Approaches to Binaural Synthesis 4861 (K-3)

Jean-Marc Jot and Scott Wardle
Joint E-mu/Creative Technology Center
Scotts Valley, CA 95067, USA

Véronique Larcher
IRCAM
Paris F-75004, France

Presented at the 105th Convention 1998 September 26–29 San Francisco, California

AES

This preprint has been reproduced from the author's advance manuscript, without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents.

Additional preprints may be obtained by sending request and remittance to the Audio Engineering Society, 60 East 42nd St., Suite 2520, New York, New York 10165-2520, USA.

AN AUDIO ENGINEERING SOCIETY PREPRINT

Approaches to binaural synthesis

Jean-Marc Jot, Scott Wardle
Joint E-mu/Creative Technology Center. 1600 Green Hills Rd. Scotts Valley, CA 95067, USA.
jmj@emu.com, scottw@emu.com, http://www.emu.com

Veronique Larcher
IRCAM. 1 Place Igor-Stravinsky. Paris, F 75004, France.
larcher@ircam.fr, http://www.ircam.fr

Abstract

This paper addresses binaural synthesis and mixing algorithms for 3-D audio rendering of multiple sound sources over headphones or loudspeakers. Efficient strategies are discussed, whereby directional panning and mixing are accomplished in an intermediate multichannel format, while the binaural synthesis filters are direction-independent and can be applied after mixing ("post-filtering") or off-line on the source signals ("pre-filtering"). A "binaural B format" is proposed and its application to mixing, recording and reproduction over headphones or two or more loudspeakers are reviewed.

0 Introduction

Figure 1 describes the implementation of a binaural encoding module simulating a single sound source in free field. The design procedure is based on factoring each HRTF (head-related transfer function) into its minimum-phase and all-pass components, so that interaural time differences (ITD) and spectral transformations are modelled and rendered independently [1, 2]. The interaural excess phase difference derived from the left and right all-pass components is well approximated by a pure delay [2, 3]. According to this factorization, the left and right HRTF filters are decomposed as follows:

$$L(\theta, \varphi, f) = \tau_L(\theta, \varphi)\, \underline{L}(\theta, \varphi, f) \qquad R(\theta, \varphi, f) = \tau_R(\theta, \varphi)\, \underline{R}(\theta, \varphi, f) \tag{1}$$

where f denotes frequency, θ and φ being respectively the azimuth and elevation of the sound source expressed in radians. $\tau(\theta, \varphi)$ represents a frequency-independent delay. $\underline{L}$ represents the minimum-phase filter having the same magnitude frequency response as L.

In order to produce dynamic variation of the apparent direction of the sound without audible artifacts, the delay lines and the filters in Figure 1 must be made variable. The computational cost of a variable fractional delay line can be evaluated to 13 operations (multiply-adds) per input sample, assuming 4-point Lagrange interpolation implemented with the Farrow structure [4]. A finite impulse response (FIR) implementation of the minimum-phase filters can be made time-variant by continuously ramping the impulse response coefficients, or, equivalently, by a cross-fading technique using paired filters. The cross-fade technique can also be used with an infinite impulse response (IIR) implementation of the filters, provided that care be taken to conceal the transient response of the filters when coefficients are updated, by proper scheduling of the cross-fading coefficients. In both cases, the computational cost of a filter is aproximately doubled when it is made time-variant. If 12th-order IIR filters are used [1, 5], the total cost of this binaural encoding module is about 140 operations per input sample with a fixed-point programmable processor. With an optimized implementation using only one variable delay line and three filters instead of four, as illustrated in Figure 2, this cost can be reduced to about 100 operations per sample, i. e. about 4.8 MIPS (million instructions per second) for a sample rate of 48 kHz.

When multiple sounds must be encoded to different directions and mixed to render a complex sound scene, this cost is multiplied by the number of elementary source signals to be processed. This motivates a search for alternative designs in which the incremental expense per additional sound source is substantially less than 100 operations per sample. The following sections of this paper discuss several approaches for this purpose, making use of an intermediate directional encoding and mixing format where directional panning can be implemented more efficiently. Some previously published approaches are reviewed, as well as a "binaural B format" which provides improved performance by decoupling ITD and spectral cues while offering some of the advantages of the B format for recording, mixing and reproduction of three-dimensional sound scenes over headphones or loudspeakers.

