Inverse Filter Design for Immersive Audio Rendering Over Loudspeakers

Article *in* IEEE Transactions on Multimedia · July 2000

DOI: 10.1109/6046.845012 · Source: IEEE Xplore

3 authors:

Athanasios Mouchtaris
Foundation for Research and Technology - Hellas
124 PUBLICATIONS **990** CITATIONS

SEE PROFILE

Panagiotis Reveliotis
Cisco Systems, Inc
8 PUBLICATIONS **103** CITATIONS

SEE PROFILE

Chris Kyriakakis
University of Southern California
138 PUBLICATIONS **1,525** CITATIONS

SEE PROFILE

Inverse Filter Design for Immersive Audio Rendering Over Loudspeakers

Athanasios Mouchtaris, Panagiotis Reveliotis, and Chris Kyriakakis, *Member, IEEE*

***Abstract*—Immersive audio systems can be used to render virtual sound sources in three-dimensional (3-D) space around a listener. This is achieved by simulating the head-related transfer function (HRTF) amplitude and phase characteristics using digital filters. In this paper, we examine certain key signal processing considerations in spatial sound rendering over headphones and loudspeakers. We address the problem of crosstalk inherent in loudspeaker rendering and examine two methods for implementing crosstalk cancellation and loudspeaker frequency response inversion in real time. We demonstrate that it is possible to achieve crosstalk cancellation of 30 dB using both methods, but one of the two (the Fast RLS Transversal Filter Method) offers a significant advantage in terms of computational efficiency. Our analysis is easily extendable to nonsymmetric listening positions and moving listeners.**

***Index Terms*—Crosstalk cancellation, head-related transfer function, 3-D audio signal processing.**

I. Introduction

ACCURATE spatial reproduction of sound can significantly enhance the visualization of three-dimensional (3-D) multimedia information particularly for applications in which it is important to achieve sound localization relative to visual images. Such applications include immersive telepresence; augmented and virtual reality for manufacturing and entertainment; air traffic control, pilot warning, and guidance systems; displays for the visually- or aurally- impaired; home entertainment; and distance learning.

Sound perception is based on a multiplicity of cues that include level and time differences, and direction-dependent frequency response effects caused by sound reflection in the outer ear cumulatively referred to as the head-related transfer function (HRTF). The outer ear can be modeled as a linear time-invariant system that is fully characterized by the HRTF in the frequency domain [1].

Using immersive audio techniques it is possible to render virtual sound sources in 3-D space using a set of loudspeakers or headphones (for a review, see [2]). The goal of such systems is to reproduce the same sound pressure level at the listener's eardrums that would be present if a real sound source was placed in the location of the virtual sound source. In order to achieve this, the key characteristics of human sound localization that are based on the spectral information introduced by the head-related transfer function must be considered [3]–[6].

The spectral information provided by the HRTF can be used to implement a set of filters that alter nondirectional (monaural) sound in the same way as the real HRTF. Early attempts in this area were based on analytic calculation of the attenuation and delay caused to the soundfield by the head, assuming a simplified spherical model of the head [7], [8]. More recent methods are based on the measurement of individual or averaged HRTF's for each desired virtual sound source direction [5], [9], [10]. In our implementation we use a pair of HRTF's (one for each ear) that are measured for each desired virtual source direction using a microphone placed in each ear canal of a mannequin (KEMAR). The main advantage of measured HRTF's compared to analytical models is that this method accounts for the pinnae, diffraction from the irregular surface of the human head, and reflections from the upper body.

Several practical problems that arise when attempting to implement digital HRTF filters for immersive audio rendering using headphones or loudspeakers are examined in this paper. In the case of headphone rendering, undesired frequency-dependent distortion is introduced to the binaural signal that is due to anomalies in the headphone frequency response. The inverse filter methods that we present in this paper can be used to remove these frequency response distortions from the headphones.

When rendering immersive audio using loudspeakers, direction dependent spectral information is introduced to the input signal due to the fact that the sound is generated from a specific direction (the direction of the loudspeakers). In addition, just as in the headphones case, the loudspeakers generally do not have an ideal flat frequency response and therefore must be compensated to reduce frequency response distortion. A key issue in loudspeaker-based immersive audio arises from the fact that each ear receives sound from both loudspeakers resulting in undesirable acoustic crosstalk. We examine the relative advantages of two inverse filter methods for crosstalk cancellation and identify one (the Fast RLS Transversal Filtering Method) that is particularly well-suited for real time applications in which the listener may be moving with respect to the loudspeakers. Adaptive inverse filters for traditional stereophonic reproduction have been studied extensively by Nelson *et al.* [11]. In that work, the authors examined the general problem of room inversion, but did not specifically address the problem of HRTF-based rendering. The work presented in this paper is an extension into HRTF-based spatial audio rendering in which the ultimate goal

Manuscript received February 15, 2000; revised March 23, 2000. This work was supported in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, the Annenberg Center for Communication at USC, and the California Trade and Commerce Agency. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jeff A. Bilmes.

The authors are with the Integrated Media Systems Center, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mail: ckyriak@imsc.usc.edu).

Publisher Item Identifier S 1520-9210(00)04705-2.

is to achieve real-time filter synthesis for interactive applications.

In this paper, we refer to monaural sound as nondirectional sound. Binaural sound represents sound that has been recorded with a dummy-head or has been generated through convolution with the appropriate HRTF's for the left and right ears.

The paper is organized as follows. We first formulate the problem mathematically in Section II for both headphone and loudspeaker rendering. The monaural and binaural input cases are treated separately. In Section III, we propose two methods that can be used to address the filter inversion problems that arise due to the nonminimum phase characteristics of the transfer functions involved. Finally, in Section IV, we examine the performance of these two methods by comparing the generated HRTF's to the original measured HRTF's.

II. Problem Specification: Headphone and Loudspeaker Rendering

Binaural methods attempt to accurately reproduce at each eardrum of the listener the sound pressure generated by a set of sources and their interactions with the acoustic environment [12]–[15]. Binaural recordings can be made with specially-designed probe microphones that are inserted in the listener's ear canal, or by using a dummy-head microphone system that is based on average human characteristics. Sound recorded using binaural methods is then reproduced through headphones that deliver the desired sound to each ear. Alternatively, a monaural sound source can be convolved with the HRTF's for a particular azimuth and elevation angle in order to generate binaural sound. It was concluded from early experiments that in order to achieve the desired degree of realism using binaural methods, the required frequency response accuracy of the transfer function was ± 1 dB [16].

When headphones are used for immersive audio rendering, their frequency response is included in the frequency response of the signal that reaches the eardrums. Ideally, a filter that inverts the frequency response of the headphones is required so that the monaural signal will be convolved not only with the HRTF's of the virtual source, but also with this filter. In the frequency domain, if H_p is the frequency response of the headphones and H_x the HRTF for a specific direction and ear, the inversion of the headphones' response can be accomplished in two ways, depending on whether the input to the designed filter is monaural or binaural sound. In the monaural input case we design the inverse filter $H_{inv} = H_x/H_p$. The monaural signal (S) is processed by this filter and then by the headphones' transfer function, so the response S_e at the eardrum will be

$$S_e = H_p H_{inv} S = H_p \frac{H_x}{H_p} S = H_x S \tag{1}$$

which is exactly the desired response (S is the monaural signal to be spatialized).

