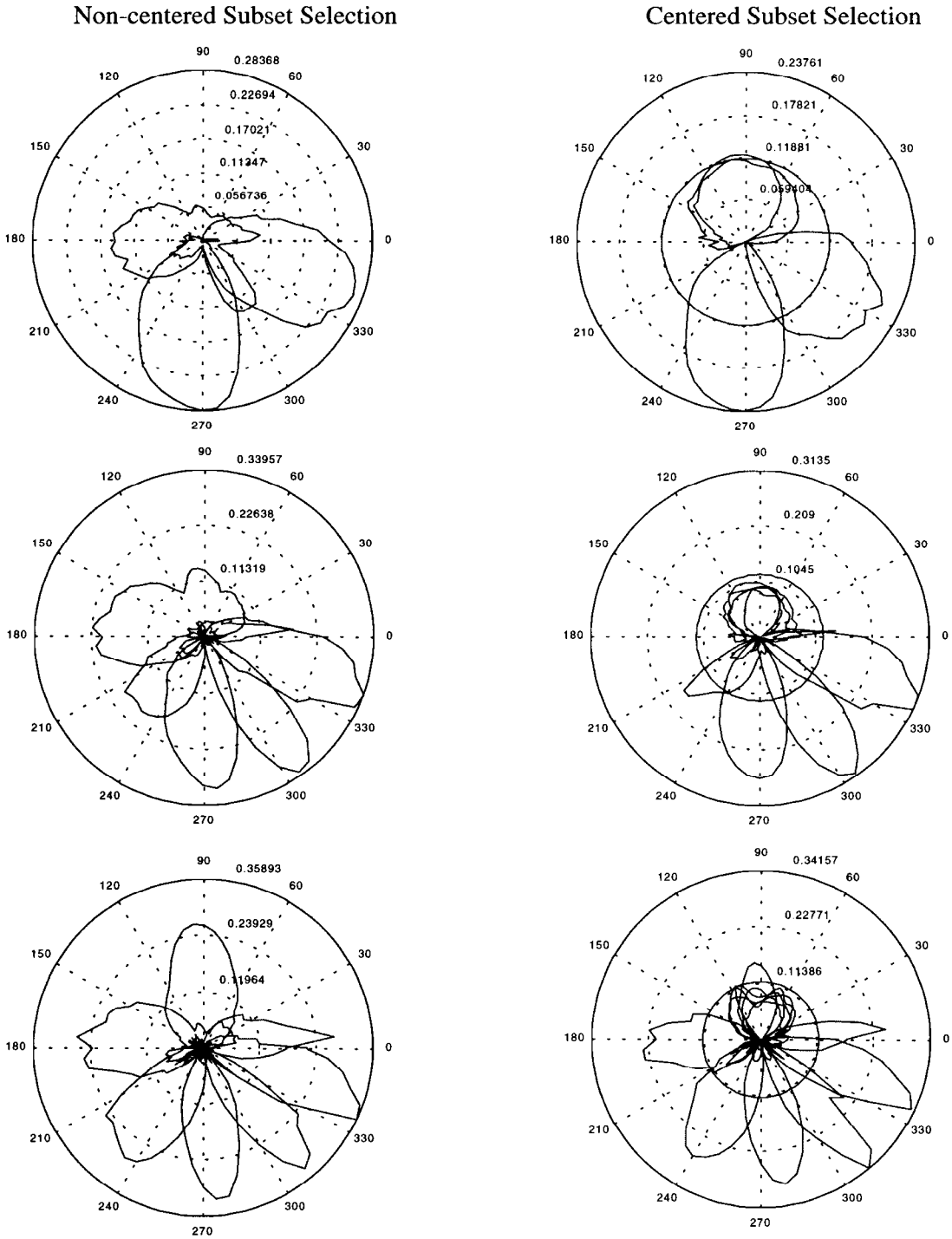

Figure11: Spatial Functions given by subset selection decomposition of centered and non-centered HRTF, for several orders (from top to bottom): 3,5 and 7 channels.

