NON-MINIMUM PHASE INVERSE FILTER METHODS FOR IMMERSIVE AUDIO RENDERING

A. Mouchtaris, P. Reveliotis, and C. Kyriakakis

Integrated Media Systems Center
University of Southern California
3740 McClintock Ave., EEB 432
Los Angeles, California 90089-2564, USA

ABSTRACT

Immersive audio systems are being envisioned for applications that include teleconferencing and telepresence; augmented and virtual reality for manufacturing and entertainment; 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. The principal function of such systems is to synthesize, manipulate and render sound fields in real time. In this paper we examine several signal processing considerations in spatial sound rendering over loudspeakers. We propose two methods that can be used to implement the necessary filters for generating virtual sound sources based on synthetic head-related transfer functions with the same spectral characteristics as those of the real source.

1. INTRODUCTION

Immersive audio seeks to render virtual sound sources from a particular direction using a set of loudspeakers. The basic concept is to reproduce the sound pressure level that would reach the eardrum if the sound were actually coming from the direction 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 pinnae must be considered [1, 2].

It is necessary, therefore, to implement a filter that alters non-directional (monaural) sound in the same way as the pinnae. 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 [3, 4]. More recent methods are based on the measurement (or simulation) of the impulse response of the pinnae for each desired direction. In this paper we propose a method that is based on measurements of the impulse response for each ear using a microphone placed inside the ear of a human subject or a mannequin. The impulse response is usually mentioned as the Head Related Transfer Function (HRTF) both in the time and frequency domains and it contains all the important spectral information for localization. The main advantage of this method compared to analytical models is that it accounts for the pinnae, the imperfections of the shape of human head, and the effects of the upper body. Several practical problems that arise when using the HRTF approach for immersive audio rendering are examined in this paper.

In the case of rendering immersive audio using two 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.

Another issue for loudspeaker-based immersive audio arises from the fact that each ear receives sound from both loudspeakers resulting in acoustic crosstalk. In order to deliver the desired signal to each ear it is necessary to implement a crosstalk cancellation method.

In this paper we refer to monaural sound as non-directional 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. When binaural sound is reproduced through headphones, there is no crosstalk and the desired directionality can be achieved by inverting the transfer function of the headphones. In this paper we will focus on reproduction of immersive audio through two loudspeakers.

2. CROSSTALK CANCELLATION

Crosstalk cancellation can be achieved by eliminating the contralateral terms H_{RL} and H_{LR} (Fig. 1), so that each loudspeaker is perceived to reproduce sound only for the corresponding 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 for the listener's ears. This implies that, in the analysis that follows, if the position of the listener changes then these terms must also change so as to correspond to the new listener position. This can be achieved through tracking of the listener's head position in three-dimensional space [5, 6]. There are a number of different methods for the solution of this problem, usually attempting to approximate H_{LL}, $H_{RR,}$ H_{LR} and H_{RL} using a model. One simple method is to

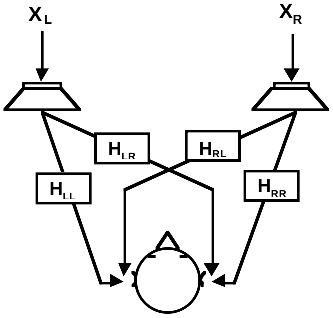

Figure 1. Contralateral and ipsilateral terms in a loudspeaker-based rendering system.

model the head as a sphere and then calculate the ipsilateral and contralateral terms [7, 8]. Another method approximates the effect of the head with a low - pass filter, a delay and a gain (less than 1) [9].

While both of the methods above have the advantage of low computational cost, the approximations involved introduce distortions particularly in the perceived timbre of virtual sound sources behind the listener. In this paper we use a different approach. In order to describe the procedure mathematically we use matrix notation and represent the loudspeaker-ear system as a two input - two output system in which we need to process the two channels simultaneously. The analysis below assumes that the loudspeakers are placed symmetrically with respect to the median plane. We are also currently working on an implementation in which the listener is tracked and the filters are computed in real time in response to changes in the listener's position.