Alternatively, a filter can be designed whose input is the binaural signal S_b that already contains the required HRTF information (i.e., $S_b = H_x S$). In this case, it is simply necessary to invert the response of the headphones and so the response of the designed filter should be

$$H_{inv} = \frac{1}{H_p}. \tag{2}$$

Then, the signal at the eardrum S_e will be

$$S_e = H_p H_{inv} S_b = H_p \frac{1}{H_p} S_b = H_x S. \tag{3}$$

A number of methods exist for implementing the filter H_{inv}. We will discuss two of these in a later section of this paper.

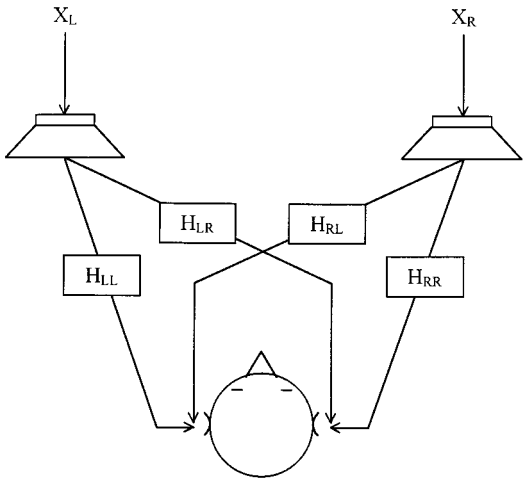

Fig. 1. Loudspeaker-based spatial audio rendering system showing the ipsilateral (H_{LL} and H_{RR}) and contralateral (H_{LR} and H_{RL}) terms.

Loudspeakers can also be used to render binaural or HRTF-processed monaural sound. In order, however, to deliver the appropriate binaural sound field to each ear 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. 1), 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. One of the key limitations of crosstalk cancellation systems arises from the fact that any listener movement that exceeds 75–100 mm completely destroys the desired spatial effect. This limitation can be overcome by tracking of the listener's head in 3-D 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 [17], [18]. A camera-based system that does not require the user to be tethered has been demonstrated for stereophonic reproduction [19], [20].

Several schemes have been proposed to address crosstalk cancellation. The first such scheme was proposed by Atal and

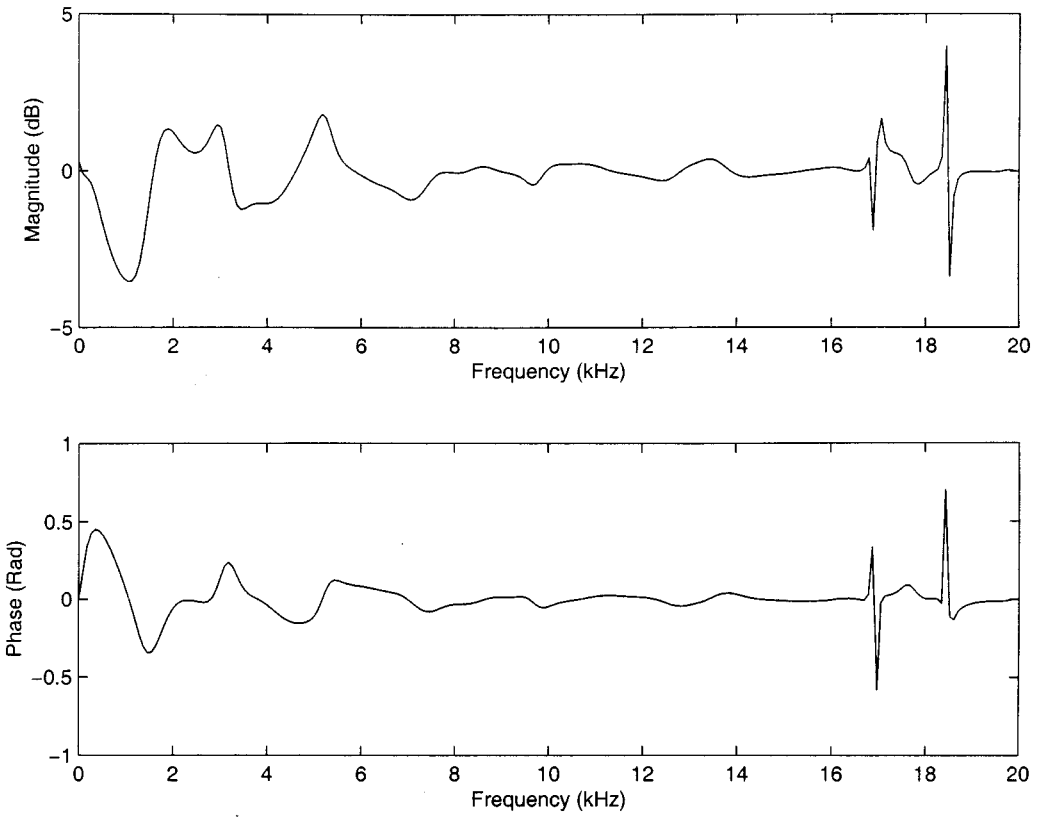

Fig. 2. Magnitude and phase response of the term $(1-(H_c/H_i)^2)^{-1}$ that is extracted as a common factor from the matrix product that describes the physical system. The assumption that the term is approximately of all-pass response is valid.

Schroeder [21] and later another was published by Damaske and Mellert [16], [22]. A method proposed by Cooper and Bauck modeled the head as a sphere and then calculated the ipsilateral and contralateral terms [23], [24]. 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) [25].

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 [23], [24]. This solution can *only* work for a nonmoving listener seated symmetrically to the loudspeakers. In this paper, we use a different approach for the analysis that can be easily generalized to the nonsymmetric case that arises when the listener is moving. While in our analysis we present the symmetric case to make the notation simpler, the methods that we propose are also valid for the nonsymmetric case. A video-based head-tracking algorithm has been developed in which the listener is tracked and the filters are computed in real time in response to changes in the listener's position [2], [19], [20]. The motivation behind the methods presented in this paper is the ability to achieve real-time performance so that the necessary filters can be calculated at each listener position.

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. In the frequency domain we define H_i as the ipsilateral term, H_c as the contralateral term, H_L as the virtual sound source HRTF for the left ear, H_R as the virtual sound source HRTF for the right ear, and S as the monaural input sound. Then the signals E_L and E_R at the left and right eardrums, respectively, are given by

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix}. \tag{4}$$

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

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix}. \tag{5}$$

In order to deliver the signals in (4), given that the physical system results in (5), preprocessing must be performed to the input S. In particular, the required preprocessing introduces the inverse of the matrix associated with the physical system, as shown below

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix}^{-1} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix}. \tag{6}$$

It can be seen that (4) and (6) are essentially the same. Solving (6) we find

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \frac{1}{H_i^2} \frac{1}{\left(1 - \frac{H_c^2}{H_i^2}\right)} \begin{bmatrix} H_i & -H_c \\ -H_c & H_i \end{bmatrix} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \tag{7}$$

which can finally be written as

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} 1 & -\frac{H_c}{H_i} \\ -\frac{H_c}{H_i} & 1 \end{bmatrix} \begin{bmatrix} \frac{H_L}{H_i} & 0 \\ 0 & \frac{H_R}{H_i} \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \tag{8}$$

assuming that

$$\frac{1}{\left(1 - \frac{H_c^2}{H_i^2}\right)} \cong 1. \tag{9}$$

This assumption is based on the fact that the contralateral term is of substantially less power that the ipsilateral term because of the shadowing caused by the head. The validity of this assumption was examined by plotting the magnitude and phase of the term in (9) and comparing them with the corresponding magnitude and phase of an all-pass filter. The term in (9) is plotted in Fig. 2 for a set of measured HRTF's. It can be seen that, indeed, this term can be considered to be of approximately all-pass response.

