IMMERSIVE SOUND RENDERING USING LASER-BASED TRACKING

Panayiotis G. Georgiou, Athanasios Mouchtaris, Stergios I. Roumeliotis, Chris Kyriakakis*

Immersive Audio Laboratory, Integrated Media Systems Center,
University of Southern California,
Los Angeles, CA 90089-2564
e-mail: georgiou@sipi.usc.edu, mouchtar@sipi.usc.edu,
stergios@robotics.usc.edu, ckyriak@imsc.usc.edu

Abstract

In this paper we describe the underlying concepts behind the spatial sound renderer built at the University of Southern California's Immersive Audio Laboratory. In creating this sound rendering system, we were faced with three main challenges. First the rendering of sound using the Head-Related Transfer Functions, second the cancellation of the crosstalk terms and third the localization of the listener's ears. To deal with the spatial rendering sound we use a two-layer method of modeling the HRTF's. The first layer accurately reproduces the ITD's and IAD's, and the second layer reproduces the spectral characteristics of the HRTF's. A novel method for generating the required crosstalk cancellation filters as the listener moves was developed based on Low-Rank modeling. Using Karhunen-Loeve expansion we can interpolate among listener positions from a small number of HRTF measurements. Finally we present a Head Detection algorithm for tracking the location of the listener's ears in real time using a laser scanner.

1 INTRODUCTION

Applications for 3-D sound rendering include teleimmersion; augmented and virtual reality for manufacturing and entertainment; teleconferencing and telepresence; air-traffic control; pilot warning and guidance systems; displays for the visually impaired; distance learning; and professional sound and picture editing for television and film.

In implementing a sound rendering system one is faced with three main challenges. First the rendering of virtual sound sources using the Head-Related Transfer Functions, second the cancellation of the crosstalk terms that is necessary for loudspeaker-based system and third the localization of the listener's ears in order to dynamically adjust both the HRTF and crosstalk cancellation filters as the listener moves.

Although localization in the horizontal plane can be attributed mainly to time and level (amplitude) differences, variations in the spectrum as a function of azimuth and elevation angles also play a key role in sound localization. These variations arise mainly from reflection and diffraction effects caused by the outer ear (pinna) that give rise to amplitude and phase changes for each angle. In fact, ITD and IAD alone are not sufficient to explain localization of sounds in the median plane, in which ITDs and IADs are both zero. These effects are described by a set of functions known as the *Head-Related Transfer Functions* (HRTF's). Each HRTF, realized as a filter, is in effect the impulse response of a pinna which is viewed as a linear and time-invariant system varying with the direction of the sound source.

In section 2 of this paper, we present a two-layer method of modeling HRTF's for immersive audio rendering systems. The first layer accurately reproduces the ITD's and IAD's, and the second layer reproduces the spectral characteristics of the HRTF's. This allows for two degrees of control over the accuracy of the model. For example, increasing the number of measured HRTF's improves the spatial resolution of the system. On the other hand, increasing the order of the model extracted from each measured HRTF improves the accuracy of the response for each of the measured directions.

*This research has been funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152.

For a given sound direction, localization can be accomplished by filtering an audio signal with the appropriate pair of HRTF's (one for each ear). This is true, though, only when using headphones for sound reproduction. When loudspeakers are used, it is clear that the physical setting introduces cross-terms, since each ear receives sound from both loudspeakers. These cross-terms need to be canceled, so that we can exert sufficient control on what reaches each of the listener's ears, as in the headphones case. Additionally, the frequency response of the loudspeakers in practice is not flat and needs to be equalized (this is also true for the response of the headphones). The approach followed here and described in section 3 realizes acoustic crosstalk cancellation based on Karhunen-Loeve expansion that was used to interpolate among listener positions from a small number of HRTF measurements.

For both the spatial rendering and the crosstalk cancellation, accurate localization of the listener's ears relevant to the sound source is required. In section 4 a Head Detection and Tracking (HDT) algorithm is implemented for this purpose. This algorithm uses data provided by a laser tracking system measuring distances of points up to 8 meters in front of the device. The significant advantage of this type of sensor compared to vision-based systems is that it does not depend on the illumination conditions of the room and therefore it can work even in total darkness. The HDT algorithm detects the head of the listener and estimates the position of his ears relevant to the sound source. The algorithm also tracks the listener's rotational motion. The tracking part of this algorithm is based on a simple model of head motion previously used for enhancing the rendering speed of a head-on display system [1].

2 Spatial Audio Rendering

One of the key drawbacks of 3-D audio rendering systems arises from the fact that each listener has HRTF's that are unique for each angle. Measurement of HRTF's is a tedious process that is impractical to perform for every possible angle around the listener. Typically, a relatively small number of angles are measured and various methods are used to generate the HRTF's for an arbitrary angle. Previous work in this area includes modeling using principal component analysis [2], as well as spatial feature extraction and regularization [3].

In this section, we present a two-layer method of modeling HRTF's for immersive audio rendering systems. Kung's method [4] was used to convert the time-domain representation of HRTF's in state-space form. The models were compared both in their *Finite Impulse Response* (FIR) filter form and their state-space form. It is clear that the state-space method can achieve greater accuracy with lower order filters. This was also shown using a balanced model truncation method [5]. Although an *Infinite Impulse Response* (IIR) equivalent of the state-space filter could be used without any theoretical loss of accuracy, it can often lead to numerical errors causing an unstable system, due to the large number of poles in the filter. State-space filters do not suffer as much from the instability problems of IIR filters, but require a larger number of parameters for a filter of the same order. However, considering that there are similarities among the impulse responses for different azimuths and elevations, a combined single system model for all directions can provide, as we will show, a significant reduction.

Previous work on HRTF modeling has mainly focused on methods that attempt to model each direction-specific transformation as a separate transfer function. In section 2 we present a method that attempts to provide a single model for the entire 3-D space. The model builds on a generalization of work by Haneda *et al.* [6], in which the authors proposed a model that shares common poles (but not zeros) for all directions. Our model uses a multiple-input single-output state-space system to create a combined model of the HRTF's for all directions simultaneously. It exploits the similarities among the different HRTF's to achieve a significant reduction in the model size with a minimum loss of accuracy.

2.1 Two Layer Model

One way to spatially render 3-D sound is to filter a monaural (non-directional) signal with the HRTF's for the desired direction. This involves a single filter per ear for each direction and a selection of the correct filter taps through a lookup table. The main disadvantage of this process is that only one direction can be rendered at a time and interpolation can be problematic. In our work we extract and model the important cues of ITD and IAD as a separate layer, thus avoiding the problem of dual half-impulse responses created by interpolation. The second layer of the interpolation deals with the angle-dependent spectrum variations (Fig. 1). This is a multiple-input single-output system (for each channel), which we created in state-space form.

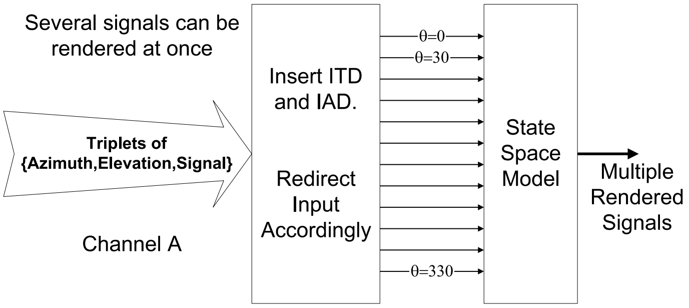
Several signals can be rendered at once
Triplets of {Azimuth,Elevation,Signal}
Channel A
Insert ITD and IAD.
Redirect Input Accordingly
θ=0
θ=30
θ=330
State Space Model
Multiple Rendered Signals

Figure 1: The unprocessed signals are passed to the algorithm along with the desired azimuth and elevation angles of projection.

