

IMMERSIVE AUDIO RENDERING ALGORITHMS USING THE TI C62 EVM BOARD

Alexei Ossadtchi, Athanasios Mouchtaris, and Chris Kyriakakis

Integrated Media Systems Center
University of Southern California
3740 McClintock Ave., EEB 432
Los Angeles, California 90089-2564, U.S.A.
Phone: (213) 740-8600, Fax: (213) 740-4651
email: ckyriak@imsc.usc.edu

Abstract

Immersive audio systems can be used to render virtual sound sources in three-dimensional 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 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 non-symmetric listening positions and moving listeners. Finally, we present our initial findings from the implementation of such systems on the TI C62 EVM platform and discuss computational complexity and performance tradeoffs.

1. Introduction

Accurate spatial reproduction of sound can significantly enhance the visualization of three-dimensional 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 three-dimensional 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 non-directional (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 loudspeakers are examined in this paper. Furthermore, we examine practical issues that relate to the realization of such filters on the TI C62 DSP platform.

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 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 stereophonic reproduction have been studied extensively by Nelson *et al.* [11]. The work presented in this paper is an extension into HRTF-based spatial audio rendering in which the goal is to achieve real-time filter synthesis for interactive applications.

2. 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].

In this paper we focus on loudspeaker methods for rendering binaural sound. In order 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 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 [17, 18]. A camera-based system that does not require that the user to be tethered is also under development [19, 20].

Several schemes have been proposed to address crosstalk cancellation. The first such scheme was proposed by Atal and 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 of Cooper and Bauck [23, 24] that the loudspeakers are placed symmetrically with respect to the median plane (*i.e.*, $H_{LR} = H_{RL}$ and $H_{LL} = H_{RR}$)

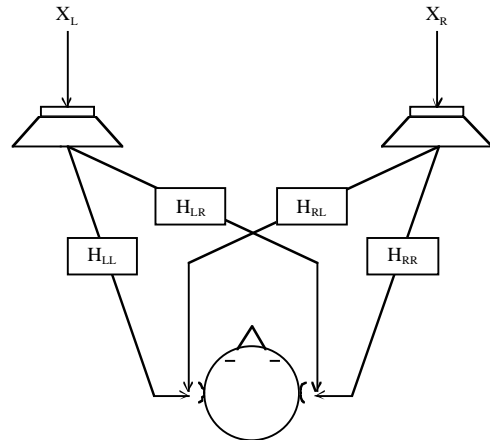


Fig. 1. Contralateral and ipsilateral terms in a loudspeaker-based rendering system.

leads to a solution that uses the diagonalized form of the matrix introduced by the physical system. 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. While in our analysis we present the symmetric case, the methods that we propose are easily extendable to the non-symmetric 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 on the TI C62 platform 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

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \quad (1)$$

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

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \quad (2)$$

In order to deliver the signals in (1), given that the physical system results in (2), pre-processing 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

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix}^{-1} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \quad (3)$$

It can be seen that equations (1) and (3) are essentially the same. Solving (3) we find

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} 1 & -\frac{H_c}{H_i} \\ -\frac{H_c}{H_i} & 1 \end{bmatrix} \begin{bmatrix} \frac{H_L}{H_i} & 0 \\ 0 & \frac{H_R}{H_i} \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \quad (4)$$

assuming that $\left(1 - \frac{H_c^2}{H_i^2}\right)^{-1} \cong 1$

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 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

$$\begin{bmatrix} X_L \\ X_R \end{bmatrix} = \begin{bmatrix} \frac{H_L}{H_i} & -\frac{H_c}{H_i} \frac{H_R}{H_i} \\ -\frac{H_c}{H_i} \frac{H_L}{H_i} & \frac{H_R}{H_i} \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \quad (5)$$

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

$$\begin{aligned} F_L &= \frac{H_L}{H_i} - \frac{H_c}{H_i} \frac{H_R}{H_i}, \\ F_R &= \frac{H_R}{H_i} - \frac{H_c}{H_i} \frac{H_L}{H_i}. \end{aligned} \quad (6)$$

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

3. Theoretical Analysis

The analysis in the previous sections has shown that crosstalk cancellation, and loudspeaker HRTF inversion, all require the implementation of preprocessing filters of the type $H_{inv} = H_x/H_i$, in which H_x is 1, H_L , H_R or H_c and H_i is the ipsilateral response. 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_i 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_i 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 non-causal but