The terms H_L/H_i and H_R/H_i in (8) 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

$$\begin{bmatrix} 1 & -\frac{H_c}{H_i} \\ -\frac{H_c}{H_i} & 1 \end{bmatrix}$$

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

$$\begin{bmatrix} X_L \\ X_R \end{bmatrix} = \begin{bmatrix} \frac{H_L}{H_i} & -\frac{H_c}{H_i}\frac{H_R}{H_i} \\ -\frac{H_c}{H_i}\frac{H_L}{H_i} & \frac{H_R}{H_i} \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \tag{10}$$

which can be written as

$$\begin{aligned} X_L &= \left(\frac{H_L}{H_i} - \frac{H_c}{H_i}\frac{H_R}{H_i}\right) S \\ X_R &= \left(\frac{H_R}{H_i} - \frac{H_c}{H_i}\frac{H_L}{H_i}\right) S. \end{aligned} \tag{11}$$

This implies that the filters F_L and F_R for the left and right channel should be

$$\begin{aligned} F_L &= \frac{H_L}{H_i} - \frac{H_c}{H_i}\frac{H_R}{H_i} \\ F_R &= \frac{H_R}{H_i} - \frac{H_c}{H_i}\frac{H_L}{H_i}. \end{aligned} \tag{12}$$

The monaural signal S passes through these filters and then each channel is led to the corresponding loudspeaker.

Similarly to the headphones inversion case described earlier, a filter can be designed for the case that the input is the binaural signal S_b instead of the monaural S. In this case, convolution with the pair of HRTF's H_L and H_R is not needed since the binaural signal already contains the directional HRTF information. For the binaural case, the matrix

$$\begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix}$$

is substituted in (7) by the identity matrix.

III. Theoretical Analysis

The analysis in the previous sections has shown that inversion of the headphones response, crosstalk cancellation, and loudspeaker HRTF inversion, all require the implementation of preprocessing filters of the type $H_{inv} = H_x/H_y$, in which H_x is 1, H_L, H_R or H_c and H_y is the headphones response H_p or the ipsilateral response H_i. 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 nonminimum 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 [26]. The drawback is that information contained in the excess phase is lost.

Here, we propose a different procedure that maintains the HRTF phase information. The procedure is to find the noncausal 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. We describe below two methods for finding this noncausal solution.

A. Least Mean Squares (LMS) Filter Design Method

Based on the previous discussion and taking into consideration the need for adding a delay in order for the preprocessing filter to be feasible (i.e., causal), we conclude that the relationship between the filters H_y, H_x and the preprocessing filter H_{inv} can be depicted as in the block diagram shown in Fig. 3.

The problem of defining the filter H_{inv} such that the mean squared error between $y(n)$ and $d(n)$ is minimized, can be classified as a combination of a system identification problem (with respect to H_x) and inverse modeling problem (with respect to H_y) and its solution can be based on standard adaptive methods such as the LMS algorithm [27]. More specifically, the taps

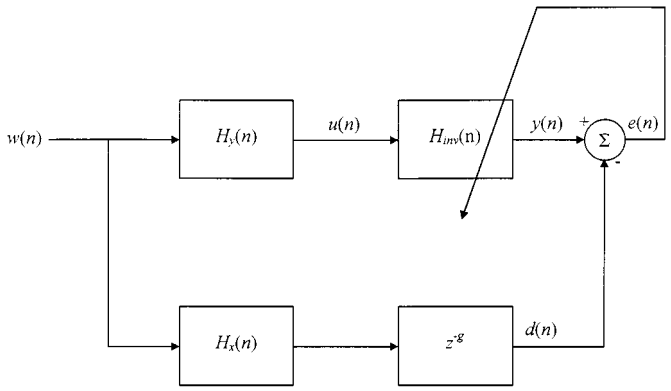

Fig. 3. Block diagram for the algorithm used to estimate of the inverse filter. The problem of finding the filter H_{inv} such that the mean squared error between $y(n)$ and $d(n)$ is minimized, is a combinations of a system identification problem (with respect to H_x) and inverse modeling problem (with respect to H_y) and its solution can be based on standard adaptive methods.

of the filter H_{inv} at iteration n can be computed based on the weight adaptation formula

$$\boldsymbol{h}_{inv}(n+1) = \boldsymbol{h}_{inv}(n) + \mu \boldsymbol{u}(n) e(n) \tag{13}$$

in which

$$e(n) = d(n) - \boldsymbol{h}_{inv}^{H}(n)\boldsymbol{u}(n). \tag{14}$$

In (14), H denotes the Hermitian of the vector $\boldsymbol{h}_{inv}$. The desired response $d(n)$ can be found from Fig. 3 to be

$$d(n) = \boldsymbol{h}_{x}^{H}(n)\boldsymbol{w}(n-g). \tag{15}$$

The notation $\boldsymbol{u}(n)$ denotes a vector of samples arranged as

$$\boldsymbol{u}(n) = [u(n) \quad u(n-1) \quad \cdots \quad u(n-M+1)]^{T} \tag{16}$$

where M is the order of the filter $\boldsymbol{h}_{inv}$. This is also true for $\boldsymbol{w}(n)$. The system input $\boldsymbol{w}(n)$ can be chosen arbitrarily, but a usual practice for system identification problems is to use white noise as the input. The reason is that white noise has an all-pass frequency response so that all frequencies are weighted equally during the adaptation procedure.

The filter length M, as well as the delay g, can be selected based on the minimization of the mean squared error. In this paper we used a variation of the LMS (the Normalized LMS) with a progressive adaptation (decrement) of the step size μ that results in faster convergence as well as smaller misadjustment. The step size μ changes at every iteration, using the update formula

$$\mu(n) = \frac{\beta}{\alpha + ||\boldsymbol{u}(n)||^2}. \tag{17}$$

In (17), β is a positive constant, usually less than two, and α is a small positive constant [27].

The resulting filter from this method is $\boldsymbol{h}_{inv}$, which in the frequency domain is equal to H_x/H_y. If the desired output is of the form $1/H_y$, (in the binaural case), $\boldsymbol{h}_x$ can be chosen to be the impulse sequence. The result in either case is an FIR filter.

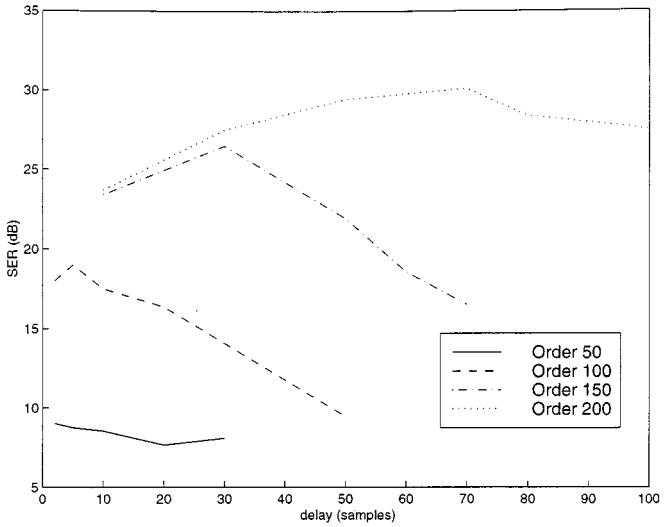

Fig. 4. SER for several choices of delay and filter order, using the LMS method. The lowest order that gives an SER of 30 dB is 200 with a corresponding optimum delay of 70 samples.

