# Scalable Multichannel Coding with HRTF Enhancement for DVD and Virtual Sound Systems*

**M. O. J. HAWKSFORD**

*Centre for Audio Research and Engineering, University of Essex, UK CO4 3SQ*

A scalable and reverse compatible multichannel method of spatial audio using transaural coding designed for multiple-loudspeaker feeds is described with a focus on attaining optimum ear signals. A Fourier transform method for computing HRTF matrices is employed, including the generation of a subset of band-limited reproduction channels. Applications considered embrace multichannel audio, DVD, virtual reality, and telepresence.

## 0 INTRODUCTION

The purpose of this paper is to investigate how transaural processing can enhance conventional multichannel audio both by embedding perceptually relevant information and by improving image stability using additional loudspeakers integrated with supplementary digital processing and coding. The key objective is to achieve scalability in spatial performance while retaining full compatibility with conventional multichannel formats. This enables the system in its most basic form with unprocessed loudspeaker feeds to be used in a conventional multichannel installation. However, by appropriate signal processing additional loudspeaker feeds can be derived, together with the option of exploiting buried data to extract more signals in order to improve spatial resolution. The system is therefore hierarchical in terms of number of loudspeakers, channels, and ultimately spatial resolution, while in its simplest incarnation it remains fully compatible with the system configurations used with multichannel DVD-A and SACD replay equipment.

The multichannel capabilities of DVD[1] technology [1], [2] were designed to enhance stereo[2] sound reproduction by offering surround image and improved envelopment capabilities. Normally multichannel audio encoded onto DVD assumes the ITU standard of a five-loudspeaker configuration driven by five discrete wide-band "loudspeaker feeds." However, a limitation of this system is the lack of a methodology to synthesize virtual images capable of three-dimensional audio (that is, a perception of direction, distance, and height together with acoustic envelopment) rather than just "sound effects" often (although not exclusively) associated with surround sound in a home theater context. The ITU five-channel loudspeaker configuration can also be poor at side image localization, although this deficiency is closely allied to a sensitivity to room acoustics. Nevertheless, DVD formats still offer only six discrete channels, which if mapped directly into loudspeaker feeds remain deficient in terms of image precision, especially if height and depth information is to be encoded.

The techniques described in this paper support scalable spatial audio that can remain compatible with conventional multichannel systems. It is shown that in this class of system, under anechoic conditions, signal processing can be used to match theoretically the ear signals to either a real or an equivalent spatially synthesized sound source. Also, in order to improve image robustness, directional sound-field encoding is retained as exploited in conventional surround sound to match the image synthesized through transaural processing. It may be argued that as the number of channels is increased, there is convergence toward wavefront synthesis [3], where by default optimum ear-signal reconstruction is achieved. However, the proposed system is positioned well into the middle territory[3] and is far removed from the array sizes required for broadband wavefront synthesis. Consequently, from the perspective of wavefront synthesis the transition frequency above which spatial aliasing occurs is located at a relatively low frequency, implying that for the proposed system the core concepts of wavefront synthesis do not apply

* Presented at the 108th Convention of the Audio Engineering Society, Paris, France, 2000 February 19–22; revised 2001 October 10 and 2002 July 15.

[1] Includes both DVD-A and SACD formats.

[2] Of Greek origin, meaning solid, stereo is applicable universally to multichannel audio.

[3] A range of 5 to 32 loudspeakers is suggested.

over much of the audio band.

It is emphasized that an *n*-channel system does not necessarily imply *n* loudspeakers. Indeed, as is well known, it is possible for a two-loudspeaker system to reproduce virtual-sound sources [4], while using more than *n* loudspeakers can help create a more robust and stable illusion. Also, the mature technology of Ambisonics [5]–[7] is scalable and can accommodate both additional loudspeakers and information channels. However, here the encoding is hierarchical in terms of spatial spherical harmonics, although no attempt is made to reconstruct the ear signal directly at the listener. Consequently the approach taken in this paper differs in a number of fundamental aspects from that of Ambisonics, especially since there is no attempt to transform a sound field directly into a spherical harmonic set. Thus it remains for future work to establish the relative merits of these approaches although, because similar loudspeaker arrays are used, there is no fundamental compromise should the system be used either for Ambisonics or for conventional surround sound encoded audio.

The method of spatial audio described in this paper uses a conventional loudspeaker array to surround the listener and to reproduce a directional sound field. In addition, ear signals are simultaneously synthesized using head-related transfer functions (HRTFs) matched to the source image, where it is assumed in all cases that loudspeaker transfer functions have been equalized or otherwise taken into account. A number of examples illustrate the computational methods, which include pairwise transaural image synthesis[4] reported in earlier work, where some preliminary experimental results were also discussed [8]–[10] to establish the efficacy of the method. This technique is especially well matched to multichannel multiloudspeaker installations, where transaural coding can be applied during encoding and recording while processing within the decoder located within the reproduction system can accommodate both additional loudspeakers and loudspeaker positions that differ from those assumed at the encoder. Consequently for an *n*-channel system it is straightforward to employ only *n* loudspeakers, although additional loudspeaker feeds can be derived, while still retaining correct ear signals, either by using matrix techniques or deterministically within the DVD-A format using additional embedded code. However, it is emphasized that in the simplest configuration, using only direct loudspeaker feeds and provided the loudspeakers are correctly located, there are no additional decoding requirements and the system remains fully compatible with all existing recordings.

Alternative technologies such as Ambisonics [11] have used fewer loudspeakers together with sophisticated matrix encoding. Also, there has been substantial research into perceptually based processing to reconstruct a three-dimensional environment using only two channels and two loudspeakers. More recently DOLBY EX[5] has been introduced as a means of synthesizing a center rear channel using nonlinear Prologic[6] processing applied to the rear two channels of a five-channel system. However, this technology is aimed principally at surround sound as conceived for cinema and home theater, with a bias toward sound effects and ambience creation. Nevertheless there exists a grey area between cinema applications, music reproduction, gaming applications, and the synthesis of virtual acoustics, especially as at their core the same multichannel carriers can link all systems. It is therefore not unreasonable to anticipate some degree of convergence as similar theoretical models apply. Also, with conventional multichannel technology it is often the listening environment and the methods used to craft the audio signals that impose the greatest performance limitations.

Multichannel stereo on DVD allows for improved methods of spatial encoding that can transcend the common studio practice of using just pairwise amplitude panning with blending to mono. It is conjectured that by including perceptually motivated processing, three-dimensional "soundscapes" can be rendered rather than just peripheral surround sound. Complex HRTF data by default encapsulate all relevant spatial information [that is, interaural amplitude difference (IAD), interaural time difference (ITD), and directional spectral weighting] and form a generalized approach. However, to reduce signal coloration, a method of HRTF equalization is proposed with an emphasis on characterizing the interear difference signal computed in the lateral plane. The extension to height information in the equalized HRTFs is also discussed briefly.

In Section 5.2.1 a special case is presented for narrow subtended angle, two-channel stereo where it is shown that ear signals derived from a real acoustic source located on the arc of the loudspeakers can be closely synthesized using a mono source with amplitude-only panning. Critically in this example, the HRTFs are defined by the actual locations of the loudspeakers, and so are matched automatically to an individual listener's HRTF characteristics. This is an important aspect of the proposal, which is directly extendable to the method of pairwise association where mismatch sensitivity between listener HRTFs and target HRTFs is reduced. This approach also encapsulates succinctly the principles of two-channel amplitude-only panning stereo while exposing inherent errors as the angle between the loudspeakers is increased.

To summarize, four core elements constitute the proposed scalable and reverse compatible spatial audio system:

- A vector component of the sound field is produced as a loudspeaker array surrounds the listener following conventional multichannel audio practice.
- Pairwise transaural techniques are used to code directional information and to create ear signals matched to the required source signal.
- Matrix processing can increase the number of loudspeakers used in the array while simultaneously preserving the ear signals resulting from transaural processing and nonoptimum loudspeaker placements.
- Embedded digital code[7] [12] is used to create additional channels for enhanced resolution while remaining com-

[4] Subject to a British Telecommunications patent application.
[5] Dolby Laboratories, channel extension technology to AC-3 perceptual coding.
[6] Registered trademark of Dolby Laboratories.
[7] Applicable only to the DVD-A format.

patible with the basic system already enhanced by pairwise transaural processing.