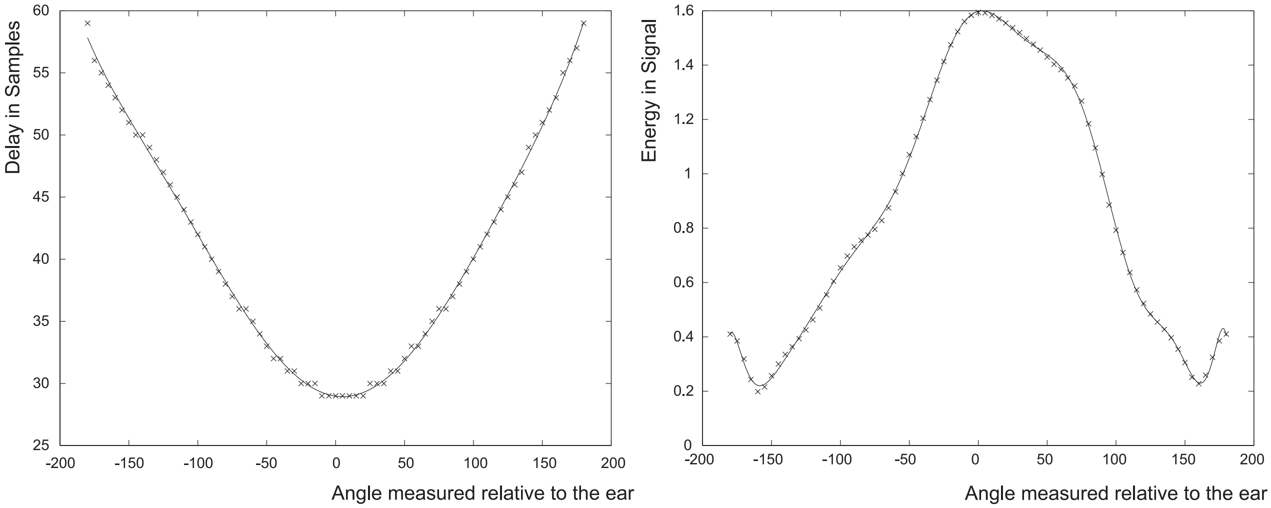

Figure 2: Extracted delay and sixth order polynomial on the left, while on the right is the extracted energy versus a twelfth order polynomial fit. fit.

The signal for any angle θ can be fed to the input corresponding to that angle, or if there is no input corresponding to θ then the signal can be split into the two adjacent inputs (or more in the case of both azimuth and elevation variations). In order to proceed with the two-layered model described above, we first extract the delay from the measured impulse responses. Fig. 2 (left) shows the delay extracted from the measurements and fitted with a sixth order polynomial.

It should be noted that here the azimuth is measured from the center of the head relative to the midcoronal and towards the face as shown in Fig. 3 and not relative to the midsagittal and clockwise as is common practice. For example, the azimuth of 270° relative to the midsagittal corresponds to 180° for the right ear but to 0° for the left ear measured with this proposed convention. This method of representation was chosen because it allows us to use two copies of the same model for both ears.

Similarly we can approximate the gain with a 14th order polynomial as in Fig. 2 (right). The advantages of polynomial fitting are not so obvious when only one elevation is considered, but become more evident when the entire 3-D space is taken into consideration.

Removing the initial delay and gain of the HRIR's of Fig. 4 (left) we are left with the set of impulse responses

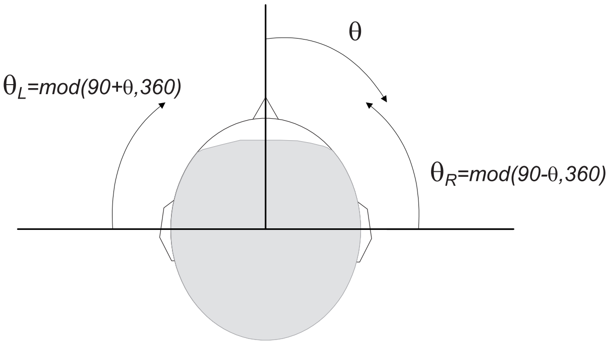

Figure 3: Proposed convention of measuring azimuth in order to have a single delay and gain function for both ears.

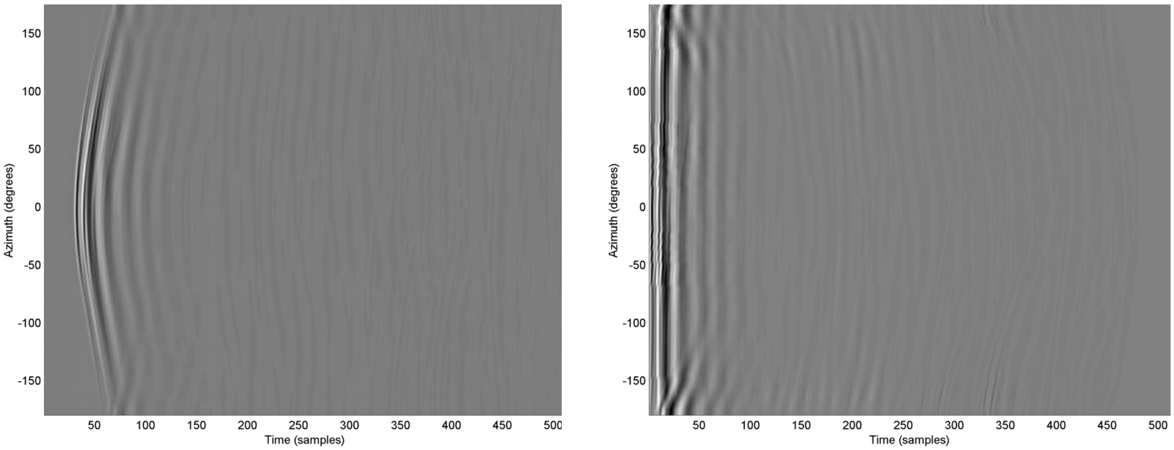

Figure 4: Left: Original HRIR's for 0° elevation and a 5° azimuth resolution. Right: HRIR's after initial delay and gain are removed for 0° elevation and a 5° azimuth resolution.

that will be modeled by the second layer. These HRIR's are very similar to each other as shown on Fig. 4 (right).

2.2 FIR Filter Reduction

The measurements used in this paper consist of impulse responses taken using a KEMAR dummy head [7]. These 512-point impulse responses can be used as an FIR model against which our comparisons will be based. In order to reduce these impulse responses we used the method first proposed by Kung [4] at the 12th Asilomar Conference on Circuits, Systems and Computers. The one input-one output case is briefly described below.

Note that alternative methods can be used (see Mackenzie *et al.* [5]). For this and other methods, the reader can refer to the original paper by Kung [4], as well as Beliczynski *et al.* [8] and references therein.

Consider an impulse response model of a causal, stable, multivariable and linear time-invariant system. If the system state space model is

$$x(n+1) = Ax(n) + Bu(n) \tag{1}$$

$$y(n) = Cx(n) + Du(n) \tag{2}$$

and an impulse is applied to the system then (assuming that $u_0 = 1$, without loss of generality):

$$\begin{array}{ccccccc} y_0 & = & D & & & \\ x_1 & = & B & y_1 & = & CB \\ x_2 & = & AB & y_2 & = & CAB \\ x_3 & = & A^2B & y_2 & = & CA^2B \\ & \cdots & & & \cdots & \\ & \cdots & & & \cdots & \\ x_N & = & A^NB & y_N & = & CA^NB \end{array} \tag{3}$$

Forming the above into a matrix:

$$\begin{bmatrix} y(n) \\ y(n+1) \\ y(n+2) \\ \vdots \end{bmatrix} = \begin{bmatrix} CB & CAB & CA^2B & \dots \\ CAB & CA^2B & CA^3B & \dots \\ CA^2B & CA^3 & \dots & \dots \\ CA^3B & CA^4B & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} u(n) \\ 0 \\ \vdots \\ \vdots \end{bmatrix}$$