B. Least-Squares Filter Design Method

Referring again to Fig. 3, another way of approaching the problem is to minimize the sum of squared errors $e(n)$ (instead of the mean squared error as in the LMS method)

$$\min_{h_{inv}(m)} \sum_{n=M}^{N} \left| \sum_{m=0}^{M} u(n-m)h_{inv}(m) - d(n) \right|^2. \tag{18}$$

The above equation can be rewritten in matrix notation as

$$\min_{\boldsymbol{h}_{inv}} ||\boldsymbol{H}\boldsymbol{h}_{inv} - \boldsymbol{h}_x||^2 \tag{19}$$

in which $\boldsymbol{H}$ is a rectangular Toeplitz matrix that can be easily derived from (18). The solution to (19) in the least-squares sense is

$$\boldsymbol{h}_{inv} = \boldsymbol{H}^{+}\boldsymbol{h}_x \tag{20}$$

in which we denote the pseudoinverse of $\boldsymbol{H}$ as $\boldsymbol{H}^{+}$. In general, (19) describes an overdetermined system for which $\boldsymbol{H}^{+}$ in (20) can be written as

$$\boldsymbol{H}^{+} = (\boldsymbol{H}^{H}\boldsymbol{H})^{-1}\boldsymbol{H}^{H}. \tag{21}$$

We denote $\boldsymbol{P} = \boldsymbol{H}^{H}\boldsymbol{H}$ which can be viewed as the time-averaged correlation matrix. The calculation of the pseudoinverse is a computationally demanding operation that is not suitable for real-time implementations. One way to overcome this problem is by calculating the pseudoinverse recursively. Specifically, we calculate the inverse of $\boldsymbol{P}$ recursively, using the well-known matrix inversion lemma. This method is known as recursive least-squares (RLS). The advantage of this method is that for most problems it requires M iterations for convergence, where M is the order of the designed filter. On the other hand, LMS usually requires a higher number of iterations for convergence. The number of iterations is a very important issue for real-time implementations, but equally important is the computational complexity of the algorithm (measured in number of multiplies and

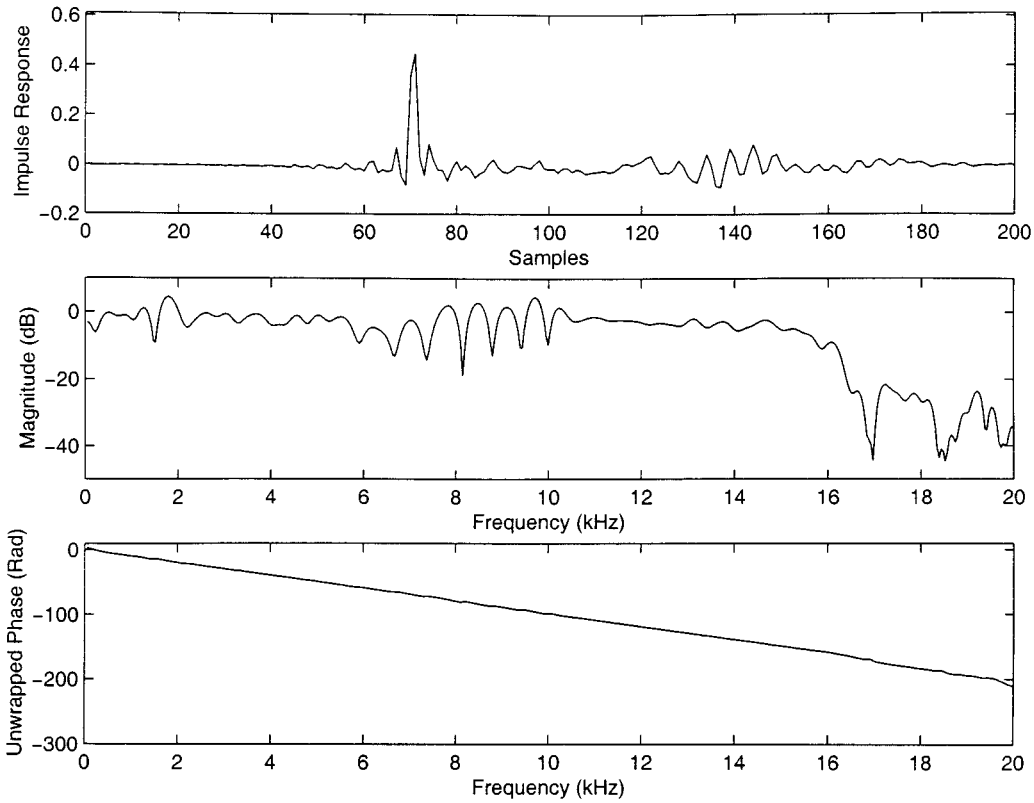

Fig. 5. Impulse response (top), magnitude response (middle) and phase response (bottom) of the designed filter H_{inv} using the LMS method.

divides for adaptive systems). Here LMS has a great advantage, requiring only $3M$ operations per iteration whereas RLS requires M^2. This problem of the RLS algorithm has motivated a lot of research to find efficient implementations with reduced computational complexity. In this paper we implemented the FTF method for RLS proposed by Cioffi and Kailath [28]. This algorithm requires $7M$ computations per iteration while it retains the fast convergence property of the RLS algorithm, thus it is highly suitable for real-time implementations. The FTF algorithm decouples the recursive calculation of the inverse matrix of $\boldsymbol{P}$ into a recursive calculation of three vectors $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, which is a procedure that requires fewer computations, since no matrix multiplication is involved.

In Section IV, we describe our findings and show that the FTF algorithm has a significant advantage over the LMS algorithm in terms of convergence rate while incurring only a moderate increase in computational complexity.

IV. Simulation Results

A. Loudspeaker Inversion

All of the filters that are of the form H_x/H_y were designed using both the LMS and least-squares methods. As discussed above, a delay is introduced to the system to satisfy causality. The coefficients of these FIR filters were designed using Matlab. The delays and lengths for the filters used were optimized to achieve maximum signal-to-error power ratio (SER) in the time domain between the filter $H_{inv}H_y$ (which we will call the *cascade* filter) and H_x. In our case, the SER is defined by

$$\frac{\sum_{k=1}^{N} h_x^2(k)}{\sum_{k=1}^{N} (h_x(k) - h_{ca}(k))^2} \tag{22}$$

in which h_{ca} is the impulse response of the cascade filter.

It is important to evaluate the error in the time-domain because a good approximation is required both in the magnitude and phase responses. Both methods worked successfully with a number of different measured HRTF's corresponding to 128 tap filters. The following simulation results were found using the 0° azimuth and 0° elevation measured HRTF of length 128 taps, corresponding to the term H_x. The HRTF measurements in this paper were performed using a KEMAR dummy-head with Etymotic Research microphones. The playback system consisted of two TMH Corp. loudspeakers placed on a table so that the center of each loudspeaker was at the same height as the center of the KEMAR pinnae for on-axis measurements. The loudspeakers spacing was 50 cm and the center of the KEMAR's head was 50 cm from the center point of the loudspeaker baffle plane. The room in which the measurements were performed has dimensions 8.5 m (L) $\times$ 7.0 m (W) $\times$ 3.5 m (H) and the reverberation time was measured using the THX R2 spectrum analyzer and found to be 0.5 s from 125 Hz to 4 kHz.