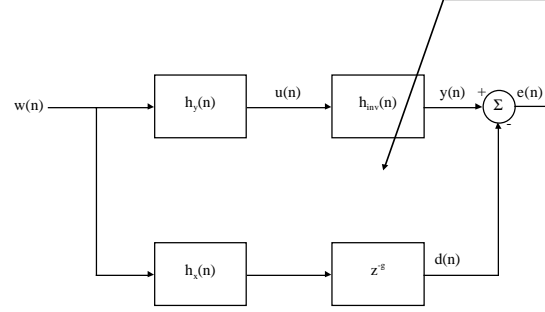


Fig. 2. Block diagram for estimation of the inverse filter. The problem of finding the filter H_{inv} such that the mean squared error between $y(n)$ and $d(n)$ is minimized, is a combination of a system identification problem (with respect to H_x) and inverse modeling problem (with respect to H_y) and its solution can be based on standard adaptive methods.

stable impulse response, which also corresponds to H_x/H_i 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 non-causal solution.

A. 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_i , H_x and the preprocessing filter H_{inv} can be depicted as in the block diagram shown in Fig. 2.

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_i) and its solution can be based on standard adaptive methods such as the LMS algorithm [27]. More specifically, the taps of the filter H_{inv} at iteration n can be computed based on the weight adaptation formula

$$\mathbf{h}_{inv}(n+1) = \mathbf{h}_{inv}(n) + \mu \mathbf{u}(n)e(n) \quad (7)$$

in which,

$$e(n) = d(n) - \mathbf{h}_{inv}^H(n)\mathbf{u}(n) \quad (8)$$

In (8) H denotes the Hermitian of the vector \mathbf{h}_{inv} . The desired response $d(n)$ can be found from Fig. 2 to be

$$d(n) = \mathbf{h}_x^H(n)\mathbf{w}(n-g) \quad (9)$$

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

$$\mathbf{u}(n) = [u(n) \quad u(n-1) \quad \dots \quad u(n-M+1)]^T \quad (10)$$

where, M is the order of the filter \mathbf{h}_{inv} . This is also true for $\mathbf{w}(n)$. The system input $\mathbf{w}(n)$ 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 fre-

quency 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

$$\mu(n) = \frac{\beta}{\alpha + \|\mathbf{u}(n)\|^2} \quad (11)$$

In (11) β is a positive constant, usually less than 2, and α is a small positive constant [27].

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

B. Least-Squares Filter Design Method

Referring again to Fig. 2, 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):

$$\min_{\mathbf{h}_{inv}(m)} \sum_{n=M}^N \left| \sum_{m=0}^M u(n-m)h_{inv}(m) - d(n) \right|^2 \quad (12)$$

The above equation can be rewritten in matrix notation as

$$\min_{\mathbf{h}_{inv}} \|\mathbf{H}\mathbf{h}_{inv} - \mathbf{h}_x\|^2 \quad (13)$$

in which \mathbf{H} is a rectangular Toeplitz matrix that can be easily derived from (12). The solution to (13) in the Least-Squares sense is

$$\mathbf{h}_{inv} = \mathbf{H}^+ \mathbf{h}_x \quad (14)$$

in which we denote the pseudoinverse of \mathbf{H} as \mathbf{H}^+ . In general, (13) describes an overdetermined system for which \mathbf{H}^+ in (14) can be written as

$$\mathbf{H}^+ = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \quad (15)$$

We denote $\mathbf{P} = \mathbf{H}^H \mathbf{H}$ 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 \mathbf{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 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 \mathbf{P} into a recursive calculation of three vectors \mathbf{A} , \mathbf{B} , and \mathbf{C} , which is a procedure that requires fewer computations, since no matrix multiplication is involved.

In section 4 we describe our findings including a discussion of the optimal values of λ and μ . We also show that there is a significant advantage of the FTF algorithm over the LMS algorithm in terms of convergence rate while incurring only a moderate increase in computational complexity.

4. Simulation Results

A. Loudspeaker Inversion

All of the filters that are of the form H_x/H_i 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 $H_{inv}H_i$ (which we will call the *cascade* filter) and H_x . In our case, the SER is defined by

$$\frac{\sum_{k=1}^N h_x^2(k)}{\sum_{k=1}^N (h_x(k) - h_{ca}(k))^2} \quad (16)$$

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's 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 seconds from 125 Hz to 4 kHz.