Separating the Hankel matrix (i.e., the matrix that is in position (i, j) is $CA^{i+j-1}B$) and expressing it in its Singular Value Decomposition (SVD) components:

$$H = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} B & AB & A^2B \dots \end{bmatrix} = \Omega\Gamma = U\Sigma V^T \tag{4}$$

where U, V are unitary matrices and Σ contains the singular values along its diagonal in decreasing magnitude, i.e.,

$$\Sigma = \mathrm{Diag}[\sigma_1,\ \sigma_2,\ \sigma_3,\ \dots\ \sigma_r,\ \sigma_{r+1},\ \dots\ \sigma_{N+1}] \tag{5}$$

and Ω and Γ are the extended observability and reachability matrices that can be expressed in terms of the SVD components of H as:

$$\Omega = U\Sigma^{\frac{1}{2}} \quad \text{and} \quad \Gamma = \Sigma^{\frac{1}{2}}V^T \tag{6}$$

One way to reduce the model is to use

$$H = \begin{bmatrix} U_n & \overline{U_n} \end{bmatrix} \cdot \begin{bmatrix} \Sigma_n & 0 \\ 0 & \overline{\Sigma_n} \end{bmatrix} \begin{bmatrix} V_n \\ \overline{V_n}^T \end{bmatrix} \tag{7}$$

and reduce Ω and Γ to:

$$\Omega_n = U_n\Sigma_n^{\frac{1}{2}} \text{ and } \Gamma_n = \Sigma_n^{\frac{1}{2}}V_n^T \tag{8}$$

This will give:

$$\begin{array}{cccccc} A & = & \Sigma^{-\frac{1}{2}}U_n^TU_n^{\uparrow}\Sigma^{\frac{1}{2}} & C & = & U_n^1\Sigma^{\frac{1}{2}} \\ B & = & \Sigma^{\frac{1}{2}}\left(V_n^1\right)^T & D & = & y_0 \end{array} \tag{9}$$

While there are several definitions (10), one that also guarantees stability is

$$U_n = \begin{bmatrix} U_n^1 \\ \vdots \\ U_n^{N-1} \\ U_n^N \end{bmatrix} \text{ and } U_n^{\uparrow} = \begin{bmatrix} U_n^2 \\ \vdots \\ U_n^N \\ 0 \end{bmatrix} \tag{10}$$

To achieve higher speeds in model creation and the ability to handle any model size, Kung's method is performed on each impulse response separately. This avoids the dimension increase of the Hankel matrix and consequently drops the computational cost of the SVD significantly since SVD is an O(3) operation. The individual state-space models are combined in a single model to form the final model. Further reduction can be achieved on the resulting model if desired.

2.3 Results

The measurements used in this paper consist of impulse responses taken using a KEMAR dummy head [7]. These 512-point impulse responses can be used as an FIR model against which our comparisons will be based. In order to reduce these impulse responses we used the method first proposed by Kung [4] at the 12th Asilomar Conference on Circuits, Systems and Computers.

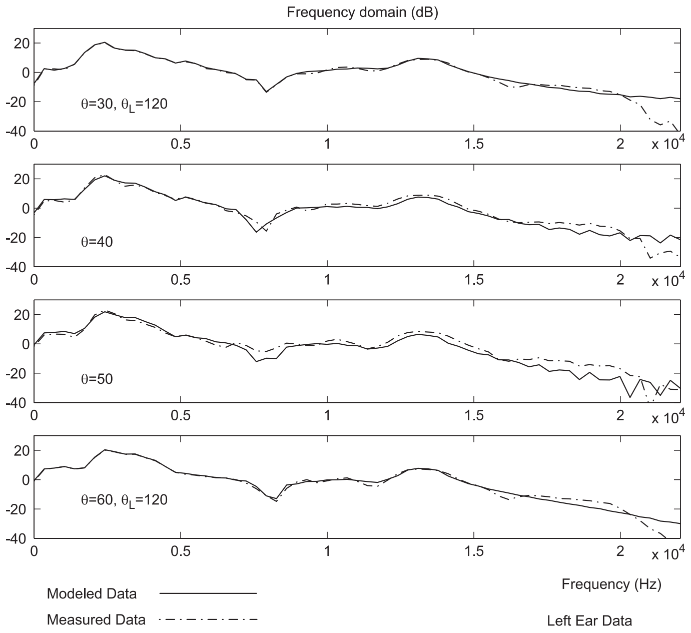

Figure 5: Frequency domain of measured and simulated impulse responses for a model created with a 30° resolution. $\theta = 40$, and $\theta = 50$ were not used for the creation of the model

Note that alternative methods can be used (see Mackenzie *et al.* [5]). For this and other methods, the reader can refer to the original paper by Kung [4], as well as Beliczynski *et al.* [8] and references therein.

To achieve higher speeds in model creation and the ability to handle any model size, Kung's method is performed on each impulse response separately. This avoids the dimension increase of the Hankel matrix and consequently drops the computational cost of the SVD significantly since SVD is an O(3) operation. The individual state-space models are combined in a single model to form the final model. Further reduction can be achieved on the resulting model if desired.

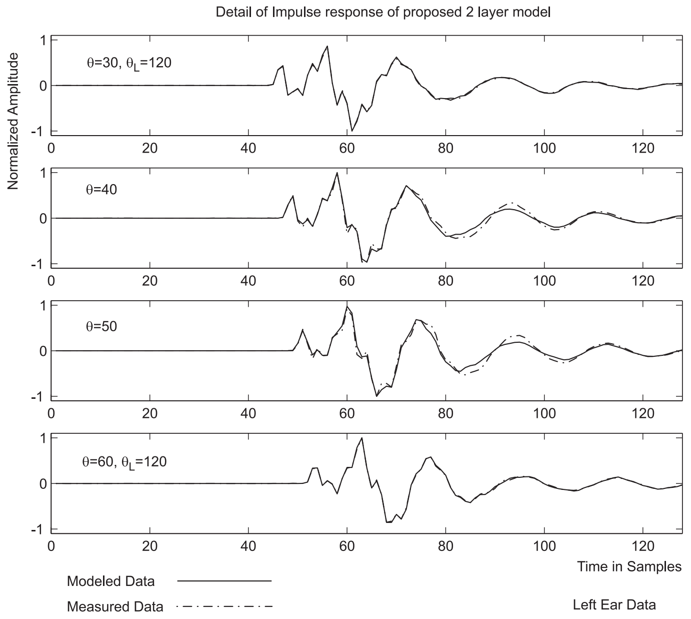

Figure 6: Detail of the time domain of Fig. 5

The advantages of the two-layer HRTF model can better be observed by examining a few representative impulse responses. Figs. 5 and 6 show the measured data with a dashed line and the simulated data with a solid line. The model was created with data measured every 30°, and therefore only data from the first and last plot of each figure were used for the creation of the model. The other two simulated responses in the plot correspond to data synthesized from the 30° and 60° inputs of the state-space model. For example, angle 40° corresponds to $\frac{2}{3}$ of the input signal being fed through the 30° input, while the remaining $\frac{1}{3}$ is input to the 60° direction. As expected, the two main cues of delay and gain were preserved in the impulse response since they are generated from a separate, very accurate layer. The second layer can then be reduced according to the desired accuracy.

Fig. 7 shows the performance of a further reduced state space model. The model was reduced to less than a third its initial size (down to 191 states from 600). The reduction was performed using techniques as described in [9] and [10]. As can be seen from the figures, there was some minor loss of accuracy. Fig. 8 displays the performance of an equivalent model size that was created by reducing each individual HRTF to a 16 state model. These models correspond to a combined model of 192 states that is of equivalent size to the previous combined model but that performs very poorly. The advantage of performing the reduction to the combined model, as described above, is clearly evident.

3 Loudspeaker-Based Rendering

In order to deliver the appropriate binaural sound field to each ear through loudspeakers, it is necessary to eliminate the crosstalk that is inherent in all loudspeaker-based systems. This limitation arises from the fact that while each loudspeaker sends the desired sound to the same-side (ipsilateral) ear, it also sends undesired sound to the opposite-side (contralateral) ear.