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, in matrix notation, the signals E_L and E_R at the left and right eardrums respectively are

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

The introduction of the contralateral and ipsilateral terms from the physical system (taking into account the assumptions made before) will introduce an additional transfer matrix

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

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

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

We see that (1) and (3) are essentially the same. Solving (3) we find

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

assuming that $1/(1-H_c^2/H_i^2) \approx 1$. 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 terms H_L/H_i and H_R/H_i in (4) correspond to the speaker position inversion. That is, the actual position of the speakers is inverted because it adds undesired spectral information to the binaural signal. The matrix

$$\begin{bmatrix} 1 & -\frac{H_c}{H_i} \\ -\frac{H_c}{H_i} & 1 \end{bmatrix} \tag{5}$$

corresponds to the actual crosstalk cancellation. In the approach described here, the crosstalk cancellation and the inversion of the loudspeakers' position are closely connected, but it is important to state the difference between these two terms. Finally, the required filters F_L and F_R for the left and right channel are

$$F_L = \frac{H_L}{H_i} - \frac{H_c}{H_i}\frac{H_R}{H_i}, \qquad F_R = \frac{H_R}{H_i} - \frac{H_c}{H_i}\frac{H_L}{H_i}. \tag{6}$$

Alternatively, a filter can be designed for the case that the input is the binaural signal S_b instead of the monaural signal S. In this case, convolution with the pair of the HRTF's H_L and H_R is not needed since the binaural signal already contains the directional HRTF information. In this case the terms H_R and H_L are simply unity. Obviously, the filters that must be designed in this case for the left and right channel are identical (because of the symmetry assumption) and equal to

$$F_L = F_R = \frac{1}{H_i} - \frac{H_c}{H_i}\frac{1}{H_i} \tag{7}$$

The above analysis has shown that crosstalk cancellation and loudspeaker position inversion require the implementation of preprocessing filters of the type H_x/H_i (in which H_x is 1, H_L, H_R or H_c), which we will denote as H_{inv}. 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 [10].

Here, we propose a different procedure that maintains the HRTF phase information. We first find the non-causal but stable impulse response, which also corresponds to H_x/H_i assuming a different Region of Convergence for the transfer function. Then, a delay is introduced in order 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. We describe below our methods for finding this non-causal solution.

3. THEORETICAL ANALYSIS

3.1 Adaptive Algorithm

Based on the discussion in the previous paragraph and taking into consideration the need for adding a delay in order for the preprocessing filter to be feasible (*i.e.* causal), as explained in 1.1, we conclude that the relationship between the filters H_i, H_x and the preprocessing filter H_{inv} can be represented as in the block diagram in Fig. 2.

The problem of defining the filter H_{inv} so that the mean squared error $E[e(n)]$ is minimized, can be classified as a combination of system identification (with respect to H_x) and inverse modeling (with respect to H_i) problem and its solution can be found using adaptive methods such as the LMS algorithm [11].

More specifically the taps of the filter H_{inv} can be defined based on the weight adaptation formula

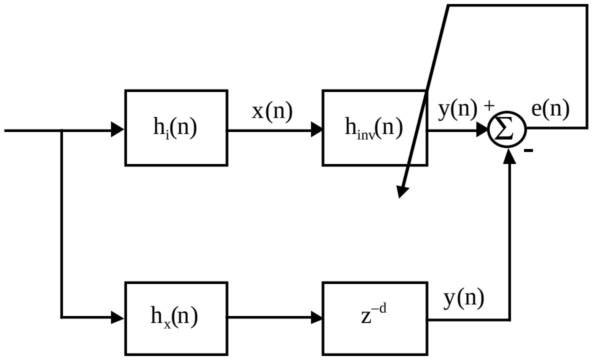

Figure 2. LMS block diagram for estimation of the inverse filter.

$$h_{inv}(n+1) = h_{inv}(n) + \mu h_i(n) e(n) \tag{8}$$

in which,

$$e(n) = h_x(n-d) - h_{inv}(n) * h_i(n) \tag{9}$$

The filter length, as well as the delay d, can be selected based on the minimization of the mean squared error. Moreover, progressive adaptation (decrement) of the step size μ could lead to faster convergence as well as less misadjustment. This method can be used either off-line or in real time to account for movement of the virtual sound source position and movement of the listener's head.

The output filter of this method is $h_{inv}(n)$, which in the frequency domain is equal to H_x/H_i. If the desired output is $1/H_i$, $h_x(n)$ can be chosen to be the impulse sequence. The result is an FIR filter.