1 Decomposition of HRTFs using a set of spatial functions

The general approach considered here for optimizing binaural synthesis with multiple sources makes use of a linear decomposition of the left and right HRTF filters into N components:

$$L(\theta, \varphi, f) = \textstyle\sum_i g_i(\theta, \varphi)\, L_i(f) \qquad \text{for } i = 1..\text{N} \tag{2}$$

$$R(\theta, \varphi, f) = L(2\pi - \theta, \varphi, f) \tag{3}$$

As expressed by Equation (3), we will assume in the remainder of this paper that the left and right ear directivity characteristics are symmetric with respect to the median plane (θ = 0). The decomposition (2) involves a set of binaural synthesis filters $L_i(f)$ which apply to any sound source, irrespective of its position, while the terms $g_i(\theta, \varphi)$ are frequency-independent spatial functions. This can be exploited in two alternative implementation approaches:

a) "post-filtering" approach (Figure 3): the source signals are directionally panned and mixed in the intermediate N-channel format using the directional panning functions $g_i(\theta, \varphi)$, and the resulting signal is then transcoded into a binaural signal (for headphones or loudspeakers) by a common module involving time-invariant implementations of the binaural synthesis filters L_i.

b) "pre-filtering" approach (Figure 4): the source signals are processed off-line through the binaural synthesis filters L_i, so that the real-time process reduces to directional panning and mixing involving frequency-independent coefficients $g_i(\theta, \varphi)$.

An identical structure can be used to synthesize the right ear signal, using the panning functions $g_i(2\pi-\theta, \varphi)$. If each of the panning functions $g_i(\theta, \varphi)$ is either symmetric or antisymmetric with respect to the median plane, i. e. $g_i(2\pi-\theta, \varphi) = \pm g_i(\theta, \varphi)$, the left and right ear signals can both be recovered from the same set of N signals.

Several methods can be used for achieving the decomposition (2):

(I) Binaural downmix method

This method consists of exploiting a known multichannel panning technique for reproduction over loudspeakers, and using binaural synthesis filters to produce a binaural downmix for reproduction over headphones. In this approach, the filters L_i involve the free-field transfer functions corresponding to the directions of the loudspeakers. Kendall & al. used pairwise amplitude panning over 12 virtual loudspeakers surrounding the listener in the horizontal plane [6]. More recently, the Ambisonic panning technique has been proposed for the same purpose. The Ambisonic approach uses the B format as the intermediate multichannel encoding format. The B format offers the possibility of mixing with an ambience recorded with a Sounfield microphone, and compensating for the rotations of the listener's head after mixing by applying a rotation matrix before transcoding to binaural format [7, 8, 9].

These two methods can be evaluated in terms of their ability to approximate the free-filed HRTF for any direction. The use of a pairwise amplitude panning technique is equivalent to a "local interpolation" method for reconstructing HRTFs for any direction based on a set of particular directions. A relatively large number of channels (at least 6) is necessary in order to produce convincing images over a 360° sound stage [10], and avoid noticeable comb-filtering effects in the interpolated HRTF spectra [1]. Additionnal channels are necessary for panning over a full sphere.

The Ambisonic approach, on the other hand, can be compared to a "global interpolation" method based on spherical harmonics decomposition if the HRTF data, as will be seen below. In this approach, the output decoding stage involves an Ambisonic decoder [11] prior to the simulation of the virtual loudspeakers, although the two operations can still be lumped into a parallel bank of 4 filters L_i as shown in Figure 3. Limiting the B format to its first-order components, all directions on a sphere can be encoded by use of 4 signals only (conventionally denoted W, X, Y and Z):

$$\begin{aligned} g_1(\theta, \varphi) &= W(\theta, \varphi) = 1 \\ g_2(\theta, \varphi) &= X(\theta, \varphi) = \cos(\varphi)\cos(\theta) \\ g_3(\theta, \varphi) &= Y(\theta, \varphi) = \cos(\varphi)\sin(\theta) \\ g_4(\theta, \varphi) &= Z(\theta, \varphi) = \sin(\varphi) \end{aligned} \tag{4}$$