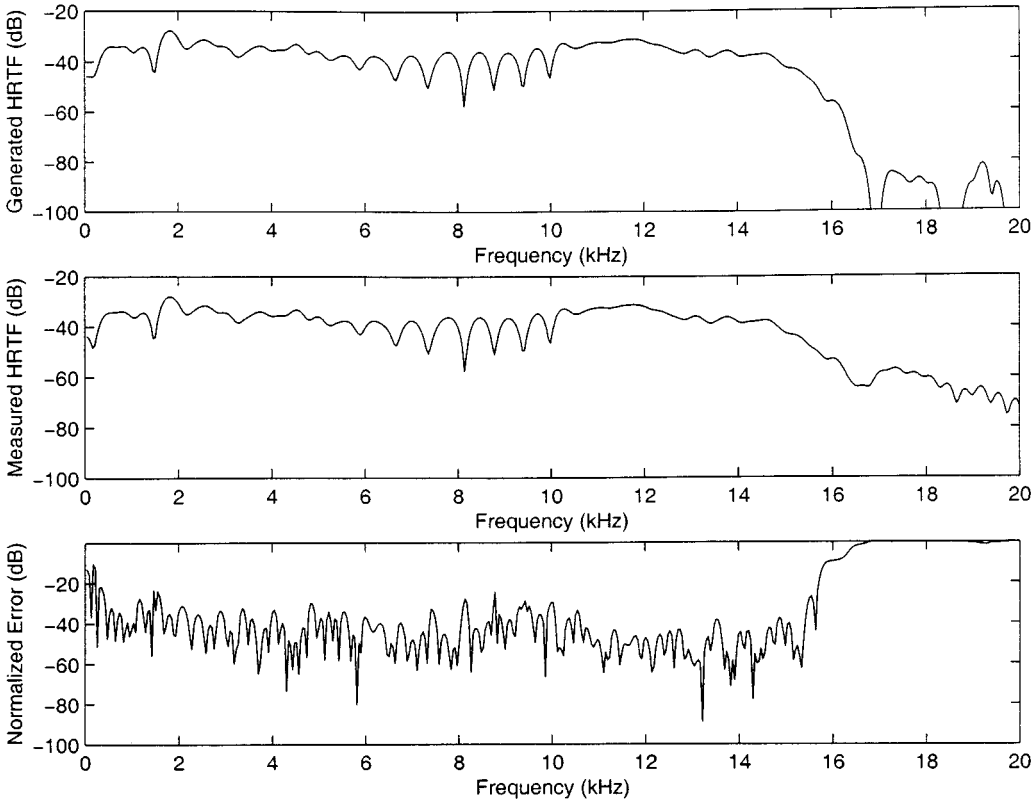

Fig. 6. The HRTF generated from the inverse filter using the LMS method is shown in the upper plot. The measured HRTF (0° azimuth and 0° elevation) is shown in the middle and the relative error between the two is shown in the bottom plot.

For the monaural input case, an inverse filter of 200 taps was designed, that introduced a delay of 70 samples (1.6 ms at a sampling rate of 44.1 kHz). These were the optimum values of filter length and delay that gave rise to an SER of better than 30 dB. The tradeoffs in SER, filter length, and delay are shown in Figs. 4 and 7 for the LMS and least-squares (RLS) methods, respectively. It is interesting to note that the optimal choices of filter order and delay are the same for both methods. The filter order can, of course, be chosen arbitrarily, but we found that for a given order, the corresponding delay is the same for both methods. The SER in the time domain for this case was 30.3 dB for the LMS method and 31.5 dB for the least-squares method. The results for the LMS method can be seen in Figs. 5 and 6. In Fig. 5 the resulting filter H_{inv} is plotted in both the time and frequency domains. In Fig. 6, a comparison is made between the magnitude of the measured HRTF and the HRTF generated using our inverse filter. Because the approximation of the two filters is made in the time domain, it was expected that their phase responses would be practically identical. The same plots are shown in Figs. 8 and 9 for the least-squares case. The required number of iterations for the two algorithms is in agreement with what was mentioned in Section III. The LMS algorithm required 5000 iterations in order to reach the 30 dB SER criterion, while the least-squares method required only 500 iterations for the same error. This result, along with the relatively small increase in computational requirements of the FTF algorithm, justifies the claim that this method is highly suitable for a real-time implementation in which the filter parameters are updated in response to head-tracking information.

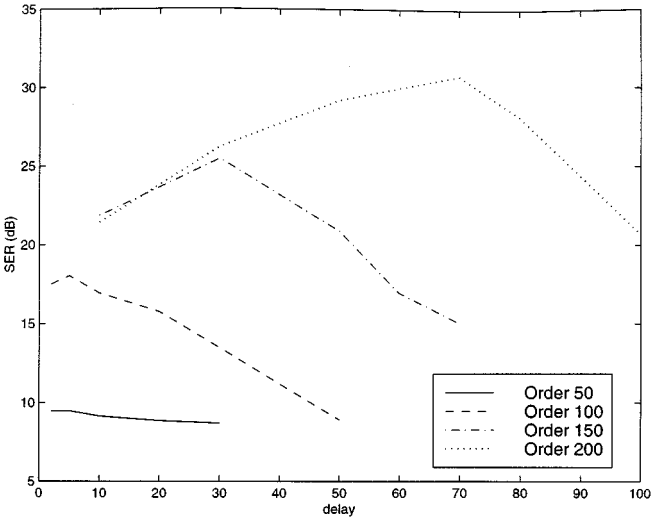

Fig. 7. SER for several choices of delay and filter order, using the least-squares method. As in the LMS case, the lowest order that gives an SER of 30 dB is 200 with corresponding delay of 70 samples.

It should be noted that for frequencies above 15 kHz, the associated wavelengths are less than 20 mm. In this range it is practically impossible to accurately place the listener's ears in the desired location for which the filters have been designed. For this reason the degradation of the normalized error above 15 kHz (as seen in Figs. 6 and 9) is acceptable since listener position errors will dominate.