Crosstalk cancellation can be achieved by eliminating the terms H_{RL} and H_{LR} (Fig. 9), so that each loudspeaker is perceived to produce sound only for the corresponding ipsilateral ear. Note that the ipsilateral terms (H_{LL}, H_{RR}) and the contralateral terms (H_{RL}, H_{LR}) are just the HRTF's associated with the position of the two loudspeakers with respect to a specified position of the listener's ears. This implies that if the position of the listener changes then these terms must also change so as to correspond to the HRTF's for the new listener position.

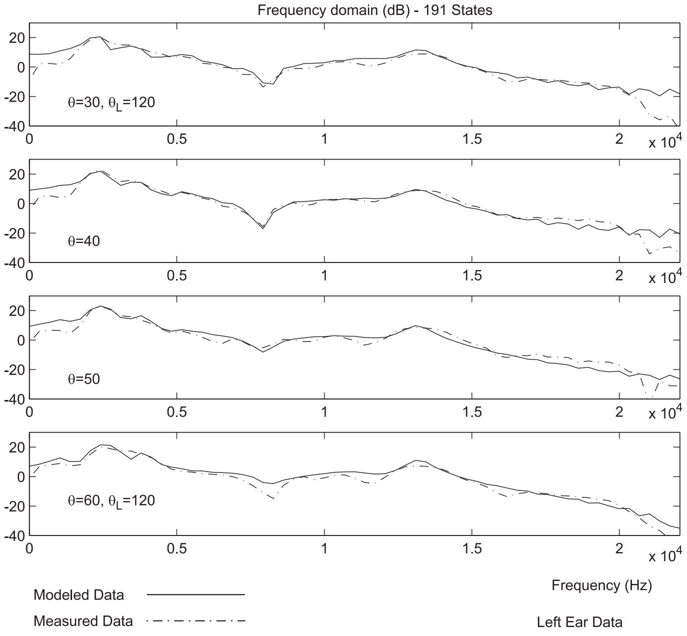

Figure 7: Model used is reduced down to 191 states from an original size of 600 states. Accuracy has not decreased significantly.

One of the key limitations of crosstalk cancellation systems arises from the fact that any listener movement that exceeds 75 to 100 mm completely destroys the desired spatial effect. This limitation can be overcome by tracking of the listener's head in three-dimensional space. A prototype system that used a magnetic tracker and adjusted the HRTF filters based on the location of the listener was demonstrated by Gardner [11, 12].

Several schemes have been proposed to address crosstalk cancellation. The first such scheme was proposed by Atal and Schroeder [13] and later another was published by Damaske and Mellert [14, 15]. A method proposed by Cooper and Bauck modeled the head as a sphere and then calculated the ipsilateral and contralateral terms [16, 17]. They showed that under the assumption of left-right symmetry a much simpler shuffler filter can be used to implement crosstalk cancellation as well as synthesize virtual loudspeakers in arbitrary positions. Another method by Gardner approximates the effect of the head with a low-pass filter, a delay and a gain (less than 1) [18].

While these methods have the advantage of low computational cost, the spherical head approximations can introduce distortions particularly in the perceived timbre of virtual sound sources behind the listener. Furthermore, the assumption that the loudspeakers are placed symmetrically with respect to the median plane (i.e., $H_{LR} = H_{RL}$ and $H_{LL} = H_{RR}$) leads to a solution that uses the diagonalized form of the matrix introduced by the physical system [16, 17]. This solution can only work for a non-moving listener seated symmetrically to the loudspeakers. In this paper, we use a different approach for the analysis that can be easily generalized to the non-symmetric case that arises when the listener is moving [19]. In our analysis we present the non-symmetric case for a listener placed at the center between two loudspeakers but being able to do rotational movement (assuming the ears of the listener are at the same level as the loudspeakers for simplicity purposes only). For this case, the terms ipsilateral and contralateral terms are not equal, however they still correspond to the HRTF's of a specific angle. This angle can be found based on the angle of the loudspeaker placement with respect to the listener and to the angle of rotation of the listener with respect to the fully symmetric case. If the angle of the loudspeakers with respect to the median plane is a degrees referring to the right loudspeaker, and the listener has rotated b degrees clockwise, then the required HRTF's will be as follows: H_{RR} will be the ipsilateral HRTF for the angle $\alpha - \beta \mod 360^\circ$ and H_{RL}

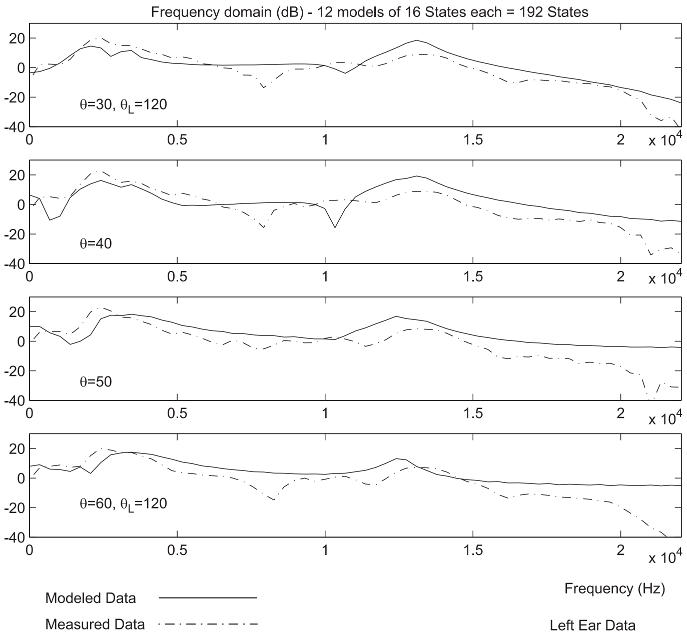

Figure 8: 12 Models of total 192 states. Accuracy has dropped significantly in comparison with Fig. 7 although model size is the same.

will be the corresponding contralateral HRTF, while the same is true for H_{LL} and H_{LR} but for the angle $\alpha + \beta \mod 360°$ (mod stands for the modulo operation). For the system described, the range of β can be from 0 to 90 degrees for both clockwise and counter-clockwise rotation.

We can use matrix notation to represent the loudspeaker-ear system as a two input-two output system in which the two outputs must be processed simultaneously. The following analysis corresponds to the frequency domain. We define H_L as the virtual sound source HRTF for the left ear, H_R as the virtual sound source HRTF for the right ear, H_{RR} , H_{RL}, H_{LL} and H_{LR} as described above, and S as the monaural input sound. Then the signals E_L and E_R at the left and right eardrums respectively should, ideally, be equal with the HRTF-processed monaural sound S_L and S_R (that is the input to the crosstalk canceling system) and are given by

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} S_L \\ S_R \end{bmatrix} = \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \tag{11}$$

The introduction of the contralateral and ipsilateral terms from the physical system (the loudspeakers) will introduce an additional transfer matrix

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} S_L \\ S_R \end{bmatrix} \tag{12}$$

In order to deliver the signals in (11), given that the physical system results in (12), pre-processing must be performed to the binaural input. In particular, the required preprocessing introduces the inverse of the matrix associated with the physical system, as shown below

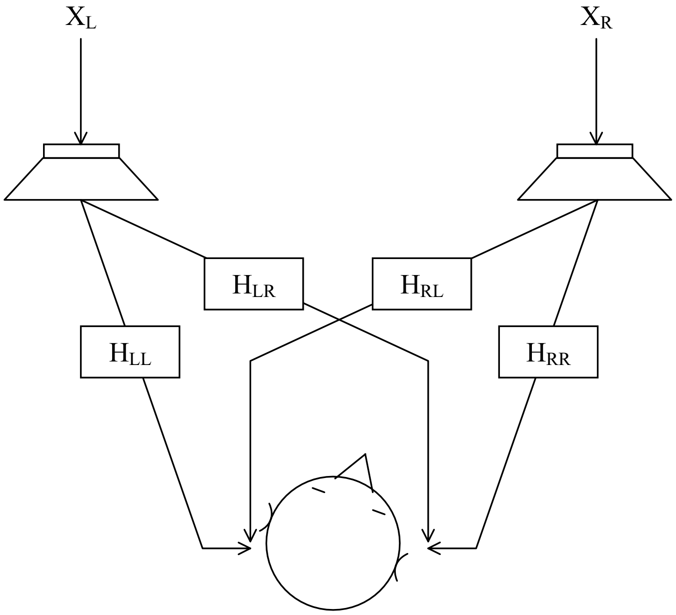

Figure 9: Two loudspeaker-based spatial audio rendering system showing the ipsilateral (H_{LL} and H_{RR}) and contralateral (H_{LR} and H_{RL}) terms for a rotating listener.