## 1 HRTF NOTATION

A set of HRTFs is unique to an individual and describes a continuum of acoustic transfer functions linking every point in space to the listener's ears. HRTFs depend on the relative position of the source to the listener and are influenced by distance, reflection, and diffraction around the head, pinna, and torso. In this paper HRTFs were derived from measurements taken at BT Laboratories (BTL) utilizing an artificial head and small microphones mounted at the entrance of each ear canal. Measurements of head-related impulse responses (HRIRs) were performed in an anechoic chamber at 10° intervals using a maximum-length-sequence (MLS) excitation, and the corresponding HRTFs were computed using a time window and the Fourier transform.

To define the nomenclature used for various HRTF subfunctions, consider the arrangement shown in Fig. 1, where the listener's ears are labeled A (left) and B (right) when viewed from above. In a sound reproduction system all sound sources and loudspeakers have associated pairs of HRTFs, uniquely linking them to the listener, whereas in this paper these transfer functions are called the HRTF coordinates for each object given. In Fig. 1 the single sound source $X$ has the HRTF coordinates $\{h_{xa}, h_{xb}\}$, while the three loudspeakers 1, 2, and $n$, with arbitrary positions, have the coordinates $\{h_a(1), h_b(1)\}$, $\{h_a(2), h_b(2)\}$, and $\{h_a(n), h_b(n)\}$. In specifying the loudspeaker HRTF coordinates, a left–right designation can be included when the loudspeaker array is known to be symmetrical about the centerline. Consequently $h_{la}(r)$ denotes the HRTF between the left-hand loudspeaker $r$ and the left-hand ear A. However, for arrays having only three symmetrically positioned loudspeakers (left, center, and right) a simpler notation is used in Section 3, namely, $\{h_{la}, h_{lb}\}$, $\{h_{ca}, h_{cb}\}$, and $\{h_{ra}, h_{rb}\}$. It should be observed that in typical HRTF calculations, such as the evaluation of the positional transfer functions $G_R$ and $G_L$ used in transaural signal processing (see Section 3.1), ear-canal equalization need not be incorporated, provided the same set of HRTFs is used for both loudspeaker and image locations. Then the ear-canal transfer functions cancel, assuming they are not directionally encoded. For example, in Section 4 Eqs. (23a) and (23b) describe typical transaural processing to derive the positional transfer functions $G_R$ and $G_L$, where any transfer function components common to all HRTFs cancel.

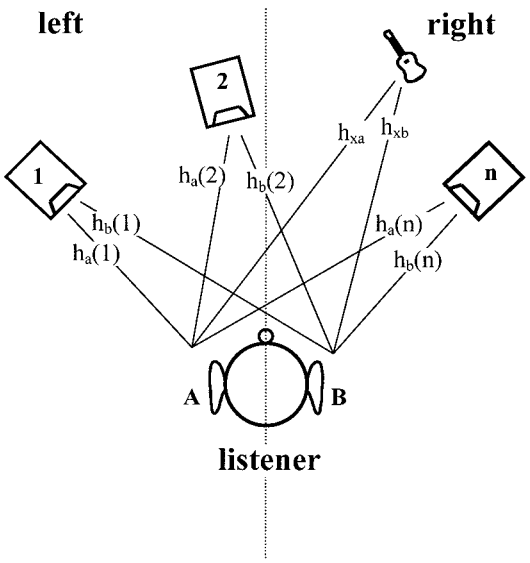



Fig. 1. Definition of HRTF notation for $n$ loudspeakers positioned arbitrarily around the listener.

## 2 HRTF EQUALIZATION

The HRTFs used in transaural processing reveal frequency response variations that may contribute tonal coloration when sound is reproduced. In this section a strategy for equalization is studied that reduces the overall spectral variation, yet retains the key attributes deemed essential for localization. A simplified form of HRTF is also defined, which can prove useful in multiloudspeaker systems.

### 2.1 Methods of Equalization

When sound is reproduced over a conventional multiloudspeaker array, where for example a signal is spatialized using pairwise amplitude panning, equalization as a function of direction is not normally employed. In such a system sound is perceived generally as uncolored, even though the ears, head, and torso impose direction-specific spectral weighting. However, although HRTFs used in transaural processing take account of both the source location and the loudspeakers, reducing frequency response variations can ameliorate tonal variance, which may become accentuated as the phantom image moves away from the loudspeaker locations and when the listener turns away from the optimum forward orientation. Also, systems are rarely optimally aligned and exhibit sensitivity to small head motions, both of which map into frequency response errors in the reconstructed ear signals. Consequently the aim is to introduce minimum spectral modifications commensurate with achieving spatialization.

It is proposed that for image localization within the lateral plane the relationship between the complex interaural difference signal and the signal components common to both ears is the critical factor. This conjecture is based on the premise that for lateral images, spectral components common to both ears relate closely to the source spectrum whereas the interaural spectrum is strongly influenced by source direction. Consequently it is argued that modification to the common spectrum causes principally tonal coloration, whereas the relationship between common spectrum and interaural spectrum is more critical to localization, even though spectral cues embedded in the source can induce an illusion of height. This approach may be extendable to include height localization, although it is recognized that additional spectral weighting of the monaural component can be required following, for example, the boosted-band experiments performed by Blauert [13].

### 2.1.1 Lateral-Plane HRTF Equalization

The proposed method of lateral-plane HRTF equalization first transforms each HRTF pair into sum and difference (M-S) coordinates and then performs equalization on the corresponding pair by dividing by the corresponding sum spectrum. It is proposed that all HRTFs in a set should be equalized using this technique in order to maintain relative group delay and, with appropriate weighting, relative level.

As defined in Section 1, let the HRTFs for a given source location $X$ be $h_{xa}$ and $h_{xb}$, and let the corresponding complex sum and difference transforms be $\mathrm{HSUM}_x$ and $\mathrm{HDIFF}_x$. Thus

$$\mathrm{HSUM}_x = h_{xa} + h_{xb} \tag{1}$$

$$\mathrm{HDIFF}_x = h_{xa} - h_{xb} \, . \tag{2}$$

Four methods of HRTF equalization that match this objective are identified, where $\{h_{xea}, h_{xeb}\}$ are the resulting HRTFs after equalization.

*Method 1*: Equalization by the modulus of the complex sum spectrum,

$$h_{xea} = \frac{h_{xa}}{|h_{xa} + h_{xb}|} W_{nx} \tag{3a}$$

$$h_{xeb} = \frac{h_{xb}}{|h_{xa} + h_{xb}|} W_{nx} \, . \tag{3b}$$

*Method 2*: Equalization by complex sum spectrum,

$$h_{xea} = \frac{h_{xa}}{h_{xa} + h_{xb}} W_{nx} \tag{4a}$$

$$h_{xeb} = \frac{h_{xb}}{h_{xa} + h_{xb}} W_{nx} \, . \tag{4b}$$

*Method 3*: Equalization by the derived minimum-phase spectrum of the complex sum spectrum,

$$h_{xea} = \frac{h_{xa}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa} + h_{xb})))))} W_{nx} \tag{5a}$$

$$h_{xeb} = \frac{h_{xb}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa} + h_{xb})))))} W_{nx} \, . \tag{5b}$$

Here $W_{nx}$ are the normalization coefficients calculated to maintain the relative levels after equalization of all HRTF coordinates in the set. Each form of equalization delivers identical magnitude spectra in the HRTFs, although there are variations in the time-domain waveforms resulting from phase response differences. To illustrate these variations, consider an example HRTF pair corresponding to a nominal 30° off-axis image source. Fig. 2(a) shows the measured HRIRs, whereas Fig. 2(b)–(d) presents the impulse responses resulting from each form of equalization in order that both pre- and postring can be compared. Fig. 3 shows the corresponding amplitude spectra before and after equalization, and Fig. 4 illustrates the sum and difference spectra, again before and after equalization.

In selecting a potential equalization strategy, it is a necessary condition that the relative time difference between HRTF pairs be maintained. Also, the time-domain waveforms should not accentuate or exhibit excessive prering or postring, as this can produce unnatural sound coloration. Although each equalization method meets the principal objective, the technique of forging the denominator from the minimum phase of the sum spectrum yields results with minimum prering. In essence, the minimum-phase information common to both ear signals is removed, leaving mainly excess phase components to carry the essential time-delay information. Equalization using the complex sum spectrum (method 2) also yields results close to the requirements. However, inspection of Fig. 2(c) shows that the right-ear response, which in this case has the greater delay, exhibits prering extending back in time to the commencement of the left HRIR.

However, experience gained with equalization has revealed that certain image locations, particularly toward the center rear, can yield excessive ringing after equalization. Consequently a further equalization variant is proposed. This is similar to method 3, but it differs in the way the sum spectrum is computed and is defined as follows.