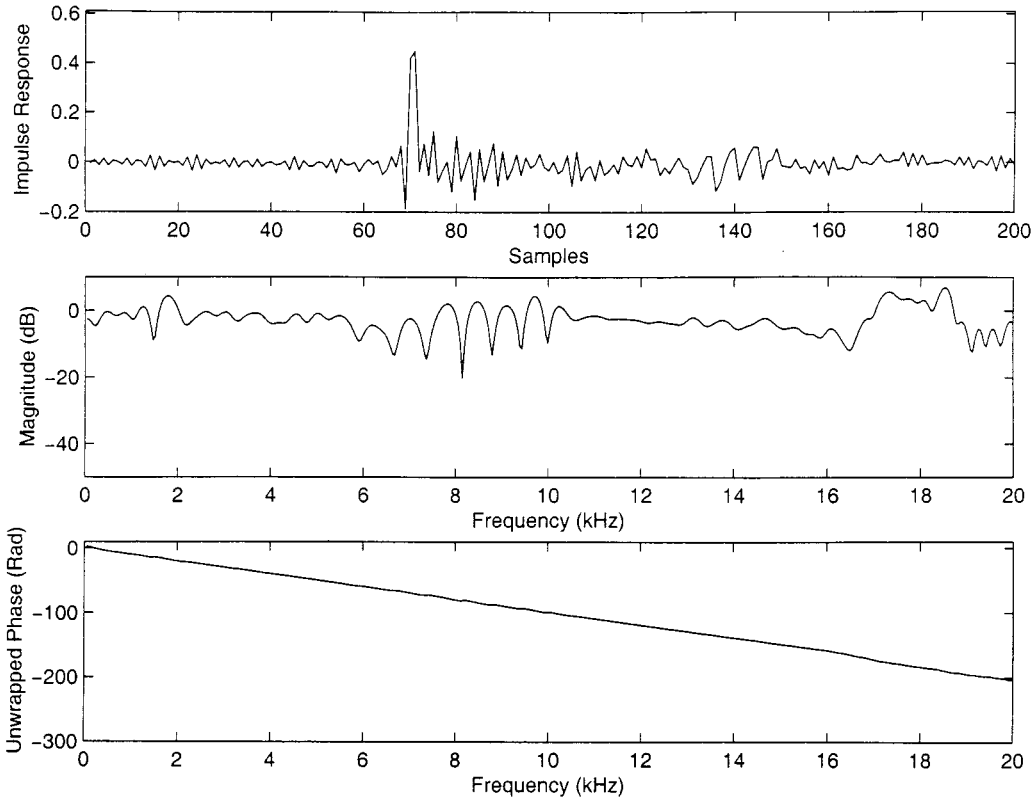

Fig. 8. (Top) Impulse response, (middle) magnitude response, and (bottom) phase response of the designed filter H_{inv} using the least-squares method.

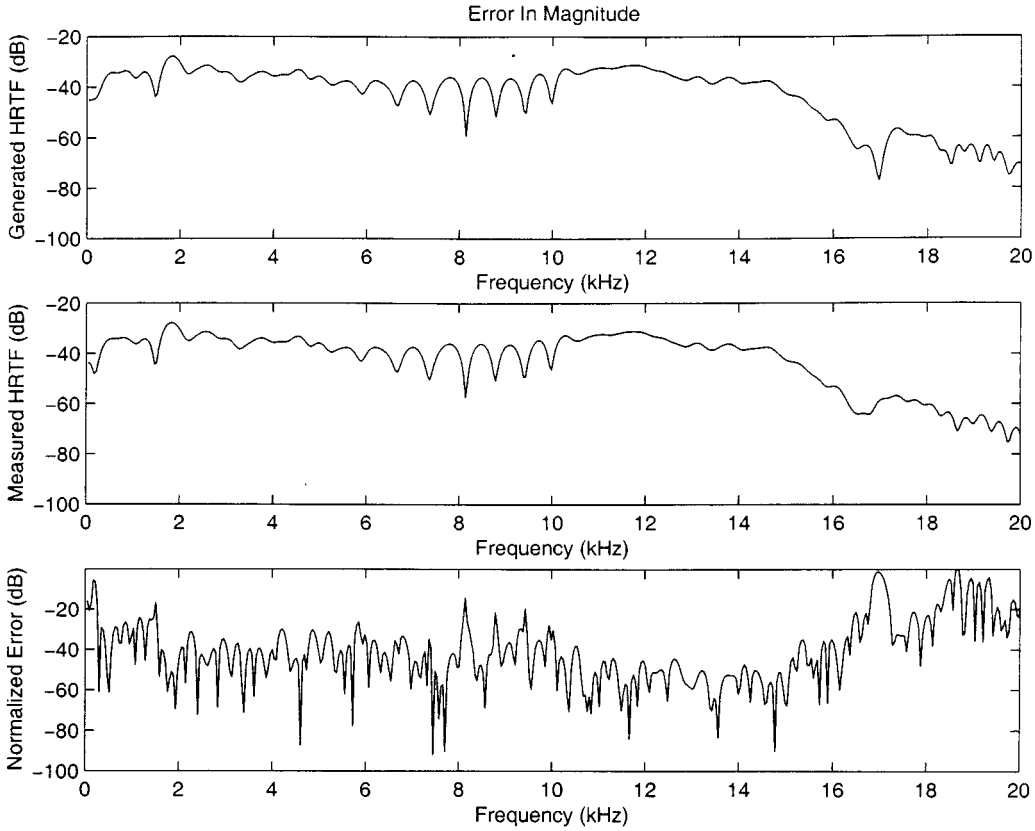

Fig. 9. The HRTF generated from the inverse filter using the least-squares method is shown in the upper plot. The measured HRTF (0° azimuth and 0° elevation) is shown in the middle and the relative error in the bottom plot.

If inversion of the type $1/H_i$ is required (binaural input), the cascade filter should be of exactly all-pass response. This case proved to be more demanding than the monaural input case. In order to get the desired SER of 30 dB in the time domain we had to increase the filter length to 400 taps (with a corresponding delay of 160 samples). Alternatively, it is possible to design a filter of the form of H_a/H_i where H_a has an all-pass response up to 15 kHz. Using this approximation, we were able to achieve the 30 dB requirement in SER with a filter length of 200 taps and a delay of 70 samples. In listening tests there was no perceptible difference in using this method compared to the full spectrum all-pass.

B. Crosstalk Cancellation

If we denote in the upper equation (12) the delay introduced by H_c/H_i as d_1 and the delay introduced by H_R/H_i as d_2 then, in the z-domain, we find that the filter can be written as

$$F_L = \frac{H_L}{H_i} z^{-(d_1+d_2)} - \frac{H_c}{H_i} z^{-d_1} \frac{H_R}{H_i} z^{-d_2}. \qquad (23)$$

Note that the delay for H_L/H_i in (23) must be equal to the sum of d_1 and d_2. The delay introduced by the filter F_R should also be equal to $d_1 + d_2$. In the time domain (23) becomes

$$\begin{aligned} \boldsymbol{f}_l &= \boldsymbol{h}_{li} - \boldsymbol{h}_{ci} * \boldsymbol{h}_{ri} \\ \boldsymbol{f}_r &= \boldsymbol{h}_{ri} - \boldsymbol{h}_{ci} * \boldsymbol{h}_{li} \end{aligned} \qquad (24)$$

in which $*$ denotes convolution.

In order to design the filter for each channel, each of the three filters $\boldsymbol{h}_{li}$, $\boldsymbol{h}_{ci}$ and $\boldsymbol{h}_{ri}$ can be designed separately, and then be combined using (24) to obtain the desired final filter. This method is preferable when H_L, H_c, and H_R are given in the time domain (e.g., from a measurement). In this case note that the delay introduced by $\boldsymbol{h}_{li}$ in $\boldsymbol{f}_l$ is $d_1 + d_2$ while in $\boldsymbol{f}_r$ it is d_2. A similar argument holds for $\boldsymbol{h}_{ri}$. This means that the filters $\boldsymbol{h}_{li}$ and $\boldsymbol{h}_{ri}$ required for $\boldsymbol{f}_l$ will be different from the filters $\boldsymbol{h}_{li}$ and $\boldsymbol{h}_{ri}$ required for $\boldsymbol{f}_r$. Only the filter $\boldsymbol{h}_{ci}$ can be the same. Also, filter lengths should be chosen accordingly, since convolution of two filters with lengths l and p results in a filter with length $l + p - 1$ and in order to subtract two filters they should be of the same length.

An interesting test of the performance of the methods described is to measure the crosstalk cancellation that is achieved. That is, when both loudspeakers produce sound, the sound pressure level at the contralateral ear must be very low compared with the sound pressure level at the ipsilateral ear. A certain degree of crosstalk cancellation is achieved even with no filtering due to the head shadowing, particularly at higher frequencies (Fig. 10). Toole [29], [30] and Walker [31] studied the psychoacoustic effects of early reflections and in small rooms found that in order to remain inaudible they must be at least 15 dB below the direct sound in spectrum level. A successful crosstalk cancellation scheme should therefore result in at least a 15 dB attenuation of the crosstalk term.