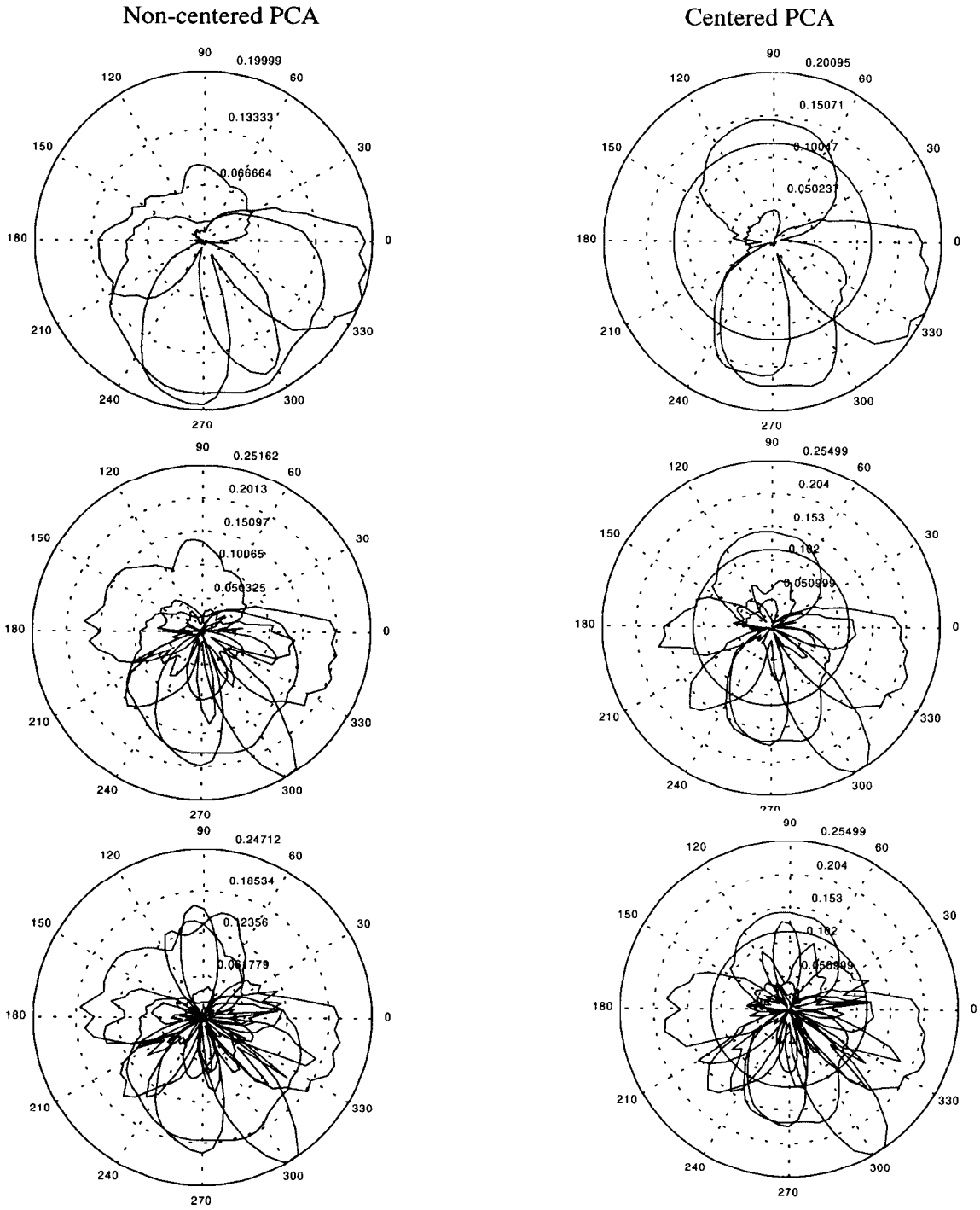

Figure12: Spatial Functions given by PCA decomposition of centered and non-centered HRTF, for several orders (from top to bottom): 3,5 and 7 channels.