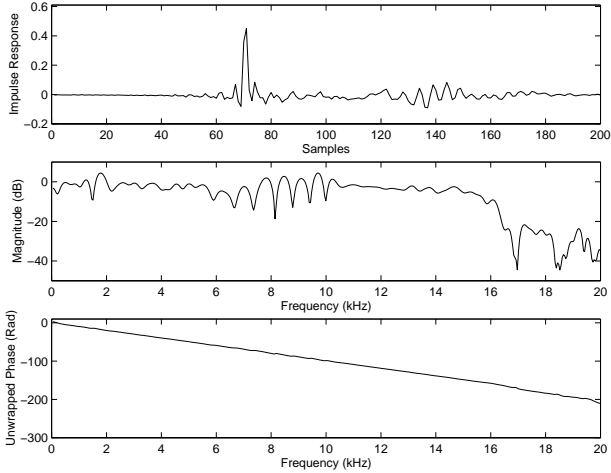


Fig. 3. Impulse response (top), magnitude response (middle) and phase response (bottom) of the designed filter H_{inv} using the LMS method.

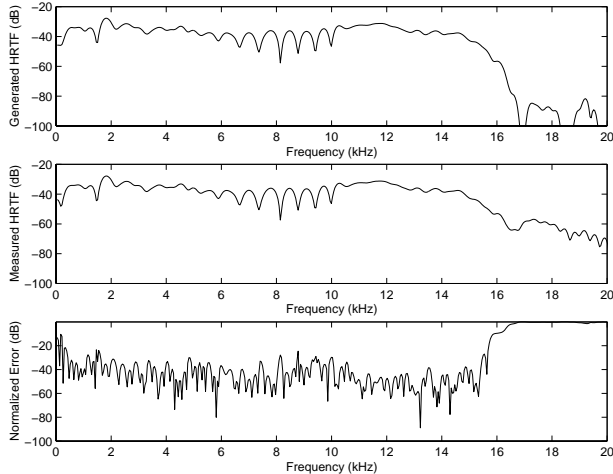


Fig. 4. The HRTF generated from the inverse filter using the LMS method is shown in the upper plot. The measured HRTF (0° azimuth and 0° elevation) is shown in the middle and the relative error in the bottom plot.

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 results for the LMS method can be seen in Figs. 3 and 4. Similar results were obtained for the Least-Squares case. 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.

It should be noted that for frequencies above 15 kHz, the associated wavelengths are less than 20 mm. In this range it is practi-

cally 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 Fig. 4) is acceptable since listener position errors will dominate.

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 of (6) 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

$$F_L = \frac{H_L}{H_i} z^{-(d_1+d_2)} - \frac{H_c}{H_i} z^{-d_1} \frac{H_R}{H_i} z^{-d_2} \quad (17)$$

Note that the delay for H_L/H_i in (17) 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 (17) becomes

$$\begin{aligned} \mathbf{f}_l &= \mathbf{h}_{li} - \mathbf{h}_{ci} * \mathbf{h}_{ri} \\ \mathbf{f}_r &= \mathbf{h}_{ri} - \mathbf{h}_{ci} * \mathbf{h}_{li} \end{aligned} \quad (18)$$

in which $*$ denotes convolution.

In order to design the filter for each channel, each of the three filters \mathbf{h}_{li} , \mathbf{h}_{ci} and \mathbf{h}_{ri} can be designed separately, and then be combined using (18) 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 \mathbf{h}_{li} in \mathbf{f}_l is $d_1 + d_2$ while in \mathbf{f}_r it is d_2 . A similar argument holds for \mathbf{h}_{ri} . This means that the filters \mathbf{h}_{li} and \mathbf{h}_{ri} required for \mathbf{f}_l will be different from the filters \mathbf{h}_{li} and \mathbf{h}_{ri} required for \mathbf{f}_r . Only the filter \mathbf{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. 5). Toole [29, 30] and Walker [31] studied the psychoacoustic effects of early reflections and 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.