*Method 4*: Equalization by the derived minimum-phase sum of the moduli of each complex spectrum,

$$h_{xea} = \frac{h_{xa}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa}) + \mathrm{abs}(h_{xb})))))} W_{nx} \tag{6a}$$

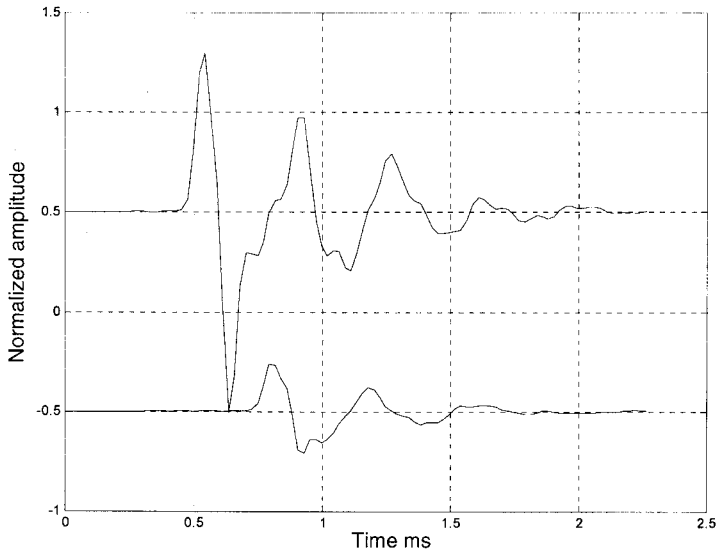
$$h_{xeb} = \frac{h_{xb}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa}) + \mathrm{abs}(h_{xb})))))} W_{nx} \, . \tag{6b}$$

In the denominator this algorithm factors out the interaural time difference between left and right signals, which otherwise map into artificial amplitude response variations in the complex sum spectrum. As such this procedure could be argued to be a better estimator of the common spectrum, as human auditory processing does not sum ear signals directly. Overall the effect on HRTFs is minor. Fig. 5 presents results that should be compared directly with those in Fig. 3.

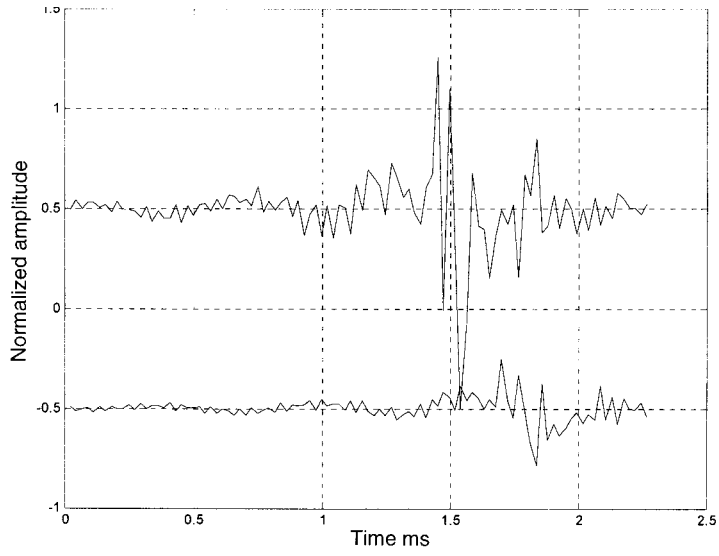
Finally a further variant of equalization is where an average of all sum spectra is formed and the HRTFs are modified following procedures similar to those reported in this section but with particular emphasis on the minimum-phase and sum-of-moduli techniques. However, in this case, since all HRTFs are modified by a common equalization function in a way similar to ear-canal equalization, when positional transfer functions are calculated, their form is unchanged.

### 2.1.2 Equalization with the Addition of Height Cues

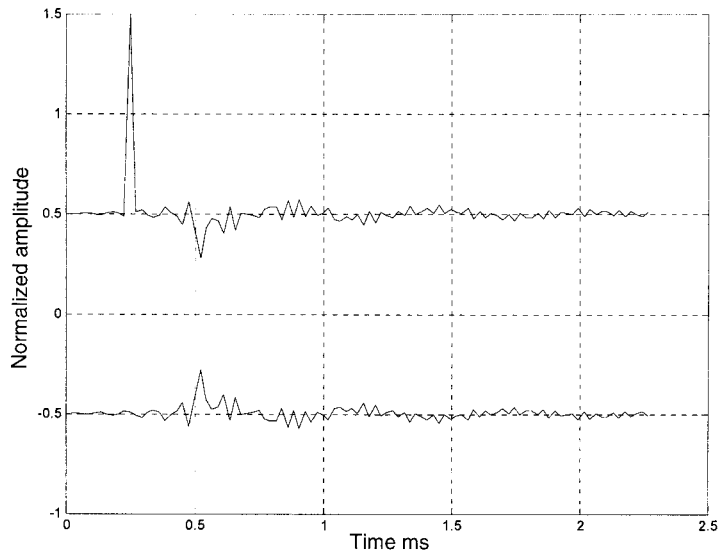
Research by Blauert [13] has shown that by introducing specific frequency-dependent characteristics into the HRTFs a sensation of height is achievable. However, the




(a)

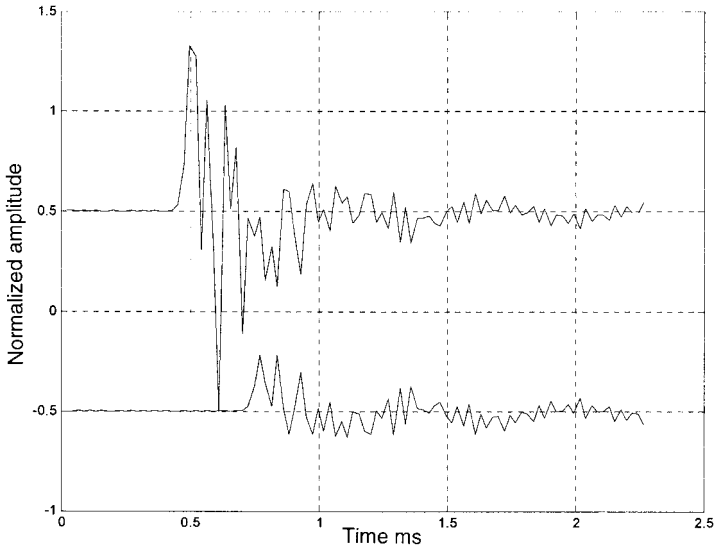



(b)



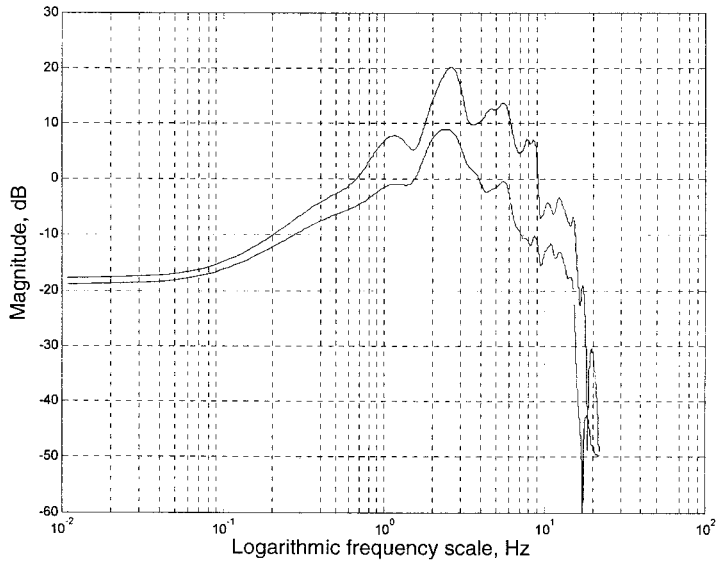

(c)

Fig. 2. Normalized time-domain left–right HRTFs at 30°. Top—left ear; bottom—right ear. (a) Measured. Observe relative time displacement revealing ITD and lack of prering in natural responses. (b) Method 1 equalized. Observe excessive prering that blurs commencement of the two HRIRs. (c) Method 2 equalized. HRIRs exhibit mirror images except for initial impulse. (d) Method 3 equalized. Prering reduced and initial ITD of HRIRs maintained.