These four signals completely characterize the sound field (acoustic pressure and velocity) at the position of the center of the head. This information is sufficient to determine the pressure signals at the two ears for frequencies where the wavelength is substantially larger than the size of the head, i. e. below about 700 Hz [12]. As frequency increases, growing errors can be expected in the reconstruction of both interaural time differences and HRTF spectral cues.

(II) Projection over a set of spectral functions

Principal Component Analysis (PCA) has been used by several authors for representing HRTF spectra, measured on a population of subjects, by decomposition over a set of orthogonal spectral functions [13, 14]. The initial studies were applied to logarithm magnitude frequency responses (dB spectra). This provides a decomposition of the form (2) for the HRTFs represented in dB, therefore yielding a possible binaural synthesis implementation by cascading the basis filters. However, in such an implementation, the spatial variation cannot be conveniently implemented and the computation of the basis filters cannot be shared between all sound sources.

However, applying the PCA to a linear representation of the HRTFs (complex frequency spectrum or impulse response) leads to a parallel decomposition of the form (2). This method was proposed in [15] for implementing the binaural synthesis by the "prefiltering" approach.

(III) Projection over a set of spatial functions

Although PCA is usually viewed as a method for decomposing HRTFs over a set of basis filters, the result can be viewed just as well as a decomposition over a set of spatial functions, as can be seen from Equation (2). In this approach, neither the spectral basis functions nor the spatial basis functions are known in advance: they are determined by an optimization process, which aims at explaining the variations observed in the data.

Alternately, the set of spatial functions $g_i(\theta, \varphi)$ could be chosen arbitrarily. If this set is orthogonal, an approximate reconstruction of the directivity pattern of the ear can be obtained by orthogonal projection of the set of measured HRTFs over the different spatial basis functions. This approach has been described in [16], where the spatial functions used are the spherical harmonics, and the projection is performed in the time domain on the measured impulse responses. The HRTF data set is thus represented as an impulse response in which every coefficient is a function of the spatial variables θ and φ, and this function is represented as a weighted combination of spherical harmonics. As a result, a sampled time function is associated to each spherical harmonic $g_i(\theta, \varphi)$, which defines the impulse response of the corresponding filter $L_i(f)$ in Equation (2). The decomposition could be obtained similarly by applying the projection in the frequency domain.

This technique provides a possible method for decoding the B-format directly for reproduction over headphones, bypassing the steps of Ambisonic decoding and virtual louspeaker simulation. Still, this approach will entail the same limitations as mentionned earlier above 700 Hz.

2 Decoupling ITD and spectral cues: the "binaural B format"

In any of the approaches (I)-(III) described in the previous section for achieving the decomposition (2), the performance is limited whenever an attempt is made to reconstruct mixed-phase HRTFs (including the delay component) by linear combination of other mixed-phase HRTFs or mixed-phase functions with different delays. The consequence is that a large number of channels (and filters) is necessary to achieve a close approximation of the original HRTFs (depending on the context, this may mean a large number of virtual loudspeakers, a large number of principal components, or high-order spherical harmonics).

The reconstruction of the HRTF magnitude spectra can be improved, with a reduced number of intermediate encoding channels, by applying the linear decomposition (2) to the minimum-phase components $\underline{L}(\theta, \varphi, f)$ and $\underline{R}(\theta, \varphi, f)$ of the left and right HRTFs. The interaural time difference (ITD) must then be synthesized explicitly at the panning stage for each sound source. Two independent sets of N signals must be used to encode the left and right ear signals, and $2N$ filters are required for transcoding to binaural format in the "post-filtering" approach.