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix}^{-1} \begin{bmatrix} S_L \\ S_R \end{bmatrix} \tag{13}$$

It can be seen that equations (11) and (13) are essentially the same. Solving (13) we find

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \frac{1}{H_{RR}H_{LL}} \frac{1}{\left(1 - \frac{H_{RL}}{H_{RR}}\frac{H_{LR}}{H_{LL}}\right)} \begin{bmatrix} H_{RR} & -H_{LR} \\ -H_{RL} & H_{LL} \end{bmatrix} \begin{bmatrix} S_L \\ S_R \end{bmatrix} \tag{14}$$

which can finally be written as

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} 1 & -\frac{H_{LR}}{H_{LL}} \\ -\frac{H_{RL}}{H_{RR}} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{H_{LL}} & 0 \\ 0 & \frac{1}{H_{RR}} \end{bmatrix} \begin{bmatrix} S_L \\ S_R \end{bmatrix} \tag{15}$$

assuming that

$$\frac{1}{\left(1 - \frac{H_{RL}}{H_{RR}}\frac{H_{LR}}{H_{LL}}\right)} \approx 1 \tag{16}$$

This assumption is based on the fact that the contralateral term is of substantially less power than the ipsilateral term because of the shadowing caused by the head.

The terms $1/H_{LL}$ and $1/H_{RR}$ in (15) correspond to the loudspeaker inversion. That is, the HRTF's corresponding to the actual position of the loudspeakers are inverted since they add spectral information that is not in the binaural signal of the virtual source. The matrix

$$\begin{bmatrix} 1 & -\frac{H_{LR}}{H_{LL}} \\ -\frac{H_{RL}}{H_{RR}} & 1 \end{bmatrix}$$

corresponds to the crosstalk cancellation. In the approach described here, the crosstalk cancellation and the inversion of the loudspeakers' response are closely connected, but it is important to note the difference between these two terms. Finally, the signals X_L and X_R that have to be presented to the left and right loudspeaker respectively in order to render the virtual source at the desired location are given by

$$\begin{bmatrix} X_L \\ X_R \end{bmatrix} = \begin{bmatrix} \frac{1}{H_{LL}} & -\frac{H_{LR}}{H_{LL}}\frac{1}{H_{RR}} \\ -\frac{H_{RL}}{H_{RR}}\frac{1}{H_{LL}} & \frac{1}{H_{RR}} \end{bmatrix} \begin{bmatrix} S_L \\ S_R \end{bmatrix} \tag{17}$$

which can be written as

$$\begin{aligned} X_L &= \frac{1}{H_{LL}}S_L - \frac{H_{LR}}{H_{LL}}\frac{1}{H_{RR}}S_R \\ X_L &= \frac{1}{H_{RR}}S_R - \frac{H_{RL}}{H_{RR}}\frac{1}{H_{LL}}S_L \end{aligned} \tag{18}$$

This implies that four different filters should be designed as follows:

$$\begin{aligned} F_L &= \frac{1}{H_{LL}} \\ F_L &= -\frac{H_{RL}}{H_{LL}} \\ F_L &= \frac{1}{H_{RR}} \\ F_L &= -\frac{H_{RL}}{H_{RR}} \end{aligned} \tag{19}$$

The binaural signals pass through these filters, which should obviously form a lattice structure, and then each channel is led to the corresponding loudspeaker.

3.1 Theoretical Analysis

The analysis in the previous sections has shown that crosstalk cancellation requires the implementation of preprocessing filters of the type $H_{\text{inv}} = H_x/H_y$. There are a number of methods for implementing the filter H_{inv}. The most direct method would be to simply divide the two filters in the frequency domain. However, H_y is in general a non-minimum phase filter, and thus the filter H_{inv} designed with this method will be unstable. A usual solution to this problem is to use cepstrum analysis in order to design a new filter with the same magnitude as H_y but being minimum phase [20]. The drawback is that information contained in the excess phase is lost.

In another publication [19] we have introduced an alternative method that maintains the HRTF phase information. The procedure is to find the non-causal but stable impulse response, which also corresponds to H_x/H_y assuming a different Region of Convergence for the transfer function, and then add a delay to make the filter causal. The trade-off and the corresponding challenge is to make the delay small enough to be imperceptible to the listener while maintaining low computational cost. Although this method has been chose because of its advantage of combining computational efficiency and fast convergence, it is still quite demanding for a real-time implementation, especially in the case it has to be combined with other modules, such as head tracking and HRTF modeling. The solution that we present in this paper is based on pre-computing the crosstalk filters for all the possible angles and then use low-rank modeling for the purpose of data reduction and position interpolation.

3.1.1 Low-Rank Modeling

We use Karhunen-Loeve Expansion (KLE) [21] for the purpose of modeling the resulting filters in a low-dimensional space and, additionally, for interpolating between the available listener positions (the listener positions for which the crosstalk filters have been calculated). For this purpose, each crosstalk filter is treated as a vector of measurements and is denoted by $\boldsymbol{h}_j$, where j ranges from 1 to P and P is the number of all the crosstalk filters used for all the desired listener rotation angles (4 filters for each angle).

In KLE, a vector of measurements can be expanded into an orthonormal basis, which actually consists of the eigenvectors of the covariance matrix that describes the measurement process. In the case of multiple vectors that we examine, a usual procedure [22] is to define a time-averaged covariance matrix such as

$$\boldsymbol{R} = \frac{1}{P}\sum_{j-1}^{P}\left(\boldsymbol{h}_j - \boldsymbol{h}_{\mathrm{av}}\right)\left(\boldsymbol{h}_j - \boldsymbol{h}_{\mathrm{av}}\right)^{\mathrm{T}} \tag{20}$$

where,

$$\boldsymbol{h}_{\mathrm{av}} = \frac{1}{P}\sum_{j-1}^{P}\boldsymbol{h}_j \tag{21}$$

is the average vector of all the vectors used. Then, the vector $\boldsymbol{h}_j$ can be represented as an expansion of orthonormal vectors as

$$\boldsymbol{h}_j = \boldsymbol{Q}\boldsymbol{w}_j + \boldsymbol{h}_{\mathrm{av}} \tag{22}$$

In (22) $\boldsymbol{Q}$ is a matrix whose columns are the eigenvectors of $\boldsymbol{R}$, and $\boldsymbol{w}_j$ are the corresponding coefficients, given by

$$\boldsymbol{w}_j = \boldsymbol{Q}^{\mathrm{T}}\left(\boldsymbol{h}_j - \boldsymbol{h}_{\mathrm{av}}\right) \tag{23}$$

For the listener rotation angles that there did not exist any HRTF measurements and, therefore, no crosstalk filters could be designed, it is possible to calculate corresponding coefficients by linearly interpolating between the two closest angles for which coefficients were found by using (23). Additionally, low rank modeling of the crosstalk filters can be used, that is, only K eigenvectors of R can be used corresponding to the K largest eigenvalues ($K < P$). Then $\boldsymbol{Q}$ contains only these K eigenvectors and its dimensions become P by K instead of P by P. This is especially effective for the case that the rest $P - K$ eigenvalues are very close to zero, since the modeling error between the vector $\boldsymbol{h}_j$ and the reconstructed $\boldsymbol{h}_{\boldsymbol{r}j}$ from the low-rank model is analogous to the sum of the $P - K$ eigenvalues that were considered small. As it will be shown in the next paragraph, this is true for the application that we examine and for the specific filters designed only a very small number of eigenvalues were substantially greater than zero.