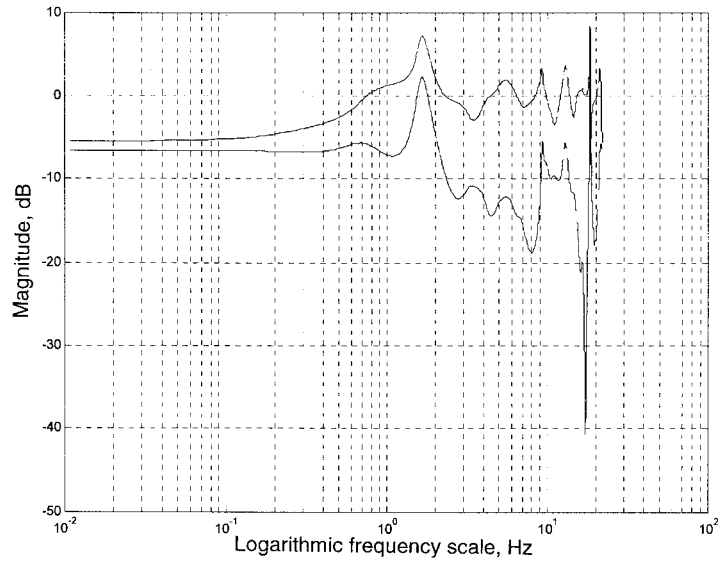



(d)

Fig. 2. *Continued*

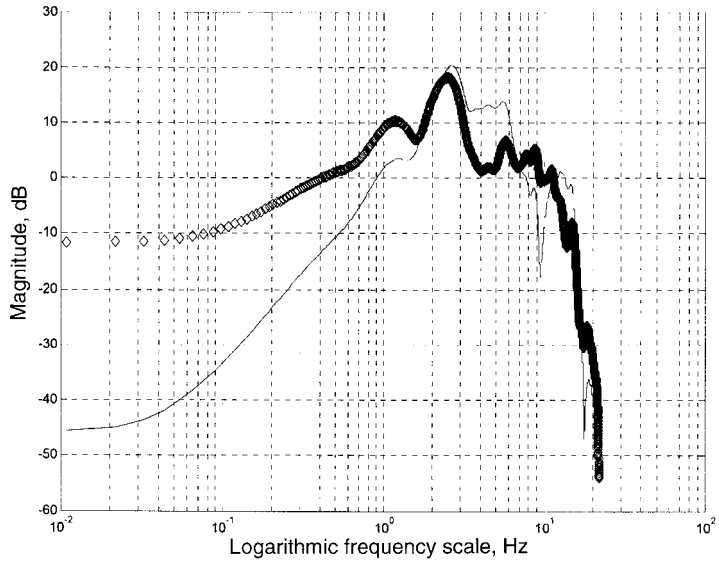



(a)

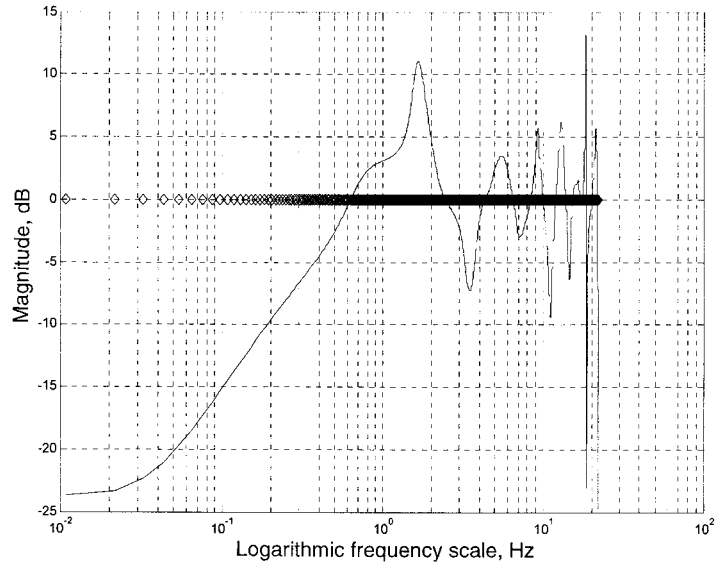



(b)

Fig. 3. Magnitude HRTF pair at 30°. Top—left ear; bottom—right ear. (a) Measured. (b). Equalized. Results are identical for methods 1, 2, and 3.




(a)




(b)

Fig. 4. Magnitude HRTF sum and difference spectra at 30°. Sum—"diamond" line; difference—continuous line. (a) Unequalized. (b) Equalized, applicable to methods 1, 2, and 3. (Note constant-level sum spectrum following equalization.)

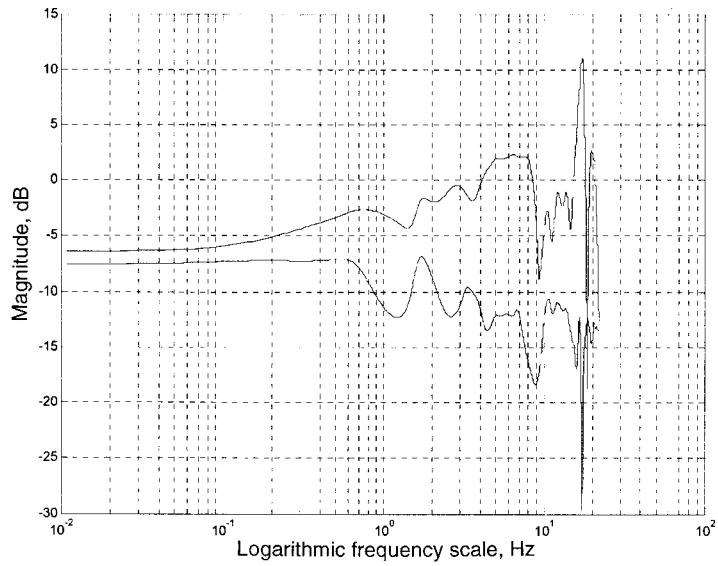



Fig. 5. Equalized HRTF pair at 30°, applicable to method 4 only. Top—left ear; bottom—right ear.

question arises as to whether this modification is compatible with the equalization strategies presented in Section 2.1.1. For example, the following questions need to be considered:

- Is it sufficient to measure the HRTF coordinates only at the required location above the lateral plane and then apply equalization, and will then sufficient information remain buried in the interear difference signal with a unique characterization to discriminate against lateral images with equivalent interaural time differences?
- Does the absolute amplitude response variation, rather than just the difference response variation inherent in the HRTFs, represent a major factor in producing height cues?
- Are there secondary factors, such as ground reflections, which introduce additional cues to aid height localization? Effectively this would require at least two interfering sets of HRTFs to be summed.

A full investigation of these points relating to height is beyond the scope of the present study. However, if the ground reflection model were responsible, then the equalization methods could be applied individually to the direct source and to the ground reflection, with the results combined by taking the path difference into account.

### 2.2 Simplified HRTF Models

In applications where phantom images are positioned close to the locality of the loudspeakers, it may be sufficient to use a simple form of HRTF. This is particularly applicable with multiloudspeaker arrays where a vector component already forms a strong localization clue. Fig. 6 shows a source image at angle θ defined by $h_{xa}$ and $h_{xb}$ with respect to a human head of diameter $d$ meters.

#### 2.2.1 Simple HRTF Model 1

The model ignores head shadowing and assumes that only the interaural time difference is significant. Hence for a source at angle θ from the forward position, the respective HRTF coordinates are approximately

$$h_{xa} = \exp\left\{-j2\pi f\left[T - \frac{d}{2c}\sin(\theta)\right]\right\} \quad (7a)$$

$$h_{xb} = \exp\left\{-j2\pi f\left[T + \frac{d}{2c}\sin(\theta)\right]\right\} \quad (7b)$$

where the velocity of sound is $c$ m/s and $T$ s is the time delay from the source to the center of the head. In this model the advantage of equalization method 3 is evident as no equalization need be applied.

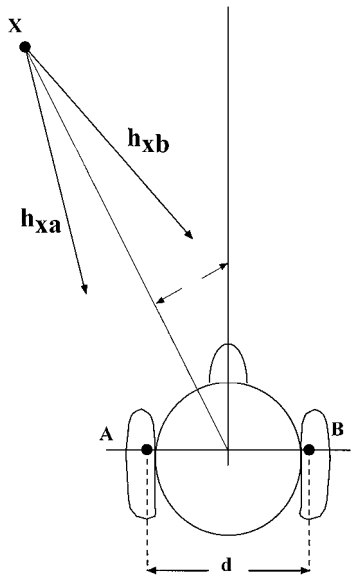



Fig. 6. Sound source $X$ and simplified head model used to derive approximate HRTFs.