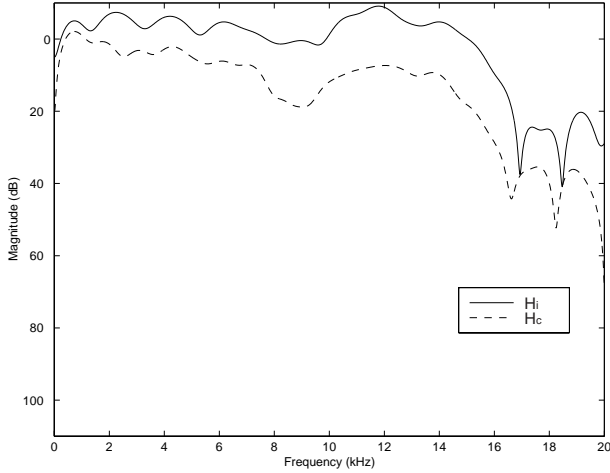


Fig. 5. The difference in dB between the ipsilateral (H_i) and the contralateral (H_c) terms shows the effect of head shadowing with no crosstalk cancellation. In this set-up the loudspeakers were 50 cm apart and the head was located in the symmetric (center) position at a distance of 50 cm from the loudspeaker baffle plane.

For the symmetric positioning of the listener that we have examined, we saw that for the binaural input case we can set $H_L = H_R = 1$ in (4) since the virtual source HRTF's are already contained in the binaural signal. Then, (8) becomes

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} 1 & -\frac{H_c}{H_i} \\ -\frac{H_c}{H_i} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{H_i} & 0 \\ 0 & \frac{1}{H_i} \end{bmatrix} \begin{bmatrix} S_{b1} \\ S_{b2} \end{bmatrix} \quad (19)$$

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 (19) can be written as

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} F_{ii} & F_{ci} \\ F_{ci} & F_{ii} \end{bmatrix} \begin{bmatrix} S_{b1} \\ S_{b2} \end{bmatrix} \quad (20)$$

which finally becomes

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i F_{ii} + H_c F_{ci} & H_i F_{ci} + H_c F_{ii} \\ H_i F_{ci} + H_c F_{ii} & H_i F_{ii} + H_c F_{ci} \end{bmatrix} \begin{bmatrix} S_{b1} \\ S_{b2} \end{bmatrix} \quad (21)$$

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 $H_i F_{ii} + H_c F_{ci}$ must be 1 (this would mean that the loudspeaker frequency response inversion has succeeded) and the off-diagonal term $H_i F_{ci} + H_c F_{ii}$ 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 f_{ii} using a length of 349 taps, introducing a delay of 140 samples and an SER of 44.1 dB. For the filter 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 men-

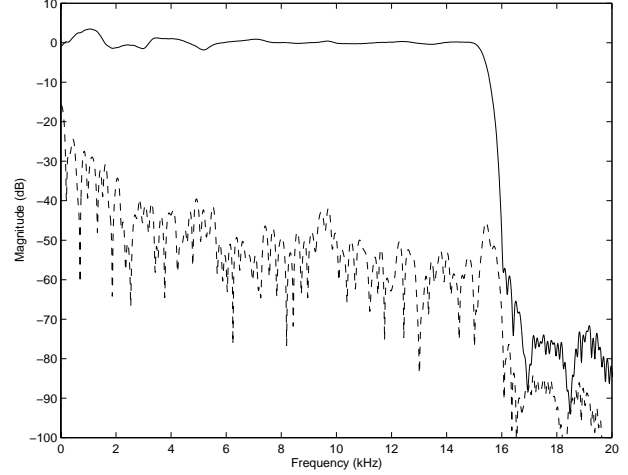


Fig. 6. Measured HRTF data from the loudspeakers (H_i and H_c) were used to simulate the physical system and design a set of filters to eliminate the crosstalk. The resulting diagonal (solid line) and off-diagonal (dotted line) terms produced by our simulation using the LMS method are plotted above. The diagonal term is very close to 1 (0 dB) from 2 kHz 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 kHz to 15 kHz.

tioned 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. 6, 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 kHz 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 kHz to 15 kHz. For the Least-Squares method, we designed the filter f_{ii} using a length of 349 taps, introducing a delay of 140 samples and an SER of 44.9 dB. The filter 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. 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.

5. Implementation on the TI C62 EVM Board

An experimental audio rendering system was implemented on the C62 EVM to evaluate the performance of the algorithms described above. The C6x EVM includes a CD quality, 16-bit audio interface with stereo microphone and line-level inputs and a stereo line level output. In order to accommodate the filter lengths required for effective rendering, a convolution scheme based on the time invariance property of convolution was implemented as described by

$$h_k[n] = h[n - kN]$$

$$y_k[n] = x[n] * h_k[n], k = 0..P-1, P = \frac{L}{N} \quad (22)$$

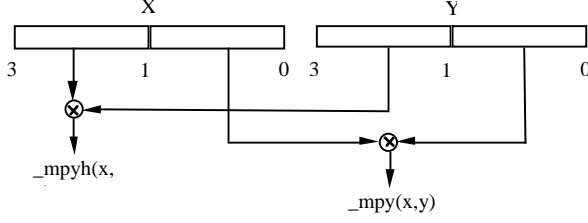


Fig. 7 TMS intrinsics `int_mpy(int x,int y)` and `int_mpyh(int x, int y)`.