For the symmetric positioning of the listener that we have examined, we saw that for the binaural input case we can set

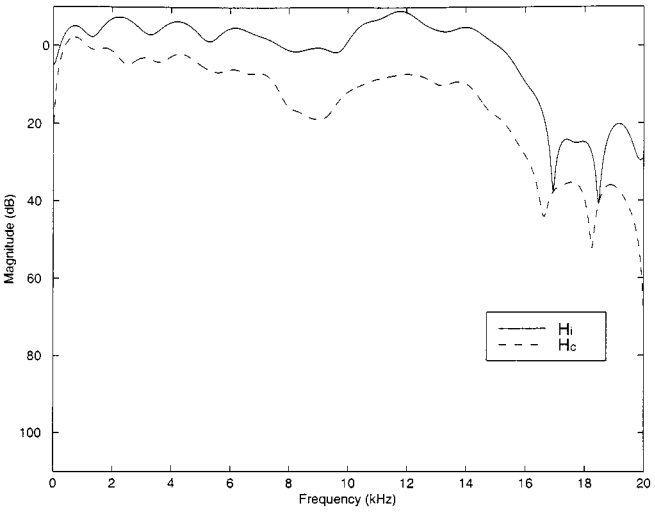

Fig. 10. The difference in dB between the ipsilateral (H_i) and the contralateral (H_c) terms shows the effect of head shadowing with no crosstalk cancellation. In this setup the loudspeakers were 50 cm apart and the head was located in the symmetric (center) position at a distance of 50 cm from the loudspeaker baffle plane.

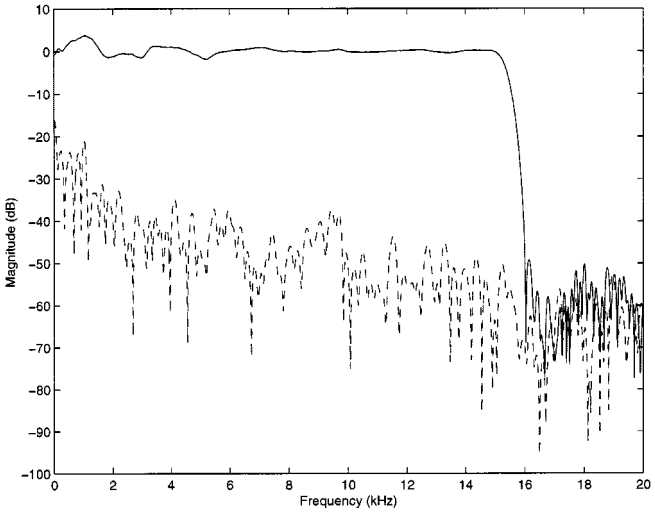

Fig. 11. Measured HRTF data from the loudspeakers (H_i and H_c) were used to simulate the physical system and design a set of filters to eliminate the crosstalk. The resulting diagonal (solid line) and off-diagonal (dotted line) terms produced by our simulation using the LMS method are plotted above. The diagonal term is very close to 1 (0 dB) from 2 to 15 kHz and deviates only slightly in the region below 1 kHz. The off-diagonal term starts at -15 dB and remains below -30 dB from 1 to 15 kHz.

$H_L = H_R = 1$ in (8) since the virtual source HRTF's are already contained in the binaural signal. Then, (8) becomes

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} 1 & -\frac{H_c}{H_i} \\ -\frac{H_c}{H_i} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{H_i} & 0 \\ 0 & \frac{1}{H_i} \end{bmatrix} \begin{bmatrix} S_{b1} \\ S_{b2} \end{bmatrix} \qquad (25)$$

in which ideally $E_L = S_{b1}$ and $E_R = S_{b2}$. If we define the filters $F_{ii} = 1/H_i$ and $F_{ci} = -H_c/H_i^2$, then (25) can be written as

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} F_{ii} & F_{ci} \\ F_{ci} & F_{ii} \end{bmatrix} \begin{bmatrix} S_{b1} \\ S_{b2} \end{bmatrix} \qquad (26)$$

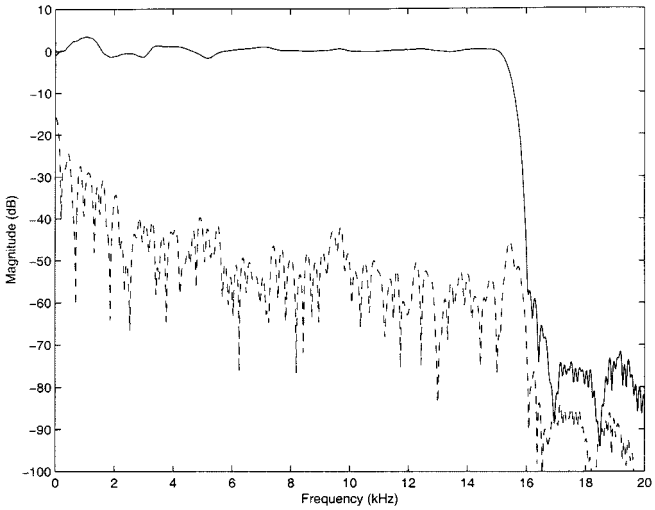

Fig. 12. Measured HRTF data from the loudspeakers (H_i and H_c) were used to simulate the physical system and design a set of filters to eliminate the crosstalk. The resulting diagonal (solid line) and off-diagonal (dotted line) terms produced by our simulation using the Least Squares method are plotted above. The diagonal term is very close to 1 (0 dB) from 2–15 kHz and deviates only slightly in the region below 1 kHz. The off-diagonal term starts at -15 dB and remains below -30 dB from 1–15 kHz.

which finally becomes

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_iF_{ii} + H_cF_{ci} & H_iF_{ci} + H_cF_{ii} \\ H_iF_{ci} + H_cF_{ii} & H_iF_{ii} + H_cF_{ci} \end{bmatrix} \begin{bmatrix} S_{b1} \\ S_{b2} \end{bmatrix}. \quad (27)$$

In order to deliver the desired binaural signal to each ear (i.e., $E_L = S_{b1}$ and $E_R = S_{b2}$) the diagonal terms $H_iF_{ii} + H_cF_{ci}$ must be 1 (this would mean that the loudspeaker frequency response inversion has succeeded) and the off-diagonal term $H_iF_{ci} + H_cF_{ii}$ must be 0 (this would mean that the crosstalk cancellation has succeeded).

We designed the filters F_{ii} and F_{ci} using both LMS and least-squares methods. For the LMS method, we designed the filter $\boldsymbol{f}_{ii}$ using a length of 349 taps, introducing a delay of 140 samples and an SER of 44.1 dB. For the filter $\boldsymbol{f}_{ci}$ we designed a filter of 150 taps length, delay of 70 samples and a resulting SER of 31.4 dB with frequency response H_c/H_i, and a filter of 200 taps length, delay of 70 samples and SER of 31.6 dB with frequency response $1/H_i$, and then convolved their time domain responses. As mentioned earlier, this procedure is preferable when the HRTF's are given in the time domain. We used the measured HRTF data from the loudspeakers (H_i and H_c) to simulate the physical system and designed a set of filters to eliminate the crosstalk. The resulting diagonal and off-diagonal terms produced by our simulation are plotted in Fig. 11, in which the diagonal term is plotted as a solid line and the off-diagonal term as a dotted line. As can be seen in the plot, the diagonal term is very close to 1 (0 dB) from 2 to 15 kHz and deviates only slightly in the region below 1 kHz. The off-diagonal term starts at -15 dB and remains below -30 dB from 1 to 15 kHz. For the least-squares method, we designed the filter $\boldsymbol{f}_{ii}$ using a length of 349 taps, introducing a delay of 140 samples and an SER of 44.9 dB. The filter $\boldsymbol{f}_{ci}$ was designed using a filter of 150 taps length, a delay of 70 samples and SER of 31.6 dB with frequency response H_c/H_i, and a filter of frequency response of 200 taps length, delay of 70 samples and SER of 33 dB and then convolved their time domain responses. The resulting diagonal and off-diagonal terms are plotted in Fig. 12, in which the diagonal term is plotted as a solid line and the off-diagonal term as a dotted line. As in the LMS case, the diagonal term is near 1 (0 dB) in the range of 20 Hz to 15 kHz and the off-diagonal term starts at -15 dB and remains below -30 dB up to 15 kHz.