#### 2.2.2 Simple HRTF Model 2

In this second model both the front and the back waves are considered, where the back wave results from head defraction. In this representation a wave incident on ear A produces, by head defraction, a secondary signal at ear B. The head diffraction transfer function from ears A to B, $DH_{A-B}(r, \theta, \phi)$, is a function of the direction of the incident wave defined by the spherical coordinates $\{r, \theta, \phi\}$. A similar function linking ears B to A is defined, $DH_{B-A}(r, \theta, \phi)$. Hence for the source HRTF coordinates $\{h_{xa}, h_{xb}\}$,

$$h_{xa} = \exp\left\{-j2\pi f\left[T - \frac{d}{2c}\sin(\theta)\right]\right\} + DH_{B-A}(r,\theta,\phi)\exp\left\{-j2\pi f\left[T + \frac{d}{2c}\sin(\theta)\right]\right\} \quad (8a)$$

$$h_{xb} = \exp\left\{-j2\pi f\left[T + \frac{d}{2c}\sin(\theta)\right]\right\} + DH_{A-B}(r,\theta,\phi)\exp\left\{-j2\pi f\left[T - \frac{d}{2c}\sin(\theta)\right]\right\}. \quad (8b)$$

In a simple model the diffraction transfer functions could be represented as attenuation $DH_k$ with a time delay of approximately the interaural time delay $\Delta T_{A-B}$,

$$DH_{A-B}(r,\theta,\phi) = DH_{B-A}(r,\theta,\phi) \approx DH_k \exp(-j2\pi f \Delta T_{A-B}). \quad (9)$$

## 3 MULTILOUDSPEAKER ARRAYS IN TWO-CHANNEL STEREO

This section introduces variants to two-channel, two-loudspeaker transaural processing to demonstrate how a two-channel signal format can be mapped into $n$ feeds to drive a multiloudspeaker array [14]. It is assumed that more than two loudspeakers are driven simultaneously by signals derived from a single-point sound source, while formal methods show that the correct ear signals can be retained. Besides supporting stand-alone applications, these transformations are relevant in the development of multichannel transaural stereo, as described in Section 5.

The outputs of an *n*-array of loudspeakers combine by acoustical superposition at the entrance to each ear canal. The principal condition for accurate sound localization is that these signals match the signals that would have been generated by a real sound source, both in the static case and in the case for small head rotations. Also, by using several loudspeakers placed to surround the listener, sound-field direction can make the system more tolerant to head motion. Consequently changes in ear signals with head motion match more closely those of a real image.

A static sound source, whatever its size and physical location, produces two ear signals that fully define the event, provided the relative head position to source is fixed. In practice it is possible to generate the correct ear signal from two or more noncoincident loudspeakers that can take any arbitrary position around the head. However, if the position of the head moves, then a change in the ear signals results, which no longer match the phantom image correctly, and a localization error is perceived. In a system of wavefront reconstruction this distortion is minimized, although the penalty is a large number of loudspeakers and channels. However, if a limited number *n* of loudspeakers is used (for example, $n = 12$), then although image position distortion still occurs, the effect is reduced as there is a robust directional component. Also, when a phantom image coincides with a loudspeaker position, the positional distortion as a function of head position tends to zero, although it is debatable whether this is a desirable situation as images from other locations are represented differently in terms of their radiated cones of sound.

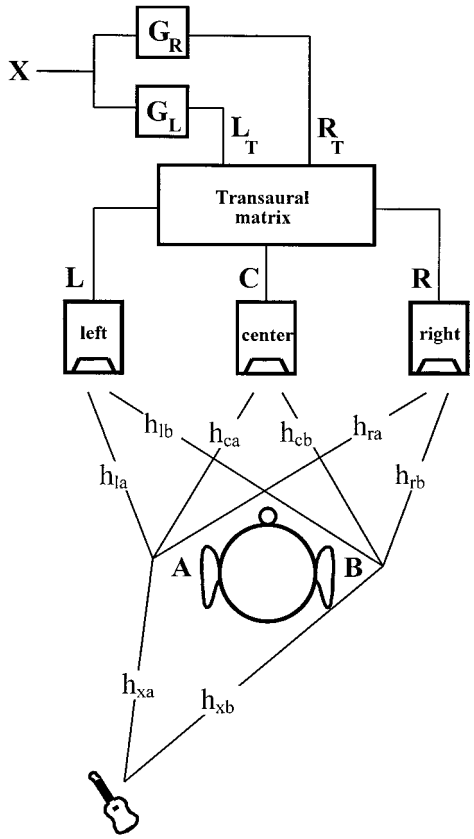



Fig. 7. Transaural processing using symmetrical three-loudspeaker array.

### 3.1 Three-Loudspeaker Transaural Processing

To illustrate how to accommodate more than two loudspeakers in an array while retaining the requirements for precise HRTF formulation, consider a three-loudspeaker array as illustrated in Fig. 7. In this system a mono source signal $X$ is filtered by the positional transfer functions $G_R$ and $G_L$ to form $L_T$ and $R_T$, which in turn form inputs to the matrix $[a]$. By way of example a Trifield[8] matrix (after Gerzon [15]) is selected, which is defined as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} = \begin{bmatrix} 0.8850 & -0.1150 \\ 0.4511 & 0.4511 \\ -0.1150 & 0.8850 \end{bmatrix}. \tag{10}$$

By applying the coefficients defined in matrix $[a]$, the three loudspeaker signals $L$, $C$, and $R$ (left, center, right) can be derived. However, to reproduce optimum localization, the system requires that the ear signals produced by the three loudspeakers match the ear signals that would be produced by the real source.

#### *3.1.1 Analysis*

The positional transfer function matrix $[G]$ converts the mono signal $X$ to $L_T$ and $R_T$ as

$$\begin{bmatrix} L_T \\ R_T \end{bmatrix} = \begin{bmatrix} G_R \\ G_L \end{bmatrix} X \ . \tag{11}$$

Using matrix $[a]$, the loudspeaker feeds $L$, $C$, and $R$ are then derived from $[G]X$,

$$\begin{bmatrix} L \\ C \\ R \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} G_L \\ G_R \end{bmatrix} X \ . \tag{12}$$

However, recalling the HRTFs as defined in Fig. 7, where $h_{xa}$ and $h_{xb}$ are the HRTF coordinates of source image $X$, then

$$\begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} X = \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} L \\ C \\ R \end{bmatrix} \tag{13}$$

where, substituting for $L$, $C$, and $R$,

$$\begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} = \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} G_L \\ G_R \end{bmatrix}. \tag{14}$$

The positional transfer functions $G_R$ and $G_L$ then follow by matrix inversion,

$$\begin{bmatrix} G_L \\ G_R \end{bmatrix} = \left\{ \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \right\}^{-1} \begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} \tag{15}$$

[8] Registered trademark describing a two-channel to three-loudspeaker mapping proposed by Gerzon [15].

from which the loudspeaker feeds $L$, $C$, and $R$ are calculated,

$$\begin{bmatrix} L \\ C \\ R \end{bmatrix} = \left\{ \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} h_{\mathrm{xa}} \\ h_{\mathrm{xb}} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_{\mathrm{la}} & h_{\mathrm{ca}} & h_{\mathrm{ra}} \\ h_{\mathrm{lb}} & h_{\mathrm{cb}} & h_{\mathrm{rb}} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \right\}^{-1} X \ . \tag{16}$$

In practice $L$, $C$, and $R$ are calculated directly using matrix inversion. However, because the transfer functions can have several thousand elements, to avoid large-dimension matrices the solution can be decomposed as follows. Define

$$\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} h_{\mathrm{la}} & h_{\mathrm{ca}} & h_{\mathrm{ra}} \\ h_{\mathrm{lb}} & h_{\mathrm{cb}} & h_{\mathrm{rb}} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \tag{17}$$

giving

$$\begin{bmatrix} h_{\mathrm{xa}} \\ h_{\mathrm{xb}} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} G_{\mathrm{L}} \\ G_{\mathrm{R}} \end{bmatrix} \tag{18}$$

where, using matrix inversion, the positional transfer functions are

$$G_{\mathrm{R}} = \frac{h_{\mathrm{xa}} h_{21} - h_{\mathrm{xr}} h_{11}}{h_{12} h_{21} - h_{11} h_{22}} \tag{19a}$$

$$G_{\mathrm{L}} = \frac{h_{\mathrm{xa}} h_{12} - h_{\mathrm{xb}} h_{22}}{h_{12} h_{21} - h_{11} h_{22}} \tag{19b}$$

which enable $L$, $C$, and $R$ to be calculated. Fig. 8 shows example transfer functions linking the system input to the three loudspeaker inputs ($L$, $C$, and $R$) located at −45°, 0°, and 45°, with a source location at 30°. Simulations confirm that the correct ear signals are produced as shown in Fig. 9, while Figs. 10 and 11 present the magnitudes of the positional transfer functions $G_{\mathrm{L}}$ and $G_{\mathrm{R}}$ and their differential phase response, respectively.