in which L is the convolution filter length and N is unrolling constant (we typically use $N=32$ and $L = 2048$). Finally, using the time invariance property we can write the convolution as

$$y[n] = \sum_{k=0}^{P-1} y_k[n - k] \quad (23)$$

Since all computations are performed using fixed point arithmetic, which is the most appropriate for the TI chip currently used, care must be taken with the normalizing constant and binary shift operations in order to avoid 32-bit number overflow.

In our implementation we used the TMS DSP intrinsics `_mpy(int x, int y)`, `_mpyh(int x, int y)` that perform multiplication of the least significant and most significant 16-bit words of the 32-bit operands. These intrinsics are outlined in Fig. 7. After performing the multiplication using the intrinsics we perform an addition and shift one bit to the right in case the result does not fit in 32 bits. This is performed by the intrinsic supported by TMS DSP `int_sadd(int a,int b)`.

Computational complexity

For an HRTF impulse response length $N = 512$ we need to perform N multiplications and $N-1$ additions. In other words for two channel loudspeaker rendering (sampled at 44.1 kHz) we need approximately $2*44100*512 = 45$ Mips (combined multiplications & addition operations). In such DSP convolution algorithm implementations the most time-consuming operation is memory access. Thus, to achieve the required performance it is necessary to minimize memory requests. This is accomplished by using 32-bit word access to retrieve two operands simultaneously, through loop unrolling to avoid memory hits, and by using intrinsics to call hardwired operations specific to the problem of elementary FIR filtering.

Taking these requirements into account we reconsidered the convolution algorithm. Using Matlab like pseudo-code we can outline the method as shown in Fig. 8. `MSB()` and `LSB()` denote the most significant 16-bit word and least significant 16-bit word of the 32-bit integer respectively. The operation `|` denotes packing of two 16-bit values into one 32-bit word. The left most operand goes to the MSB and the right most one goes to the LSB. We also use the state variables x_j, y_j in order to minimize memory

```

P = L/N
for k = 0:P-1
    c_j = h_{j+k\frac{N}{2}} | h_{j+1+k\frac{N}{2}}, j = 0..\frac{N}{2}-1
    x_0 = s_0 | 0, x_j = 0, j = 1..\frac{N}{2}
    y_0 = s_0 | s_1, y_j = 0, j = 1..\frac{N}{2}

    for i = 1:2:L_s-1
        a = s_i;
        z_{i+kN} = z_{i+kN} + _sadd(_mpy(x_0,c_0),_mpyh(x_0,c_0)) + ... _sadd(_mpy(x_N,c_N),_mpyh(x_N,c_N));
        z_{i+1+kN} = z_{i+1+kN} + _sadd(_mpy(x_0,c_0),_mpyh(x_0,c_0)) + ... _sadd(_mpy(x_N,c_N),_mpyh(x_N,c_N));

        y_N = y_N - y_1, ... y_2 = y_1, y_0 = a;
        x_N = x_N - x_1, ... x_2 = x_1, x_0 = LSB(y_1) | MSB(y_0);
    end
end

```

Fig. 8 Matlab pseudo-code for convolution

requests and let the optimizer take advantage of the processor's registers.

6. Conclusions

Several aspects in the implementation of immersive audio rendering were discussed in this paper. They include inversion of non-minimum 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 and measurements have shown that both methods provide good crosstalk cancellation results using various HRTF's.

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.

Finally, an experimental rendering system was implemented on the TI C62 EVM. This system is currently under evaluation to identify the computational bottlenecks and performance trade-offs.

7. Acknowledgments

This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center with additional support from the Annenberg Center for Communication at USC, the California Trade and Commerce Agency, and the Texas Instruments DSP Research Fund.

8. References

- [1] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, Revised Edition. Cambridge, Massachusetts: MIT Press, 1997.
- [2] C. Kyriakakis, "Fundamental and Technological Limitations of Immersive Audio Systems," IEEE Proceedings, vol. 86, pp. 941-951, 1998.
- [3] H. Moller, M. F. Sorensen, and D. Hammershoi, "Head-related transfer functions of human subjects," Journal of the Audio Engineering Society, vol. 43, pp. 300-21, 1995.