In section 3.1.2 we describe our findings by showing simulation results of the performance of the filter design method described, by comparing the designed filters with the filters produced by the low-rank modeling method and, finally, by simulating the acoustic path in order to calculate the achieved crosstalk cancellation.

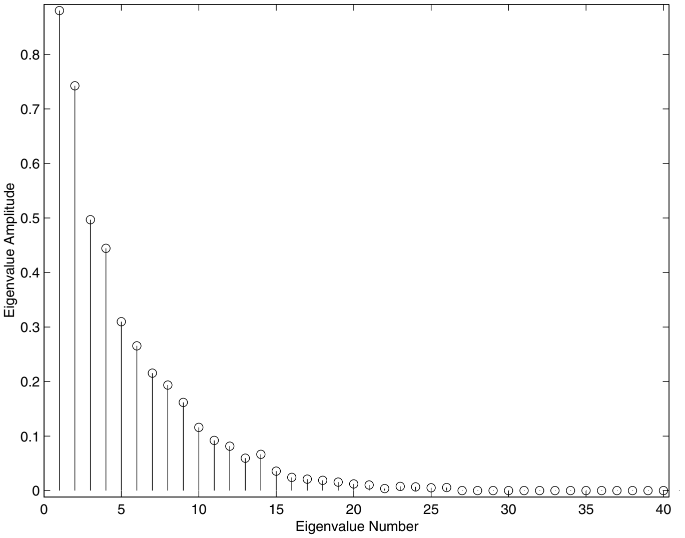

Figure 10: Two loudspeaker-based spatial audio rendering system showing the ipsilateral (H_{LL} and H_{RR}) and contralateral (H_{LR} and H_{RL}) terms for a rotating listener.

3.1.2 Simulation Results

Low-Rank Modeling

The main reason for using KLE was to interpolate between the listener positions that were available. The HRTF's that were used for the results mentioned in this section were the ones made available by MIT. The reason for this was that measurements were available every 5 degrees (assuming that the listener's ears were at the same elevation as the loudspeakers). For a more general application, the given measurements for all different elevation angles could be used (that is, for the case that the listener's level with respect to the loudspeakers is not constant). It is certainly expected that, for this case, KLE would offer a substantial improvement in storage requirements, as well as a means for interpolation. For the simple case we examine here, interpolation is the reason for applying KLE.

Using these measurements, crosstalk filters were created for a listener rotating 180 degrees (90 degrees left to 90 degrees right) for every 5 degrees. The idea, then, was to use KLE so that the listener rotation angles that were not available by measurements could be calculated by means of interpolation, using the available measurements. Instead of using synthetic HRTF's by applying KLE to the available measurements and then designing the required filters, a different approach was followed. The filters for the available angles were designed and then KLE was applied to those filters. This approach is more appropriate since, as mentioned earlier, computational cost required that pre-calculated filters to be used instead of designing the required filters in real-time.

For the listener positions that were mentioned, two different basis were designed for the two different types of filters that were to be designed using [19]. That is, KLE was applied to filers of the type $1/H_i$ and to filters H_c/H_i resulting in two different basis. The reason for this separation was that filters of the same type were quite similar in the time domain, which meant that a smaller number of basis vectors would be required for modeling the filters with the least error. The discussion that follows describes the results obtained for the filters of the type $1/H_i$.

All the filters that were designed were of 2000 taps length and 700 samples of delay. Thus, the covariance matrix $\boldsymbol{R}$ was of dimension 2000-by-2000. From its 2000 eigenvalues, the first 40 are shown in Fig. 10. It is obvious from this figure that except from the first 25 eigenvalues, all the rest can be considered as being approximately zero, resulting in a basis of 25 eigenvectors of 2000 samples each, the eigenvectors of R that correspond to these 25 eigenvalues. By using these 25 eigenvectors, a very low modeling error was achieved. In Fig. 11 one of the

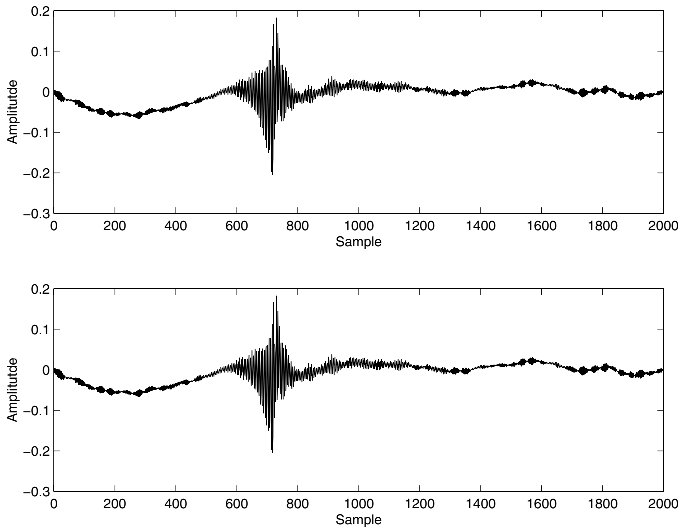

Figure 11: Two loudspeaker-based spatial audio rendering system showing the ipsilateral (H_{LL} and H_{RR}) and contralateral (H_{LR} and H_{RL}) terms for a rotating listener.

filters corresponding to 10 degrees clockwise rotation is shown (upper plot) with the filter that resulted by using the low-rank model (bottom plot). The two filters are obviously very close and their time-domain error of 45 dB (defined in 14) verifies this. In Fig. 12 their error in frequency domain is shown by plotting their normalized error at all frequencies. For a wide range of the frequency band it is obvious that the error is quite acceptable. Again, for frequencies above 15 kHz, the degradation of the error is not considered important.

4 Head Detection and Tracking

4.1 Description of the laser scanning device

For both the HRTF-based rendering and the crosstalk cancellation described above, accurate localization of the listener's ears with respect to the sound source is required. A Head Detection and Tracking (HDT) algorithm is implemented for this purpose. This algorithm uses data provided by a 2D laser radar system LMS 200[1] by SICK Optics Inc. shown in Figure 13, measuring distances of points up to 8 meters in front of the device. Each scan contains data for 180 degrees in azimuth for every 0.5 degrees (total of 361 points). The accuracy of the LMS 200 sensor is approximately 15mm for the range used. The significant advantage of this type of sensor compared to vision-based systems is that it does not depend on the illumination conditions of the room and therefore it can work even in total darkness.

4.2 Experimental setup

The laser scanner is placed such as to face the person sitting in front of the sound system at a distance of approximately 0.5 meters. Since this sensor scans in a single plane its height is appropriately adjusted so that it scans at the level where the primary features to be detected are, i.e the ears and the nose of the listener.

In Figure 14, the surfaces detected during a single scan can be seen. At the bottom of this figure the semi-ellipsoid locus of points close to $(0, 0)$ are the reflections from the head of the person sitting in front of the laser source. The first step of the HDT algorithm is to isolate these points from the reflections of other objects in the surroundings. This selection is performed based on the topology of each group of candidate reflections. A coherent

[1] This sensor complies with Class I eye safety regulations.