### 3.2 Three-Loudspeaker Matrix with Band-Limited Center Channel

This section extends multiloudspeaker transaural processing by considering a case where the center channel is band-limited by a low-pass filter with a transfer function $\lambda(f)$. For example, $\lambda(f)$ could constrain the center channel to operate only in the band where the ear and brain employ interaural time differences for localization. Alternatively the center channel may be used mainly for low-frequency reproduction.

The inclusion of $\lambda(f)$ yields effective HRTF center channel coordinates $\{h_{\mathrm{ca.}}{*}\lambda(f),\ h_{\mathrm{cb.}}{*}\lambda(f)\}$, where .* implies element-by-element vector multiplication. Hence, from the equations derived in Section 3.1, $L$, $C$, and $R$ follow,

To illustrate this system with band-limited center channel, Fig. 12 shows again the system input to the loudspeaker transfer functions for the three loudspeakers located at −45°, 0°, and 45°, with a phantom source location at 30°. The low-pass filter $\lambda(f)$ in the center channel has a cutoff frequency of 100 Hz with an asymptotic attenuation slope of 40 dB per octave. The ear signals are formed correctly and are identical to those presented in Fig. 9.

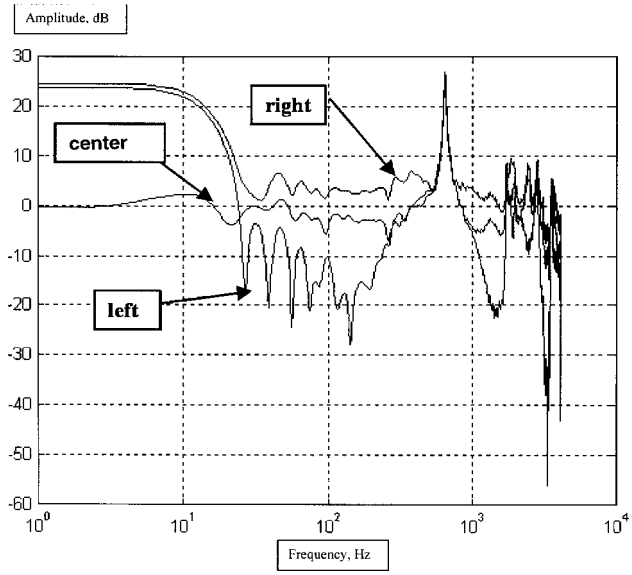



Fig. 8. Loudspeaker feed transfer functions for Trifield matrix linking input to three loudspeakers located at −45°, 0°, and 45°, image at 30°.

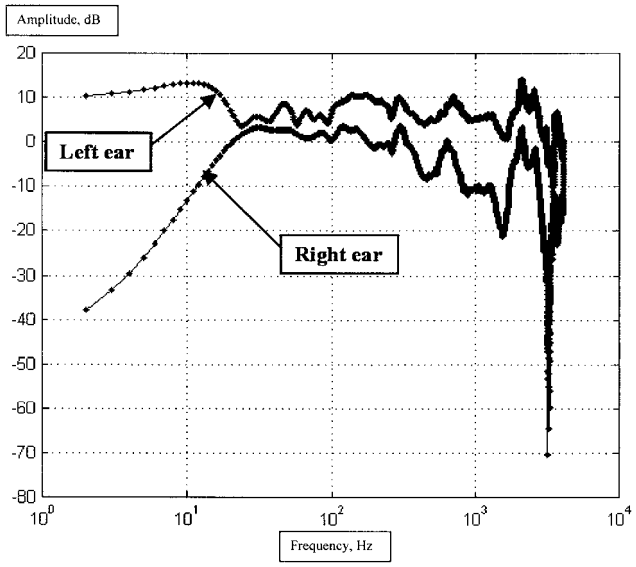



Fig. 9. Input-to-ear transfer functions for three-loudspeaker system with Trifield matrix, corresponding to functions shown in Fig. 8.

$$\begin{bmatrix} L \\ C \\ R \end{bmatrix} = \left[ \left\{ \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} h_{\mathrm{xa}} \\ h_{\mathrm{xb}} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_{\mathrm{la}} & h_{\mathrm{ca.}} * \lambda(f) & h_{\mathrm{ra}} \\ h_{\mathrm{lb}} & h_{\mathrm{cb.}} * \lambda(f) & h_{\mathrm{rb}} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \right\}^{-1} \right] \begin{bmatrix} 1 \\ \lambda(f) \\ 1 \end{bmatrix} X \ . \tag{20}$$

An attraction of this configuration is that the center channel can have a limited bandwidth while offering improvements in bass quality both in terms of power handling and by improving modal dispersion in the listening room. Alternatively, if a loss of spatial resolution at low frequency is permitted, then the center channel could function as a subwoofer with an upper response that extends only into the lower midband frequency range. The left- and right-hand loudspeakers would extend to high frequencies, although with restricted low-frequency performance.

### 3.3 *n*-Loudspeaker Array with Two-Channel Transaural Processing

The method of using more than two loudspeakers can be generalized to *n* loudspeakers while retaining only two information channels, where for example the loudspeakers surround the listener in a symmetrical array. The left- and right-hand loudspeakers in the array are fed by one of two information signals, and each loudspeaker has individual weighting defined by a coefficient matrix [*a*]

$$\begin{bmatrix} \mathrm{LS}(1) \\ \mathrm{LS}(2) \\ \vdots \\ \mathrm{LS}(n) \end{bmatrix} = \left\{ \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \begin{bmatrix} h_{\mathrm{xa}} \\ h_{\mathrm{xb}} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_{\mathrm{a}}(1) & h_{\mathrm{a}}(2) & \ldots & h_{\mathrm{a}}(n) \\ h_{\mathrm{b}}(1) & h_{\mathrm{b}}(2) & \ldots & h_{\mathrm{b}}(n) \end{bmatrix} \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \right\}^{-1} X \, . \tag{21}$$

Although this matrix equation cannot be solved in general as there are too many independent variables, solutions can be achieved when the matrix [*a*] is specified. For general two-channel stereophonic reproduction this system offers little advantage. However, in a telepresence and teleconference environment the coefficient matrix [*a*] may be transmitted for a given talker alongside the two information channels. A transaural reproduction system can then be conceived, where the coefficients are updated dynamically to enhance directional coding. This becomes particularly attractive where there are a number of talkers, as the coefficient matrix could be adjusted dynamically to enhance localization.

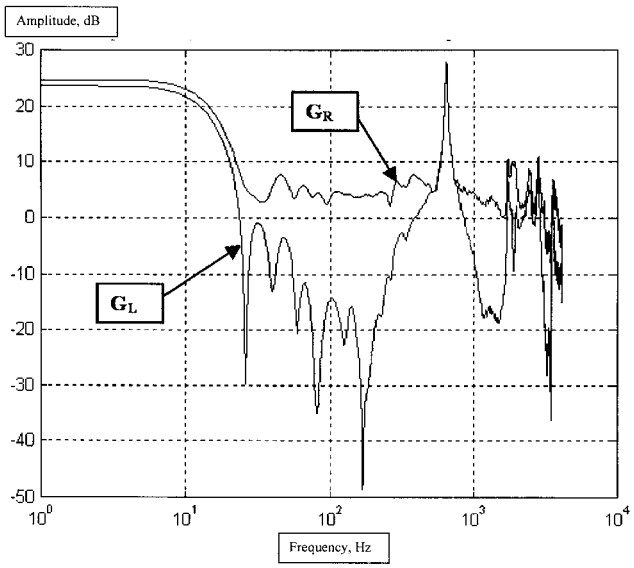



Fig. 10. Positional transfer functions $G_L$ and $G_R$, corresponding functions shown in Fig. 8.