Figure 5 illustrates this method in the case where the minimum-phase HRTFs are decomposed over spherical harmonics limited to first order. The directional encoding and mixing of the different source signals uses an 8-channel "binaural B format", and 8 filters are used to decode this format into a binaural output signal. A spherical head model provides a good approximation for synthesizing the ITD [2]:

$$ITD(\theta, \varphi) = r/\mathrm{c}\ [\ \arcsin(\cos(\varphi)\sin(\theta)) + \cos(\varphi)\sin(\theta)\ ] \qquad (5)$$

where r denotes the head radius (or distance between the two ears) and c is the speed of sound.

Figure 6 illustrates the performance of the method for reconstructing the HRTF magnitude spectra in the horizontal plane ($\varphi = 0$). For this reconstruction, only 3 channels per ear are necessary, since the Z channel is not used. The original data are diffuse-field equalized HRTFs derived from measurements on a KEMAR manikin. Due to the limitation to first-order harmonics, the reconstruction matches the original magnitude spectra reasonably well up to about 2 or 3 kHz, but the performance tends to degrade with increasing frequency. For large-scale applications, a gentle degradation at high frequencies can be acceptable, since inter-individual differences in HRTFs typically become prominent at frequencies above 5 kHz (see, e. g., [17]).

Recording in binaural B format

Encoding a sound in binaural B format approximately synthesizes a free-field recording using two Soundfield microphones located at the notional position of the two ears (in the absence of the head). The synthetic signal and the recorded signal differ only in the value of the ITD for sounds away from the median plane. Due to the absence of the head in the recording technique, the recorded ITD verifies:

$$ITD(\theta, \varphi) = 2r/\mathrm{c}\ \cos(\varphi)\sin(\theta) \qquad (6)$$

The difference betwen the synthetic ITD, Equation (5), and the recorded ITD, Equation (6), can be reduced by adjusting the distance between the two microphones to be slightly larger than the distance between the two ears of the listener.

The binaural B format recording technique is compatible with currently existing 8-channel digital recording media. The recording can be decoded for reproduction over headphones through the bank of 8 filters L_i shown on Figure 5, or decoded over two or more loudspeakers using methods to be described below. Before decoding, additional sources can be panned in binaural B format and mixed into the recording.

Decoding over headphones

An advantage of a binaural B format recording over a two-channel dummy head recording is that it does not contain spectral HRTF features. These features are only introduced at the decoding stage by the basis filters L_i. Contrary to a conventional binaural recording, a binaural B format recording can conceivably allow listener-specific adaptation at the reproduction stage, in order to reduce the occurrence of front-back reversals and in-head or elevated localization of frontal sound events. However, the adaptation is limited by the fact that ITD cues are encoded in the recording, and cannot be subsequently altered to simulate a change in microphone spacing or head size. For the same reason, rotations of the listener's head cannot be compensated for after mixing, contrary to the use of a single B format signal as described in [7] and [8].

Listener-specific adaptation can be achieved even more effectively in the context of a real-time digital mixing system, and can be conveniently implemented as it only involves the correction of the head radius r for the synthesis of ITD cues and the adaptation of the 4 binaural decoding filters L_i. If diffuse-field equalization is applied to the headphones and HRTFs (and therefore to the reconstruction filters L_i), the adaptation only needs to address direction-dependent features related to the morphology of the listener, rather than variations in HRTF measurement apparatus and conditions [17].

Decoding over two or more loudspeakers

The binaural output signal can be transcoded for reproduction over two loudspeakers by use of a transaural cross-talk canceller. Assuming left/right symmetry in the positions of the loudspeakers, this conversion only requires two additional filters [18, 1, 3]. This technique can produce convincing lateral sound images with a frontal pair of loudspeakers, covering azimuths up to about ±120°. However, lateral images tend to collapse into the loudspeakers in response to rotations and translations of the listener's head. The technique is also less effective for sound events assigned to rear or elevated positions, even when the listener sits at the "sweet spot".