V. Conclusions

Several theoretical and practical aspects in the implementation of immersive audio rendering were discussed in this paper. They include inversion of nonminimum phase filters and crosstalk cancellation that is an inherent problem in loudspeaker-based rendering. Two methods were examined to implement a set of filters that can be used to generate the necessary inverse filters required for rendering virtual sound sources, namely the least-squares and LMS algorithms. Our simulations have shown that both methods provide good crosstalk cancellation results using various HRTF's. Although mathematical measures such as the SER give an indication of relative performance among different methods, the final validation should be performed using the human ear. We are currently implementing these evaluations and will report on them in a future publication.

One of the main advantages of the FTF implementation of the least-squares algorithm is that it is highly suitable for real-time implementations. This is of particular importance for the case of a moving listener in which a different set of HRTF's must be implemented for every listener position.

Acknowledgment

The authors wish to thank Prof. R. Leahy, Director of the USC Signal and Image Processing Institute for his insights and helpful suggestions.

References

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization, Revised Edition*. Cambridge, MA: MIT Press, 1997.
[2] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," *Proc. IEEE*, vol. 86, pp. 941–951, 1998.
[3] H. Moller, M. F. Sorensen, and D. Hammershoi, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, pp. 300–321, 1995.
[4] A. D. Musicant and R. A. Butler, "The influence of pinnae-based spectral cues on sound localization," *J. Acoust. Soc. Amer.*, vol. 75, pp. 1195–1200, 1984.
[5] E. M. Wenzel, M. Arruda, and D. J. Kistler, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 94, pp. 111–123, 1993.
[6] R. A. Butler, "Spatial hearing: The psychophysics of human sound localization," *J. Acoust. Soc. Amer.*, vol. 77, pp. 334–335, 1985.
[7] D. H. Cooper, "Calculator program for head-related transfer functions," *J. Audio Eng. Soc.*, vol. 30, pp. 34–38, 1982.
[8] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 476–488, 1998.
[9] D. R. Begault, "Challenges to the successful implementation of 3-D sound," *J. Audio Eng. Soc.*, vol. 39, pp. 864–870, 1991.
[10] F. L. Wightman, D. J. Kistler, and M. Arruda, "Perceptual consequences of engineering compromises in synthesis of virtual auditory objects," *J. Acoust. Soc. Amer.*, vol. 101, pp. 1050–1063, 1992.

[11] P. A. Nelson, H. Hamada, and S. J. Elliott, "Adaptive inverse filters for stereophonic sound reproduction," *IEEE Trans. Signal Processing*, vol. 40, pp. 1621–1632, 1992.
[12] J. Blauert, H. Lehnert, W. Pompetzki, and N. Xiang, "Binaural room simulation," *Acustica*, vol. 72, pp. 295–296, 1990.
[13] H. W. Gierlich, "The application of binaural technology," *Appl. Acoust.*, vol. 36, pp. 219–243, 1992.
[14] H. Moller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, pp. 171–218, 1992.
[15] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, pp. 1648–1661, 1992.
[16] P. Damaske and V. Mellert, "A procedure for generating directionally accurate sound images in the upper half-space using two loudspeakers," *Acustica*, vol. 22, pp. 154–162, 1969.
[17] W. G. Gardner. Head-tracked 3-D audio using loudspeakers. presented at WASPAA'97
[18] ——, *3-D Audio Using Loudspeakers*. Norwell, MA: Kluwer, 1998.
[19] C. Kyriakakis, T. Holman, J.-S. Lim, H. Hong, and H. Neven, "Signal processing, acoustics, and psychoacoustics for high quality desktop audio," *J. Vis. Commun. Image Represent.*, vol. 9, pp. 51–61, 1997.
[20] C. Kyriakakis and T. Holman, "Video-based head tracking for improvements in multichannel loudspeaker audio," presented at the 105th Meeting of the Audio Engineering Society, San Francisco, CA, 1998, Preprint no. 4845.
[21] M. R. Schroeder and B. S. Atal, "Computer simulation of sound transmission in rooms," in *IEEE Int. Conv. Rec.*, vol. 7, 1963.
[22] P. Damaske, "Head related two channel stereophony with loudspeaker reproduction," *J. Acoust. Soc. Amer.*, vol. 50, pp. 1109–1115, 1971.
[23] J. Bauck and D. H. Cooper, "Generalized transaural stereo and applications," *J. Audio Eng. Soc.*, vol. 44, pp. 683–705, 1996.
[24] D. H. Cooper and J. L. Bauck, "Prospects for transaural recording," *J. Audio Eng. Soc.*, vol. 37, pp. 3–19, 1989.
[25] W. G. Gardner, "Transaural 3-D audio," MIT Media Laboratory, Cambridge, MA, Tech. Rep. no. 342, January/February 1995.
[26] A. V. Oppenheim and R. W. Shafer, *Discrete Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[27] S. Haykin, *Adaptive Filter Theory, 3rd ed.*. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[28] J. M. Cioffi and T. Kailath, "Fast, recursive least-squares transversal filters for adaptive filtering," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-32, pp. 304–307, 1984.
[29] F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences," *J. Audio Eng. Soc.*, vol. 34, pp. 227–235, 1986.
[30] F. E. Toole and S. E. Olive, "The modification of timbre by resonances: Perception and measurement," *J. Audio Eng. Soc.*, vol. 36, pp. 122–142, 1988.
[31] R. Walker, "Early reflections in studio control rooms: The results from the first controlled image design installations," presented at the 96th Meeting of the Audio Engineering Soc., Amsterdam, The Netherlands, 1994.

Athanasios Mouchtaris received the Diploma degree in electrical engineering from the Aristotle University, Thessaloniki, Greece, in 1997 and the M.S. degree in electrical engineering from the University of Southern California, Los Angeles, in 1999. He is currently pursuing the Ph.D. degree in the area of audio signal processing. His research interests include spatial audio rendering, audio signal processing, and multimedia applications.

Panagiotis Reveliotis received the Diploma degree in electrical engineering from the National Technical University of Athens (NTUA), Greece, in 1994, and the M.S. degree in electrical engineering from the University of Southern California, Los Angeles, in 1999.

He is now with Philips Research Laboratories, Briarcliff Manor, NY. His research interests include spatial audio rendering, statistical signal processing, and multimedia applications.

Chris Kyriakakis (M'96) received the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 1993.

He is the Head of the Immersive Audio Laboratory within the Integrated Media Systems Center at USC. His research interests include immersive audio signal processing, adaptive methods for audio, and immersive telepresence.