## 4 MULTICHANNEL PARADIGM EXPLOITING PAIRWISE TRANSAURAL STEREO

This section reviews a spatial audio paradigm that links multichannel audio and transaural processing. The technique augments the directional clues inherent in multichannel stereo reproduction by embedding HRTF data such that the ear signals are matched more accurately to those produced by the source image. By default, such processing includes both frequency (interaural amplitude response) and time (interaural time response) information and therefore forms an elegant method of virtual image manipulation. Also, because HRTFs vary with both angular position and distance, sound sources can be synthesized and manipulated in three-dimensional space, together with reflections spatialized using their HRTF coordinates, which further enhances this process. In a multichannel system this processing is performed during source coding and is therefore compatible with all DVD formats.

The proposal operates at six principal levels:

- Selecting a pair of loudspeakers whose subtended angle includes the position of the phantom image helps reinforce the sound direction and matches conventional mixing practice for localization in multichannel systems.
- Encoded amplitude differences in signals above about 2 kHz support localization using interaural amplitude differences.
- Encoded time differences in signals below about 2 kHz support localization using interaural time differences where an extended bass performance is desirable.
- The addition of transaural processing based on HRTF data enables the construction of ear signals that match the original event and aids localization.
- Closer spacing of loudspeakers in a multiloudspeaker array reduces sensitivity to the precise form of HRTF characteristics, thus making an averaged HRTF set more applicable to a wide range of listeners.
- The effect of moderate head motion, which is a desirable attribute for improving localization, is supported. For relatively small loudspeaker subtended angles the error in ear-signal reconstruction is reduced when the head is moved by a small angle such that the vector component reenforces localization.

As an example, consider a circular array of $n$ loudspeakers, as shown in Fig. 13. In this system pairwise coding (PWC) selects the two closest loudspeakers such that an image $X$ falls within the subtended angle at the listening position. Two-channel HRTF synthesis is then used to form the optimum ear signals. For example, if loudspeakers $r$ and $r + 1$ are selected from the $n$-array of loudspeakers, then loudspeaker feeds LS($r$) and LS($r + 1$) are computed,

$$\begin{bmatrix} \mathrm{LS}(r) \\ \mathrm{LS}(r+1) \end{bmatrix} = \begin{bmatrix} h_\mathrm{a}(r) & h_\mathrm{a}(r+1) \\ h_\mathrm{b}(r) & h_\mathrm{b}(r+1) \end{bmatrix}^{-1} \begin{bmatrix} h_\mathrm{xa} \\ h_\mathrm{xb} \end{bmatrix} X = \begin{bmatrix} G_r \\ G_{(r+1)} \end{bmatrix} X \; . \tag{22}$$

Matrix [$G$] defines a set of positional transfer functions, which are effectively filters located between source and loudspeaker feed, where the HRTF notation was defined in Section 1 with reference to Fig. 1.

Solving for the positional transfer function matrix [$G$],

$$G_r = \frac{h_\mathrm{xa} h_\mathrm{b}(r) - h_\mathrm{xb} h_\mathrm{a}(r)}{h_\mathrm{a}(r+1) h_\mathrm{b}(r) - h_\mathrm{b}(r+1) h_\mathrm{a}(r)} \tag{23a}$$

$$G_{r+1} = \frac{h_\mathrm{xb} h_\mathrm{a}(r+1) - h_\mathrm{xa} h_\mathrm{b}(r+1)}{h_\mathrm{a}(r+1) h_\mathrm{b}(r) - h_\mathrm{b}(r+1) h_\mathrm{a}(r)} \; . \tag{23b}$$

This result can be generalized for an $n$-array of loudspeakers as

$$\begin{bmatrix} \mathrm{LS}(1) \\ \mathrm{LS}(2) \\ \vdots \\ \mathrm{LS}(n) \end{bmatrix} = \left\{ \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \begin{bmatrix} h_\mathrm{xa} \\ h_\mathrm{xb} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_\mathrm{a}(1) & h_\mathrm{a}(2) & \ldots & h_\mathrm{a}(n) \\ h_\mathrm{b}(1) & h_\mathrm{b}(2) & \ldots & h_\mathrm{b}(n) \end{bmatrix} \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \right\}^{-1} X = \begin{bmatrix} G_1 \\ G_2 \\ \vdots \\ G_n \end{bmatrix} X \; . \tag{24}$$

If a sound source falls between loudspeakers $r$ and $r + 1$, then $a_r = a_{r+1} = 1$; otherwise all remaining coefficients in matrix [$a$] are set to zero. Consequently for a sound source to circumnavigate the head, the HRTF coordinates $h_\mathrm{xa}$ and $h_\mathrm{xb}$ must change dynamically, whereas as the source moves between loudspeaker pairs, the coefficient matrix is switched to redirect the sound.

A four-channel, four-loudspeaker PWC scheme is shown in Fig. 14. The positional transfer functions {$G_1$, $G_2$, $G_3$, $G_4$} are calculated for each source location, which then filters the source signal $X$ to form the loudspeaker feeds. Because transaural processing is performed at the encoder, a simple replay system is supported. Consequently complete compatibility with conventional multichannel audio is retained.

## 5 INCREASED NUMBERS OF LOUDSPEAKERS

An increase in the number of loudspeakers can achieve more even sound distribution, distribute power handling, and possibly lower the sensitivity to room acoustics. This section considers methods by which the number of loudspeakers in an array can be increased.

In the basis system, where loudspeakers are linked directly to information channels located at coordinates compatible with the encoding HRTF coordinates, the loudspeakers are designated nodal loudspeakers. An $n$-

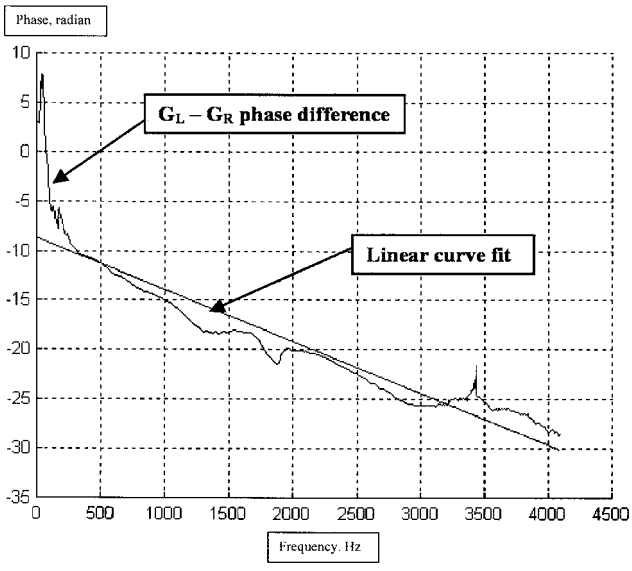



Fig. 11. Differential phase response between positional transfer functions $G_L$ and $G_R$ corresponding to functions shown in Fig 8.

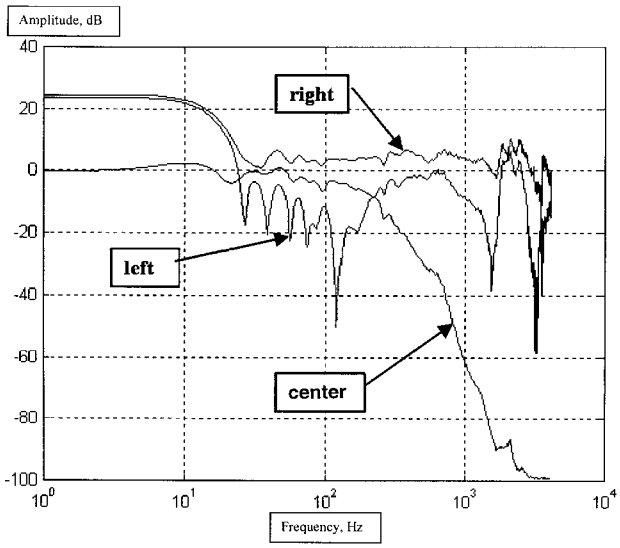



Fig. 12. Transfer functions linking input to three-loudspeaker feeds for Trifield matrix, but with center channel band-limited to 100 Hz, corresponding to functions shown in Fig. 8.

channel system therefore has *n* nodal loudspeakers, which constitute the basis array, with the corresponding drive signals, or primary signals, collectively forming the primary signal set. Loudspeakers in addition to the nodal loudspeakers are termed secondary loudspeakers.

### 5.1 Compensation for Inclusion of Secondary Loudspeakers

The objective is to derive additional signals within the decoder to drive secondary loudspeakers located between the nodal loudspeakers. However, the ear signals must be conserved and theoretically remain identical to the case where only the nodal loudspeakers are present. It is assumed here that all loudspeakers in the array have identical transfer functions and therefore do not affect the decoder process. If this is not the case, then they require individual correction. A sector of such an array is shown in Fig. 15. In this array nodal loudspeakers $r$ and $r + 1$ have the respective HRTF coordinates $[h_a(r), h_b(r)]$ and $[h_a(r + 1), h_b(r + 1)]$ at the listener, whereas the single secondary loudspeaker $p$ has the HRTF coordinates $[h_a(p), h_b(p)]$.