An advantage of the binaural B format is that it contains information for discriminating rear sounds from frontal sounds. This property can be exploited in order to overcome the limitations of 2-channel transaural reproduction, by decoding over a 4-channel loudspeaker setup. The 4-channel decoding network is shown in figure 7, and makes use of the sum and difference of the W and X signals.

The binaural signal is decomposed as follows:

$$L(\theta, \varphi, f) = LF(\theta, \varphi, f) + LB(\theta, \varphi, f) \tag{7}$$

where *LF* and *LB* are the "front" and "back" binaural signals, defined by:

$$\begin{aligned} LF(\theta, \varphi, f) &= 0.5 \; \{ \; [W(\theta, \varphi) + X(\theta, \varphi)] \; [L_W(f) + L_X(f)] + Y(\theta, \varphi) \; L_Y(f) + Z(\theta, \varphi) \; L_Z(f) \; \} \\ LB(\theta, \varphi, f) &= 0.5 \; \{ \; [W(\theta, \varphi) - X(\theta, \varphi)] \; [L_W(f) - L_X(f)] + Y(\theta, \varphi) \; L_Y(f) + Z(\theta, \varphi) \; L_Z(f) \; \} \end{aligned} \tag{8}$$

It can be verified that $LB = 0$ for $(\theta, \varphi) = (0, 0)$ and that $LF = 0$ for $(\theta, \varphi) = (\pi, 0)$. The network of Figure 7 is designed to eliminate front-back confusions, by reproducing frontal sounds over the front loudspeakers and rear sounds over the rear loudspeakers, while elevated or lateral sounds are reproduced via both pairs of loudspeakers. This significantly improves the reproduction of lateral, rear or elevated sound images compared to a 2-channel loudspeaker setup (or to 4-channel loudspeaker reproduction using conventional pairwise amplitude panning or Ambisonic techniques). The listener is also allowed to move more freely than with 2-channel loudspeaker reproduction.

This sum and difference method has been used earlier for separating sounds to be directed to front a rear louspeaker pairs [19]. The recording technique proposed by J. Bruck [19] defines a 4-channel binaural format (2 channels per ear), and makes use of two M-S microphones placed on the sides of a sphere (which plays the role of the dummy head). The front and back signals are derived by computing *M+S* and *M–S*. It can be verified that these signals coincide with the signals (*LF*, *RF*) and signals (*LB*, *RB*) as defined by Figure 7 or Equation (8). This means that the binaural B format encoding technique shown in Figure 5 is also an efficient technique for panning and mixing sounds in the 4-channel binaural format proposed in [19]. By exploiting the *Z* component, a similar approach can be used to decode the binaural B format over a 3-D loudspeaker setup (comprising loudspeakers above or below the horizontal plane).

Conclusion

By separating the left and right ear signals at the encoding stage, the 8-channel binaural B format offers –when compared to "coincident" encoding schemes– exact recontruction of ITD cues and the possibility of improved recontruction of HRTF spectral cues by use of a decoding filter bank. The choice of first-order spherical harmonics as the spatial basis functions provides an efficient panning and mixing scheme for binaural synthesis of 3-D sound scenes containing multiple sources, while allowing compatible mixing with a recording using a pair of non-coincident Soundfield microphones, and decoding over multichannel loudspeaker setups.

Acknowledgments

Methods described in section 2 of this paper are covered by a patent application. The use of three filters instead of four, as shown in Figure 2, to allow switching HRTF filters between directions was suggested by Patrick Boussard.

References

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

[2] Larcher, V., and Jot, J.-M. (1997). Techniques d'interpolation de filtres audio-numériques. Application à la reproduction spatiale des sons sur écouteurs. *Proc. 4th French Congress on Acoustics*, 97-100.

[3] Gardner, W. G. (1997). 3-D Audio Using Loudspeakers. Ph. D. Thesis. Massachussets Institute of Technology - Media Lab.

[4] Välimäki, V. (1995). Discrete-Time Modelling of Acoustic Tubes Using Fractional Delay Filters. Helsinky University of Technology, Laboratory of Acoustics and Audio Signal Processing. Report 37.