3.2 Least-Squares Method

Another way of approaching the problem is to notice that in the frequency domain ideally we must design a filter H_{inv} that satisfies

$$H_i H_{inv} = H_x \tag{10}$$

in which as discussed above H_x is 1, H_L, H_R or H_c. We will refer to the filter $H_{inv}H_i$ as the *cascade* filter. In the time domain (10) becomes a convolution relation

$$\sum_{m=0}^{M} h_i(n-m) h_{inv}(m) = h_x(n-d) \tag{11}$$

in which d is the delay introduced to satisfy the causality requirement. This equation is solved in the Least-Squares sense, that is we calculate h_{inv} such that

$$\min_{h_{inv}(m)} \sum_{n=0}^{N} \left| \sum_{m=0}^{M} h_i(n-m) h_{inv}(m) - h_x(n-d) \right|^2 \tag{12}$$

The above equation can be rewritten in matrix notation as

$$\min_{\mathbf{h}_{inv}} \left\| \mathbf{H}\mathbf{h}_{inv} - \mathbf{h}_x \right\|^2 \tag{13}$$

in which $\mathbf{H}$ is a Toeplitz matrix that can be easily derived from (11). The solution to (13) in the Least-Squares sense is

$$\mathbf{h}_{inv} = \mathbf{H}^{+}\mathbf{h}_x \tag{14}$$

in which we denote the pseudoinverse of $\boldsymbol{H}$ as $\boldsymbol{H}^{+}$. The pseudoinverse of $\boldsymbol{H}$ can be found using Singular Value Decomposition (SVD) [11]. In order to avoid very small singular values, a tolerance value is introduced so that all singular values less than the tolerance to be considered equal to zero. The result is again an FIR filter.

4. SIMULATION RESULTS

All of the filters that are encountered in (6) and (7) were designed using the Least-Squares and LMS methods. As discussed above a delay is introduced to the system. If we denote as d_1 the delay introduced by H_c/H_I in the upper part of (6) and as d_2 the delay introduced by H_R/H_i then in the z-domain we get

$$F_L = \frac{H_L}{H_i} z^{-(d_1+d_2)} - \frac{H_c}{H_i} z^{-d_1} \frac{H_R}{H_i} z^{-d_2} \tag{15}$$

Note that the delay for H_L/H_i in (15) 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 $d_1 + d_2$. A similar set of delays is used for the binaural case described in (7). 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 cascade filter $H_{inv}H_i$ and H_x (or unity for the binaural input case). It is important to evaluate the error in the time-domain because a good approximation is required both in magnitude and phase response. Both methods worked successfully with a number of different measured HRTF's.

For the monaural input case, an inverse filter of 200 taps was designed, that introduced a delay of 70 samples. The SER for this case was 41.5 dB in the time domain for the Least-Squares method and 33 dB for the LMS method. In Fig. 3 a comparison is made between the magnitude of a particular measured HRTF (0° azimuth and 0° elevation) and the HRTF generated using our inverse filter. Of course, since the approximation of the two filters is made in time domain, their phase responses are

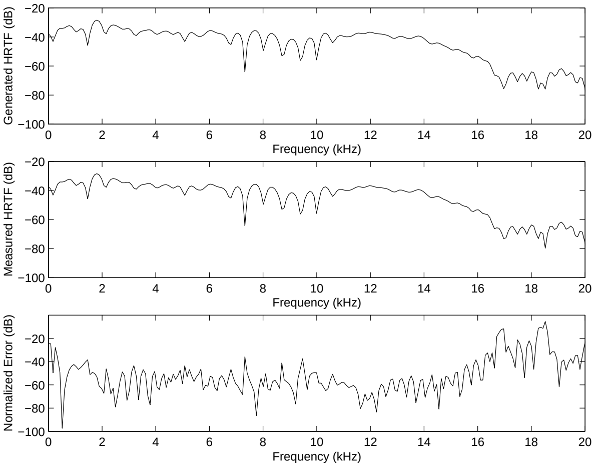

Figure 3. The HRTF generated from the inverse filter using the Least-Squares method is shown in the upper plot. The measured HRTF (0° azimuth and 0° elevation) is shown in the middle. The normalized error between the two is approximately –50 dB between 20Hz and 15 kHz.