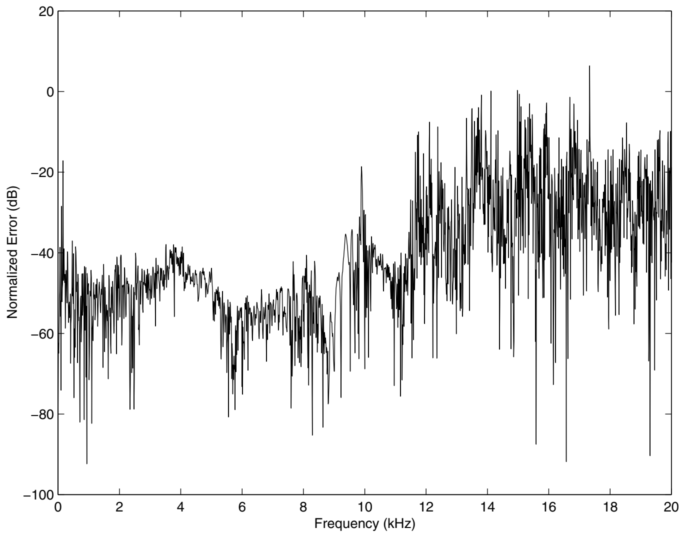

Figure 12: Two loudspeaker-based spatial audio rendering system showing the ipsilateral (H_{LL} and H_{RR}) and contralateral (H_{LR} and H_{RL}) terms for a rotating listener.

locus of points corresponding to the head should form a semi-ellipsoid with primary axes that range between 15 and 30 cm.

At the beginning of the experiment, the person is facing the laser scanner therefore allowing for a measurement of the distance between his/her ears. This is necessary since the HDT algorithm primarily estimates the location of the center of the head (x_c, y_c), defined here as the middle point between the ears, along with the orientation ϕ of the head. The distance between the two ears d_e is required for computing the position of each ear based on these estimates. If $(x_{le}, y_{le)}$ and (x_{re}, y_{re}) are the coordinates of the left and the right ear respectively, these coordinates can be calculated from the following equations:

$$x_{le} = x_c + \frac{d_e}{2} sin(\phi), \qquad y_{le} = y_c - \frac{d_e}{2} cos(\phi) \tag{24}$$

$$x_{re} = x_c - \frac{d_e}{2} sin(\phi), \qquad y_{re} = y_c + \frac{d_e}{2} cos(\phi) \tag{25}$$

As the listener's head rotates, the position of the center of the head fluctuates around its initial location while the orientation of the head varies anywhere from 0 to 180 degrees (90 degrees is the orientation when facing towards the laser scanner). The most prominent feature of the face is the nose of the individual. The HDT algorithm detects the position of its tip during each scanning interval. This information is then processed in order to track the orientation of the head. For this purpose a piecewise constant acceleration model for the rotation of the head is involved within a Kalman filter estimation scheme [1, 23].

4.3 Kalman filter based heading estimation

As we mentioned before, in order to estimate the heading of the person in front of the laser scanner a piecewise constant Wiener process acceleration model is being used for describing the rotation of his/her head. In this model the assumption is that the rotational acceleration is constant during each interval of propagation. This rotational acceleration along with the rotational velocity and heading of the person are constantly estimated based on the measurements provided by the laser scanner. The equations describing this model are:

Figure 13: The LMS 200 laser scanner.

$$x(k+1) = F\,x(k) + G\,w(k) \tag{26}$$

where $x = \left[\phi\ \dot{\phi}\ \ddot{\phi}\right]^T$ is the state vector to be estimated containing the heading ϕ of the head, its rotational velocity $\dot{\phi}$, and its rotational acceleration $\ddot{\phi}$. In this model the Gaussian process noise $w(k)$ is the rotational acceleration increment during the k^{th} sampling period and it is assumed to be a zero-mean white sequence, i.e. the rotational acceleration is a discrete time Wiener process. The system matrix F is given by:

$$F = \begin{bmatrix} 1 & T & \frac{1}{2}T^2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \tag{27}$$

where T is the sampling interval for the laser scanner which in our case it set to 0.2 seconds. The noise gain G is given by:

$$G = \begin{bmatrix} \frac{1}{2}T^2 \\ T \\ 1 \end{bmatrix} \tag{28}$$

Finally, the covariance of the process noise σ_w^2 multiplied by the gain G can be calculated as:

$$Q = G\,\sigma_w^2\,G^T = \begin{bmatrix} \frac{1}{4}T^4 & \frac{1}{2}T^3 & \frac{1}{2}T^2 \\ \frac{1}{2}T^3 & T^2 & T \\ \frac{1}{2}T^2 & T & 1 \end{bmatrix} \sigma_w^2 \tag{29}$$

The measurements provided to this model are the consecutive orientations of the head as these are calculated from the position of the nose of the person with respect to the center of the head. At each time step the nose is

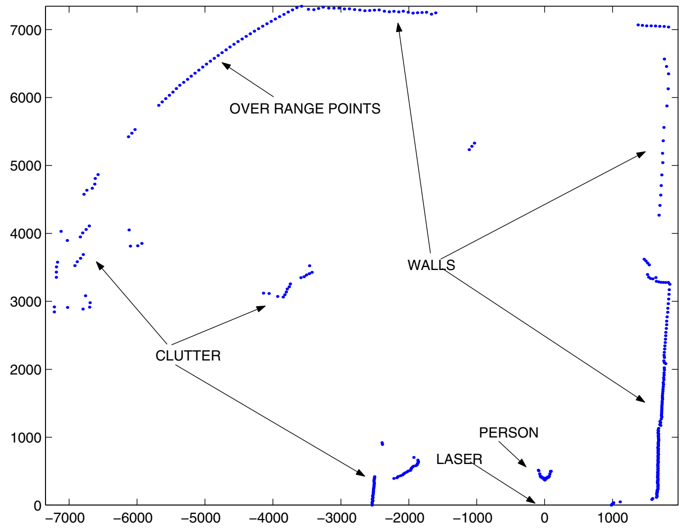

Figure 14: A laser scan of the lab environment where the experiments were contacted.

being detected as the most distant point from the center of the head. In order to avoid false positives especially when the person is facing sideways, the search for the nose feature is limited within ±10 degrees from the location that it was previously found. If the location of the center of the head is (x_c, y_c) and the position of the tip of the nose is (x_n, y_n), then the heading angle is calculated as:

$$z(k+1) = \phi_{measured}(k+1) = \tan^{-1}\left(\frac{y_c - y_n}{x_c - x_n}\right) \tag{30}$$

This measurement is incorporated in the piecewise constant acceleration model according to the following observations equation:

$$z(k+1) = H\ x(k+1) + n(k+1) \tag{31}$$

where H is the observation matrix:

$$H = \left[\begin{array}{ccc} 1 & 0 & 0 \end{array}\right] \tag{32}$$

and $n(k+1)$ is the measurement noise assumed to be a Gaussian white noise process with covariance $R = \sigma_\phi^2$.

At this point we proceed by describing the equations of the Kalman filter ([24], [25]) applied to this estimation problem. The Kalman filter is the minimum mean-square error estimator and its equations can be divided into two sets: Propagation and Update.

4.3.1 Propagation

The previous state of system $\hat{x}(k/k)$ is propagated according to the following equation:

$$\hat{x}(k+1/k) = F\ \hat{x}(k/k) \tag{33}$$

while the corresponding covariance $P(k+1/k)$ for this estimate is given by:

$$P(k+1/k) = F\;P(k/k)\;F^T(k) + Q \tag{34}$$

Every time a new heading measurement is available the update phase of the filter takes place.

4.3.2 Update

During each time step the piecewise constant acceleration model for the rotation of the head provides a prediction for the expected measurement. This is given by the observation model:

$$\hat{z}(k+1/k) = H\;\hat{x}(k+1/k) \tag{35}$$

This predicted measurement is compared to the actual measurement provided by the laser scanner and their difference is the measurement residual (innovation):

$$r(k+1) = z(k+1) - \hat{z}(k+1/k) \tag{36}$$

The covariance of the residual is calculated as:

$$S(k+1) = H\;P(k+1/k)\;H^T + R \tag{37}$$

while the Kalman filter gain is given by:

$$K(k+1) = P(k+1/k)\;H^T\;S^{-1}(k+1) \tag{38}$$

The covariance of the estimate is updated according to:

$$P(k+1/k+1) = P(k+1/k)\;-\;K(k+1)\,S(k+1)\;K^T(k+1) \tag{39}$$

Finally, the state estimate is calculated as:

$$\hat{x}(k+1/k+1) = \hat{x}(k+1/k) + K(k+1)\,r(k+1) \tag{40}$$

The HDT algorithm was tested on different sets of laser scans collected in a lab environment in order to validate its ability to detect and track the facial features of the listener (nose and ears) along with his heading. An example of a laser scan can be seen in Fig. 15. The standard deviation of the error associated with the heading estimate was found to be 4 degrees while the accuracy of the position estimates for the ears was approximately 3 cm.