[5] Huopaniemi, J. and Karjalainen, M. (1997). Review of digital filter design and implementation methods for 3-D sound. Presented at the 102nd Conv. Audio Eng. Soc. (preprint 4461).

[6] Kendall, G., Martens, W., Freed, D., Ludwig, D., and Karstens, R. (1986). Image-model reverberation from recirculating delays. Presented at the 81st Conv. Audio Eng. Soc. (preprint 2408).

[7] Malham, D. G. (1993). 3-D sound for virtual reality using ambisonic techniques. *Proc. 3rd Annual Conf. on Virtual Reality* (addendum).

[8] Travis, C. (1996). A virtual reality perspective on headphone audio. Presented at the Audio Eng. Soc. U. K. Conference 'Audio For New Media'.

[9] Farrah, K. (1979). The Soundfield microphone. *Wireless World*, 85: 48-50; 99-102.

[10] Theile, G. and Plenge, G. (1977). Localization of lateral phantom sources. *J. Audio Eng. Soc.* 25(4).

[11] Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc.* 33(11).

[12] Bamford, J., Vanderkooy, J. (1995). Ambisonic Sound For Us. Presented at the 99th Conv. Audio Eng. Soc. (preprint 4138).

[13] Martens, W. (1987). Principal components analysis and resynthesis of spectral cues to perceived direction. *Proc. International Computer Music Conference*, 274-281.

[14] Kistler, D. and Wightman, F. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.* 91(3).

[15] Marolt, M. (1995). Proc. IEEE 1995 Workshop on Applications of Signal Processing to Audio and Acoustics.

[16] Evans, M., Angus, J., and Tew, A. (1997). Spherical harmonic spectra of head-related transfer functions. Presented at the 103rd Conv. Audio Eng. Soc. (preprint 4571).

[17] Larcher, V., and Jot, J.-M. (1998). Equalization methods in binaural technology. Presented at the 105th Conv. Audio Eng. Soc.

[18] Cooper, D. H., and Bauck, J. L. (1989). Prospects for transaural recording. *J. Audio Eng. Soc.* 37(1/2).

[19] Bruck, J. (1996). Presented at the 101st Conv. Audio Eng. Soc.

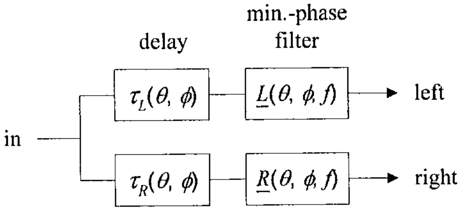

Figure 1: Implementation of a binaural encoding module for simulating a static or moving sound source in free field.

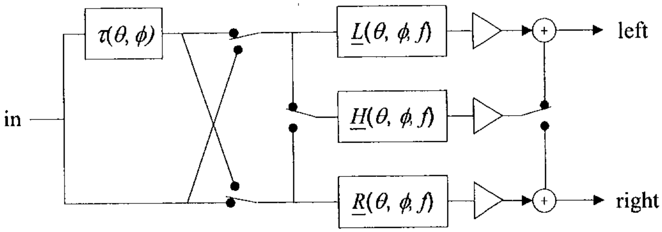
$\tau(\theta, \phi)$
$\underline{L}(\theta, \phi, f)$
+
left
in
$\underline{H}(\theta, \phi, f)$
$\underline{R}(\theta, \phi, f)$
+
right

Figure 2: Optimized implementation of binaural encoding module for simulating a moving sound source. The filter $\underline{H}$ is used for switching the filter $\underline{L}$ or the filter $\underline{R}$, alternatively, by a cross fading technique.