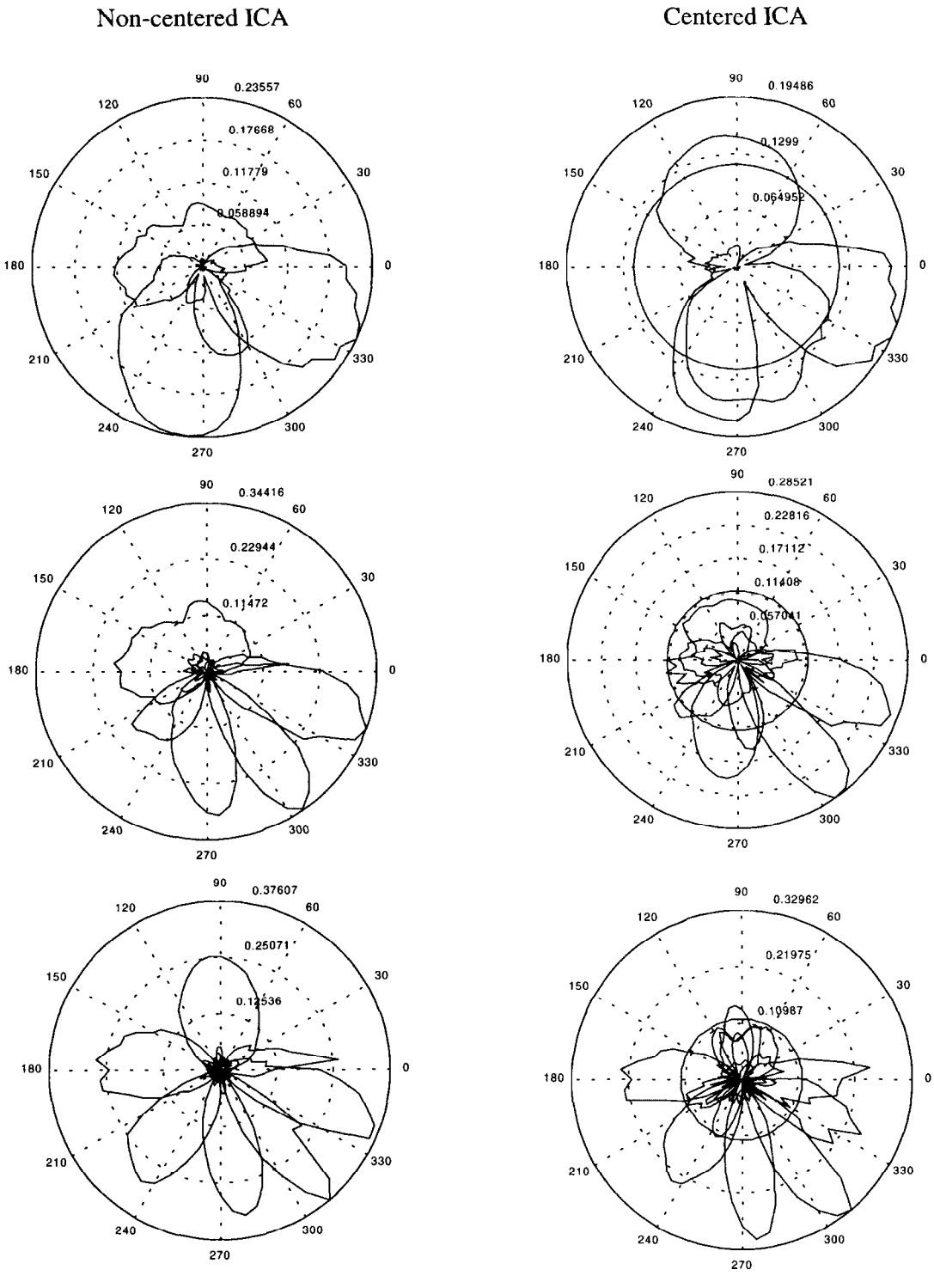

Figure13: Spatial Functions given by ICA decomposition of centered and non-centered HRTF, for several orders (from top to bottom): 3,5 and 7 channels.