- [4] A. D. Musicant and R. A. Butler, "The influence of pinnae-based spectral cues on sound localization," *Journal of the Acoustical Society of America*, vol. 75, pp. 1195-200, 1984.
- [5] E. M. Wenzel, M. Arruda, and D. J. Kistler, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, pp. 111-23, 1993.
- [6] R. A. Butler, "Spatial hearing: the psychophysics of human sound localization," *Journal of the Acoustical Society of America*, vol. 77, pp. 334-5, 1985.
- [7] D. H. Cooper, "Calculator Program for Head-Related Transfer Functions," *Journal of the Audio Engineering Society*, vol. 30, pp. 34-38, 1982.
- [8] C. P. Brown and R. O. Duda, "A Structural Model for Binaural Sound Synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 476-488, 1998.
- [9] D. R. Begault, "Challenges to the successful implementation of 3-D sound," *Journal of the Audio Engineering Society*, vol. 39, pp. 864-70, 1991.
- [10] F. L. Wightman, D. J. Kistler, and M. Arruda, "Perceptual Consequences of Engineering Compromises in Synthesis of Virtual Auditory Objects," *Journal of the Acoustical Society of America*, vol. 101, pp. 1050-1063, 1992.
- [11] P. A. Nelson, H. Hamada, and S. J. Elliott, "Adaptive Inverse Filters for Stereophonic Sound Reproduction," *IEEE Transactions on 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," *Applied Acoustics*, vol. 36, pp. 219-243, 1992.
- [14] H. Moller, "Fundamentals of Binaural Technology," *Applied Acoustics*, 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," *Journal of the Acoustical Society of America*, vol. 91, pp. 1648-61, 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-Trackable 3-D Audio Using Loudspeakers," presented at WASPAA '97, New Palz, New York, 1997.
- [18] W. G. Gardner, *3-D Audio Using Loudspeakers*. Norwell, Massachusetts: Kluwer Academic Publishers, 1998.
- [19] C. Kyriakakis, T. Holman, J.-S. Lim, H. Hong, and H. Neven, "Signal Processing, Acoustics, and Psychoacoustics for High Quality Desktop Audio," *Journal of Visual Communication and Image Representation*, vol. 9, pp. 51-61, 1997.
- [20] C. Kyriakakis and T. Holman, "Video-Based Head Tracking for Improvements in Multichannel Loudspeaker Audio," presented at 105th Meeting of the Audio Engineering Society, Preprint No. 4845, San Francisco, California, 1998.
- [21] M. R. Schroeder and B. S. Atal, "Computer Simulation of Sound Transmission in Rooms," *IEEE International Convention Record*, vol. 7, 1963.
- [22] P. Damaske, "Head Related Two Channel Stereophony with Loudspeaker Reproduction," *Journal of the Acoustical Society of America*, vol. 50, pp. 1109-1115, 1971.
- [23] J. Bauck and D. H. Cooper, "Generalized Transaural Stereo and Applications," *Journal of the Audio Engineering Society*, vol. 44, pp. 683-705, 1996.
- [24] D. H. Cooper and J. L. Bauck, "Prospects for Transaural Recording," *Journal of the Audio Engineering Society*, vol. 37, pp. 3-19, 1989.
- [25] W. G. Gardner, "Transaural 3-D Audio," MIT Media Laboratory, Technical Report No. 342, January/February, 1995.
- [26] A. V. Oppenheim and R. W. Shafer, *Discrete Time Signal Processing*: Prentice Hall, 1989.
- [27] S. Haykin, *Adaptive Filter Theory*, 3rd Edition: Prentice Hall, 1996.
- [28] J. M. Cioffi and T. Kailath, "Fast, recursive least-squares transversal filters for adaptive filtering," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-32, pp. 304-307, 1984.
- [29] F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences," *Journal of the Audio Engineering Society*, vol. 34, pp. 227-235, 1986.
- [30] F. E. Toole and S. E. Olive, "The modification of timbre by resonances: perception and measurement," *Journal of the Audio Engineering Society*, vol. 36, pp. 122-42, 1988.
- [31] R. Walker, "Early Reflections in Studio Control Rooms: The Results from the First Controlled Image Design Installations," presented at 96th Meeting of the Audio Engineering Society, Amsterdam, 1994.