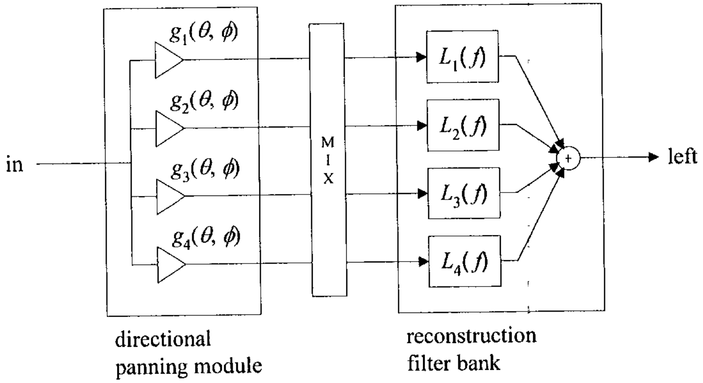

Figure 3: Binaural encoding of multiple sounds by the "post-filtering" approach, for reproduction over headphones (left signal only). The filter bank acts as an "output decoder" applied after pannig and mixing all signals.

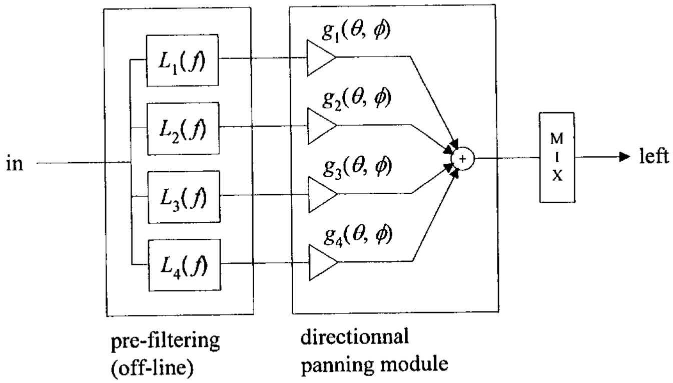

Figure 4: Binaural encoding of multiple sounds by the "pre-filtering" approach, for reproduction over headphones (left signal only). The filters are applied off-line on each source signal.

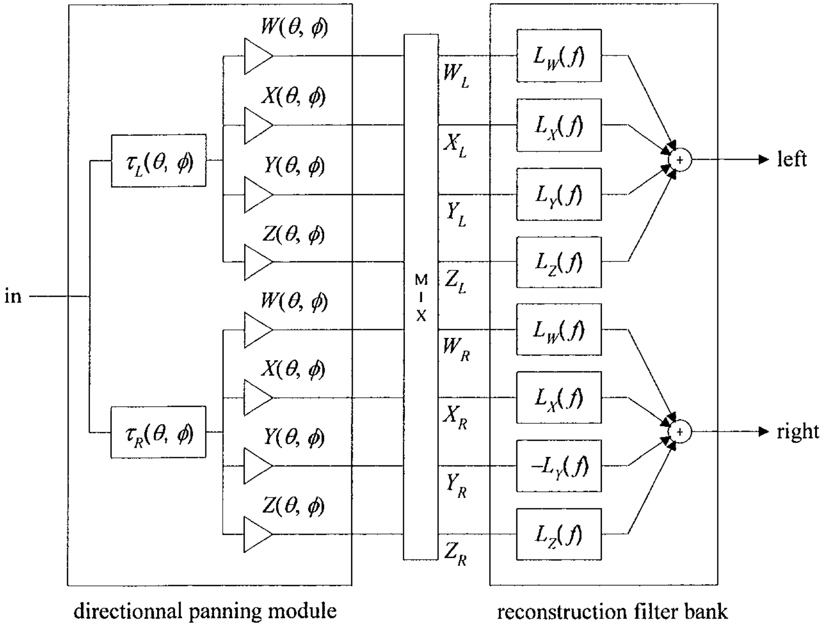

Figure 5: Implementation of the "post filtering" approach using the binaural B format.