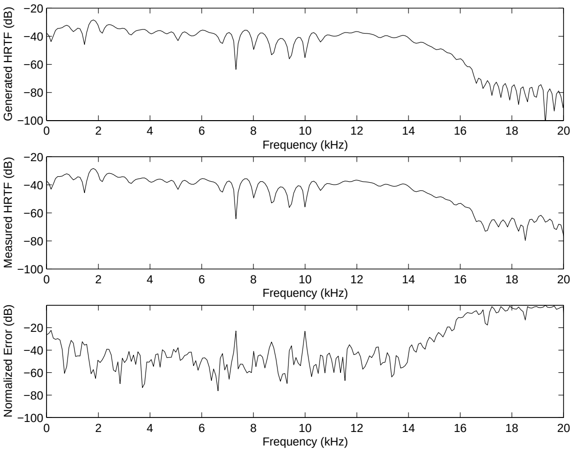

Figure 4. The HRTF generated from the inverse filter using the LMS method is shown in the upper plot. The measured HRTF (0° azimuth and 0° elevation) is shown in the middle. The normalized error between the two is approximately –40 dB between 20Hz and 15 kHz.

also almost identical. The same plot is drawn in Fig. 4 for the LMS case.

It should be noted that for frequencies above 15 kHz, the associated wavelengths are of the order of 15 mm. In this range it is practically impossible to 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 Figures 3 and 4) is acceptable.

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. The SER in the time domain is now 28 dB, but for a higher filter length of 400 taps and 160 sample delay. Alternatively, we can use a cascade filter of the form H_a/H_i where H_a has an all-pass response up to 15 kHz. Using this approximation, the resulting filters gave significantly better performance. For the Least-Squares method the SER in time domain increased to 78 dB for the same filter length and delay (400 taps, 160 sample delay). For the LMS case, the SER increased to 52 dB. In listening tests there was no perceptible difference using this method.

5. CONCLUSIONS

Important practical aspects in 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 proposed to implement a set of filters that can be used to render virtual sound sources, namely Least-Squares and LMS algorithms. Our simulations have shown that both methods give satisfactory results using various HRTF's.

Real-time performance of the methods presented here is an important issue that is currently under investigation. One of the main advantages of the LMS algorithm is that it is highly suitable for real-time implementations. We are currently examining its performance for the case of a moving listener in which a different set of HRTF's must be implemented for every position. Our experiments have shown that HRTF's with a large number of taps improve localization. We are now concentrating on the improvement of FIR to IIR conversion or on other methods that can reduce the order of the HRTF's without losing important spectral information.

6. ACKNOWLEDGMENTS

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

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 and the California Trade and Commerce Agency.

7. REFERENCES

[1] S. Mehrgard and V. Mellert, "Transformation Characteristics of the External Human Ear," *Journal of the Acoustical Society of America*, Vol. 51, pp. 1567-1576, 1977.

[2] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization, Revised Edition*, MIT Press, 1997.

[3] D. H. Cooper, "Calculator Program for Head - Related Transfer Function", *J. Audio Eng. Soc.*, Vol. 30, No. 1/2, pp. 34-381982.

[4] C. P. Brown and R. O. Duda, "A Structural Model for Binaural Sound Synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 5, pp. 476-488, 1998.

[5] 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.

[6] C. Kyriakakis, "Fundamental and Technological Limitations of Immersive Audio Systems," *IEEE Proceedings*, vol. 86, pp. 941-951, 1998.

[7] J. Bauck, D. H. Cooper, "Generalized Transaural Stereo and Applications", *J. Audio Eng. Soc.,* Vol. 44, No.9, 1996 September.

[8] D. H. Cooper, J. L. Bauck, "Prospects for Transaural Recording", *J. Audio Eng. Soc.,* Vol. 37, No. 1/2, pp. 3-19, 1989 January/February.

[9] W. G. Gardner, "Transaural 3-D audio", M.I.T. Media Laboratory, Perceptual Computing Section, Technical Report No. 342, 1995.

[10] Oppenheim A. V., Schafer R. W., *Discrete-time Signal Processing*, Prentice Hall, 1989.

[11] S. Haykin, *Adaptive Filter Theory*, 3rd Edition, Prentice Hall, New Jersey, 1996.