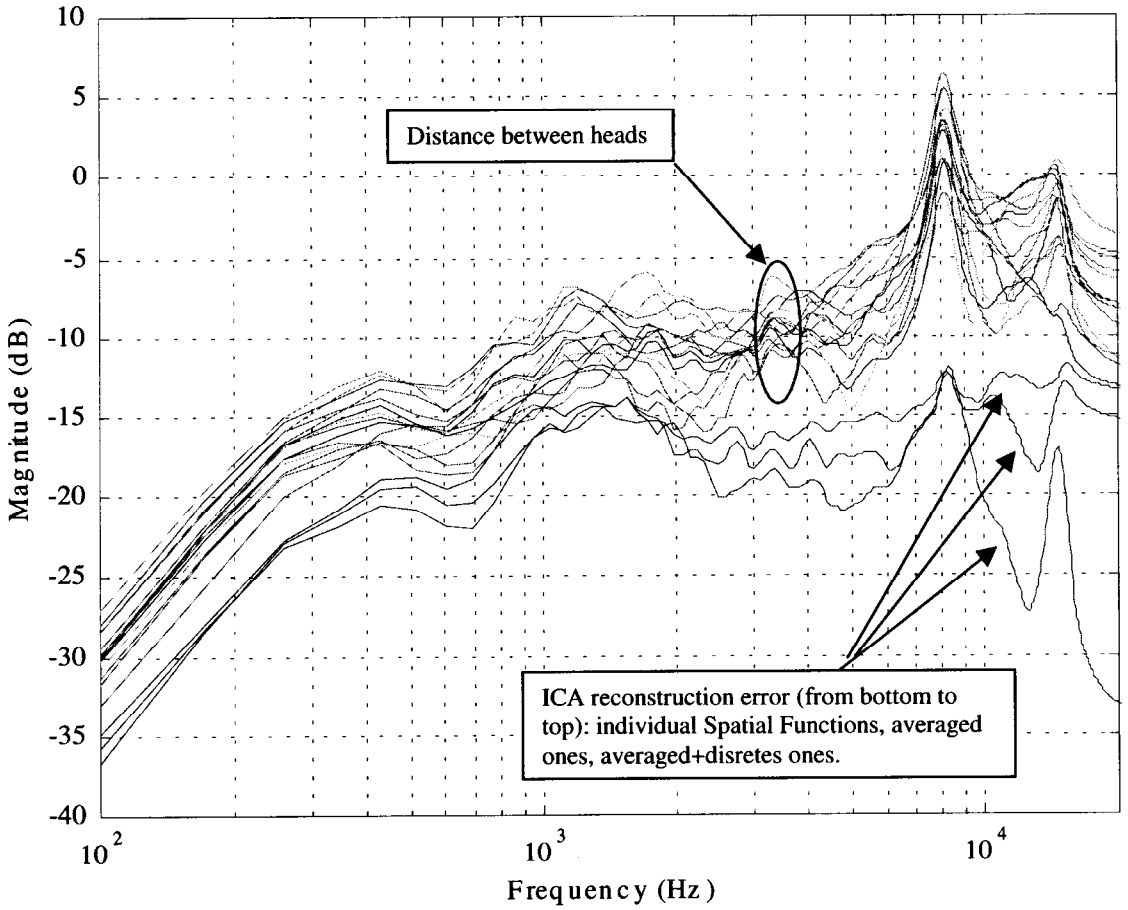

Figure14: L2 reconstruction error for one head (ICA) vs. L2 distance with other heads.