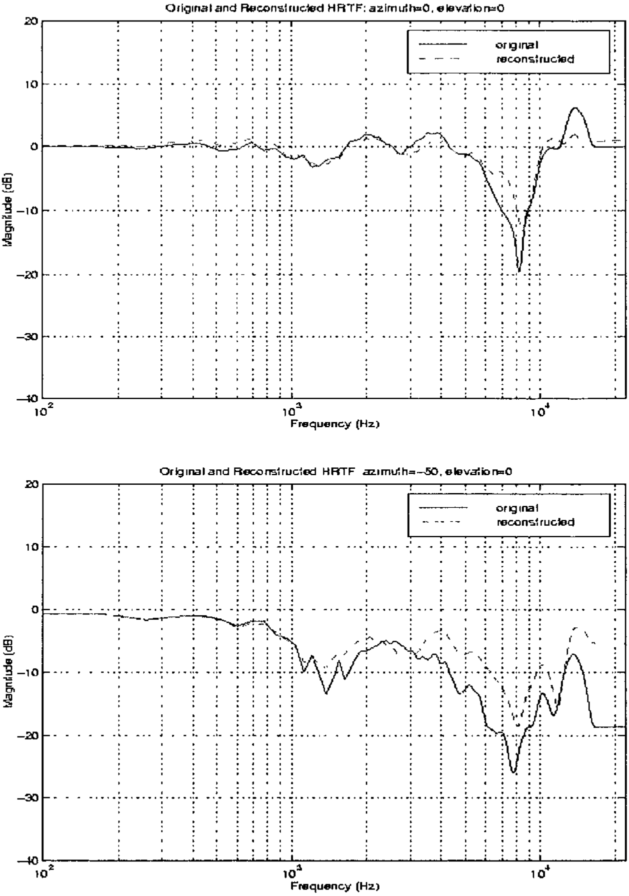

Figure 6: Comparison of original and reconstructed HRTF in the horizontal plane by projection over spherical harmonics *W*, *X* and *Y*, for two values of the azimuth angle θ.
The original HRTF (solid line) is smoothed to semitone resolution.
Top: $\theta = 0$ degrees. Bottom: $\theta = -50$ degrees.

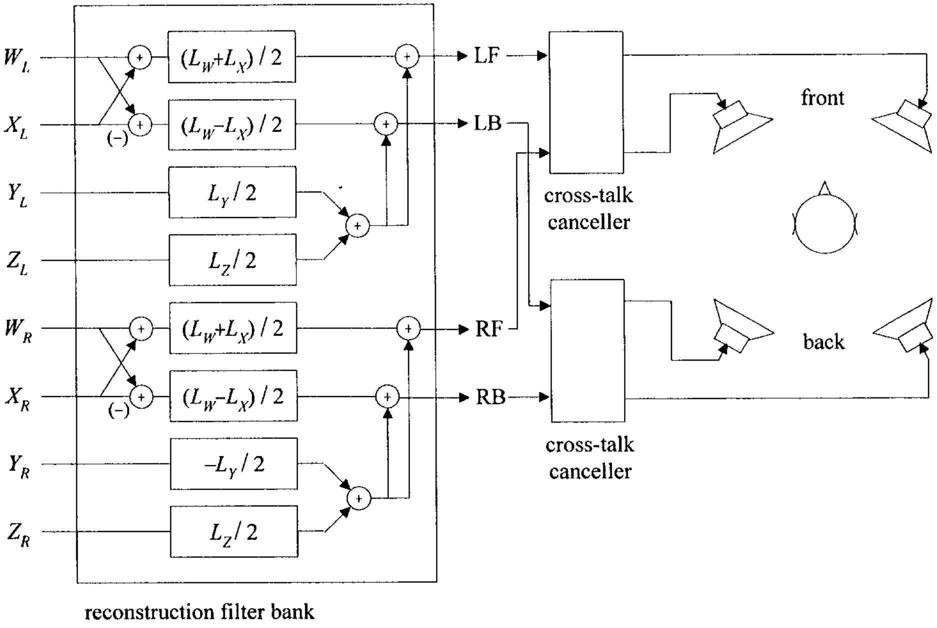

Figure 7: Decoder network for 3-D sound reproduction over 4 loudspeakers based on the binaural B format. The binaural signal is decomposed into a front signal (*LF*, *RF*) and a rear signal (*LB*, *RB*), which are respectively reproduced over the front loudspeaker pair and the rear loudspeaker pair, via two cross-talk cancellers. For simplicity, the frequency variable is omitted in the description of the filter bank