5 CONCLUSIONS

Spatially rendering of sound sources using a state space model is generally speaking more computationally expensive than an FIR filter, but it provides several advantages over the latter while avoiding some of the disadvantages of IIR filters. One advantage that comes with the use of a state-space model is memory, which eliminates the audible “clicking” noise heard when changing from coefficient to coefficient. In fact, a model with many states eliminates the need for interpolation due to the memory. Interpolation, by passing a signal to two inputs at once, is however desirable to avoid sudden jumps in space of the virtual source. We have also demonstrated that while a

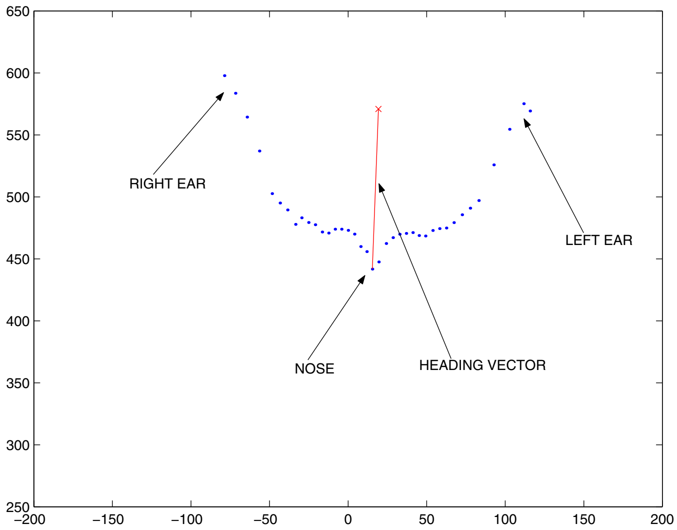

Figure 15: An example of a laser scan showing the points reflected from the face of the listener, the features detected and the heading vector.

single model for the whole space can achieve spatial rendering of multiple sources at once, it can also result in a smaller size than the individual models for all directions combined.

A novel method for generating the required crosstalk cancellation filters as the listener moves was developed based on Low-Rank modeling. Using Karhunen-Loeve expansion we can interpolate among listener positions from a small number of HRTF measurements. A set of corresponding crosstalk cancellation filters was pre-computed for the available (measured) HRTF angles and then KLE was applied to these filters. This approach significantly reduces the required computation time and is more appropriate for integration in a real-time implementation with head tracking. From our results we found that although the filters were designed with 2000 taps, the KL expansion shows that the first 25 eigenvalues are the most important and the rest can be discarded. The resulting modeled transfer functions are shown to have an error better than -45 dB compared to the measured data and so the improvement in computational performance comes at a very small cost in model quality.

Finally a Head Detection and Tracking algorithm was implemented using information provided by 2-D laser scanner. A Kalman filter based estimator was shown to be capable of tracking the location of the listener's ears and the orientation of his head. This information can be passed to both the sound rendering algorithm and the cross-talk canceler to allow spatial rendering of sound in real-time over loudspeakers.

REFERENCES

[1] Y. Akatsuka and G. Bekey, “Compensation for end to end delays in a VR system,” in *Proceedings of the 1998 IEEE Virtual Reality Annual International Symposium*, (Atlanta, GA), pp. 156–159, 14-18 March 1998.

[2] D. Kistler and F. Wightman, “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction,” *Journal of the Acoustical Society of America*, vol. 91, pp. 1637–47, March 1992.

[3] J. Chen, B. D. Van Veen, and K. E. Hecox, “A spatial feature extraction and regularization model for the head-related transfer function,” *Journal of the Acoustical Society of America*, vol. 97, pp. 439–52, January 1995.

[4] S. Kung, “A new identification and model reduction algorithm via singular value decompositions,” *Conference Record of the Twelfth Asilomar Conference on Circuits, Systems and Computers*, pp. 705–14, November 1978.

[5] J. Mackenzie, J. Huopaniemi, V. Valimaki, and I. Kale, “Low-order modeling of head-related transfer functions using balanced model truncation,” *IEEE Signal Processing Letters*, vol. 4, pp. 39–41, February 1997.

[6] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, “Common-acoustical-pole and zero modeling of head-related transfer functions,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 188–96, March 1999.

[7] B. Gardner and K. Martin, “HRTF measurements of a KEMAR dummy-head microphone,” Tech. Rep. 280, MIT Media Lab Perceptual Computing, May 1994. http://sound.media.mit.edu/KEMAR.html.

[8] B. Beliczynski, J. Gryka, and I. Kale, “Critical comparison of hankel-norm optimal approximation and balanced model truncation algorithms as vehicles for fir-to-iir filter order reduction,” *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 3, pp. 593–596, April 1994.

[9] R. Y. Chiang and M. G. Safonov, *Robust Control Toolbox User’s Guide*. The MathWorks, Inc., January 1998. Ver. 2.

[10] The MathWorks, Inc., *Control System Toolbox User’s Guide*. The MathWorks, Inc., January 1999. Fourth Printing.

[11] W. G. Gardner, “Head-tracked 3-D audio using loudspeakers,” in *Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Palz, New York), 1997.

[12] W. G. Gardner, *3-D Audio Using Loudspeakers*. Norwell, Massachusetts: Kluwer Academic Publishers, 1998.

[13] M. R. Schroeder and B. S. Atal, “Computer simulation of sound transmission in rooms,” in *IEEE International Convention Record*, vol. 7, 1963.

[14] P. Damaske and V. Mellert, “A procedure for generating directionally accurate sound images in the upper half-space using two loudspeakers,” in *Acustica*, vol. 22, pp. 154–162, 1969.

[15] P. Damaske, “Head related two channel stereophony with loudspeaker reproduction,” *Journal of the Acoustical Society of America*, vol. 50, pp. 1109–1115, 1971.

[16] J. Bauck and D. H. Cooper, “Generalized transaural stereo and applications,” *Journal of the Audio Engineering Society*, vol. 44, pp. 683–705, 1996.

[17] D. H. Cooper and J. L. Bauck, “Prospects for transaural recording,” *Journal of the Audio Engineering Society*, vol. 37, pp. 3–19, 1989.

[18] W. G. Gardner, “Transaural 3-D audio,” Tech. Rep. 342, MIT Media Lab Perceptual Computing, January/February 1995.

[19] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, “Inverse filter design for immersive audio rendering,” *IEEE Trans. on Multimedia*, vol. 2, p. 77, 2000.

[20] A. V. Oppenheim and R. W. Shafer, *Discrete Time Signal Processing*. Prentice Hall, 1989.

[21] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 3rd edition ed., 1996.

[22] Z. Wu, F. H. Y. Chan, F. K. Lam, and J. C. K. Chan, "A time domain binaural model for the head-related transfer function," *J. Acoust. Soc. Am.*, 1997.

[23] Y. Bar-Shalom and X. Li, *Estimation and Tracking : Principles, Techniques, and Software*. Boston: Artech House, 1993.

[24] S. I. Roumeliotis and G. A. Bekey, "An extended kalman filter for frequent local and infrequent global sensor data fusion," in *Proceedings of the SPIE (Sensor Fusion and Decentralized Control in Autonomous Robotic Systems)*, (Pittsburgh, PA), pp. 11–22, Oct. 1997.

[25] S. I. Roumeliotis, G. S. Sukhatme, and G. A. Bekey, "Circumventing dynamic modeling: Evaluation of the error-state kalman filter applied to mobile robot localization," in *Proceedings of the 1999 IEEE International Conference on Robotics and Automation*, vol. 2, (Detroit, MI), pp. 1656–1663, May 10-15 1999.