The synthesis of the drive signal for secondary loudspeaker $p$ employs two weighting functions $\lambda_{p1}$ and $\lambda_{p2}$ applied to the respective primary signals $LS_r$ and $LS_{r+1}$ such that $LS_p$, the drive signal to the secondary loudspeaker $p$, is

$$LS_p = \lambda_{p1}\, LS_r + \lambda_{p2}\, LS_{r+1}\,. \tag{25}$$

However, when the secondary loudspeaker enters the array, it is necessary to compensate the output from the two adjacent nodal loudspeakers in order that the ear signals remain unchanged. Ideally this needs to be achieved without knowledge of the encoding parameters. Otherwise the modified array cannot be used universally in a multichannel audio system. A scheme capable of meeting this objective is shown in Fig. 16, where the compensation transfer functions $\gamma_{p1}$ and $\gamma_{p2}$ filter the signal $LS_p$ to yield signals that are added to $LS_r$ and $LS_{r+1}$.

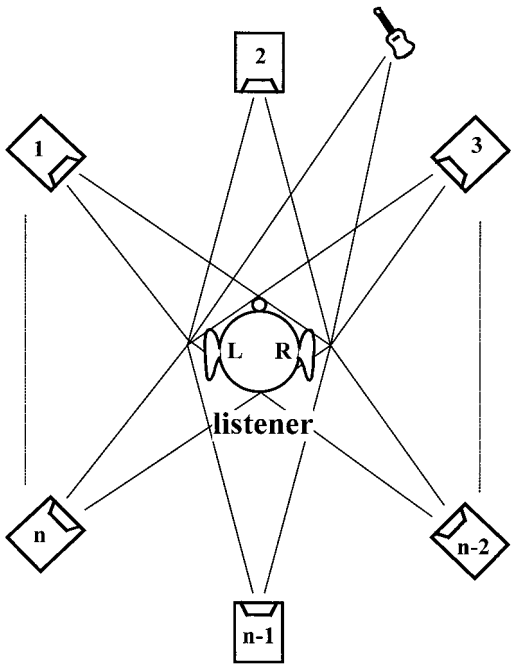



Fig. 13. *n*-loudspeaker array, suitable for transaural pairwise stereo.

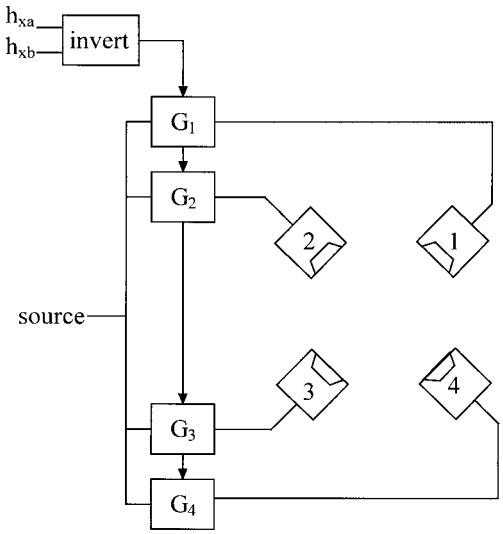



Fig. 14. Four-loudspeaker array with pairwise transaural synthesis.

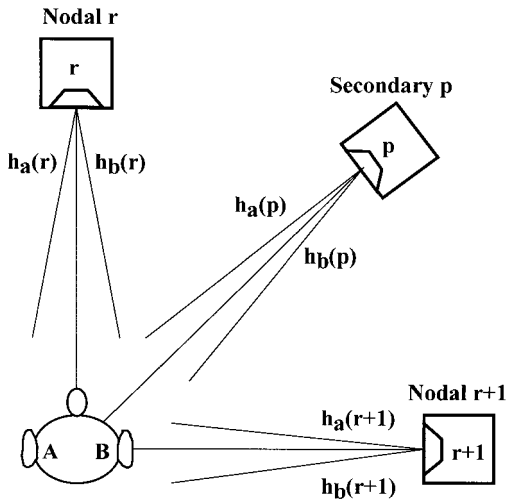



Fig. 15. Nodal loudspeakers with additional secondary loudspeaker.

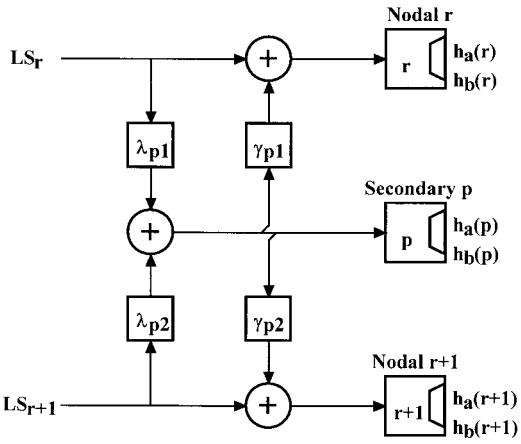



Fig. 16. Decoder processing to compensate for secondary loudspeaker *p*.

#### 5.1.1 Analysis

Consider initially the case when the secondary loudspeaker $p$ is absent from the array. The left- and right-ear signals $e_a$ and $e_b$ are expressed in terms of the HRTFs corresponding to the two adjacent nodal loudspeakers $r$ and $r + 1$,

$$e_a = \mathrm{LS}_r h_a(r) + \mathrm{LS}_{r+1} h_a(r+1) \quad (26a)$$

$$e_b = \mathrm{LS}_r h_b(r) + \mathrm{LS}_{r+1} h_b(r+1) . \quad (26b)$$

When the secondary loudspeaker $p$ is introduced into the array, the modified ear signals $e'_a$ and $e'_b$ become

$$\begin{aligned} e'_a = {} & \mathrm{LS}_r h_a(r) + \mathrm{LS}_{r+1} h_a(r+1) + \left\{\mathrm{LS}_r \lambda_{p1} + \mathrm{LS}_{r+1}\lambda_{p2}\right\}\gamma_{p1} h_a(r) \\ & + \left\{\mathrm{LS}_r \lambda_{p1} + \mathrm{LS}_{r+1}\lambda_{p2}\right\}\gamma_{p2} h_a(r+1) + \left\{\mathrm{LS}_r \lambda_{p1} + \mathrm{LS}_{r+1}\lambda_{p2}\right\} h_a(p) \end{aligned} \quad (27a)$$

$$\begin{aligned} e'_b = {} & \mathrm{LS}_r h_b(r) + \mathrm{LS}_{r+1} h_b(r+1) + \left\{\mathrm{LS}_r \lambda_{p1} + \mathrm{LS}_{r+1}\lambda_{p2}\right\}\gamma_{p1} h_b(r) \\ & + \left\{\mathrm{LS}_r \lambda_{p1} + \mathrm{LS}_{r+1}\lambda_{p2}\right\}\gamma_{p2} h_b(r+1) + \left\{\mathrm{LS}_r \lambda_{p1} + \mathrm{LS}_{r+1}\lambda_{p2}\right\} h_b(p) . \end{aligned} \quad (27b)$$

Forcing $e'_a = e_a$ and $e'_b = e_b$, then

$$\begin{bmatrix} h_a(r) & h_a(r+1) \\ h_b(r) & h_b(r+1) \end{bmatrix} \begin{bmatrix} \gamma_{p1} \\ \gamma_{p2} \end{bmatrix} = -\begin{bmatrix} h_a(p) \\ h_b(p) \end{bmatrix} \quad (28)$$

from which the correction functions $\gamma_{p1}$ and $\gamma_{p2}$ follow,

$$\gamma_{p1} = -\left[\frac{h_a(p)h_b(r+1) - h_b(p)h_a(r+1)}{h_a(r)h_b(r+1) - h_b(r)h_a(r+1)}\right] \quad (29a)$$

$$\gamma_{p2} = -\left[\frac{h_b(p)h_a(r) - h_a(p)h_b(r)}{h_a(r)h_b(r+1) - h_b(r)h_a(r+1)}\right] . \quad (29b)$$

These equations reveal that the correction functions $\gamma_{p1}$ and $\gamma_{p2}$ depend only on the HRTFs corresponding to the loudspeaker array. Consequently they are purely a function of the replay system, and its loudspeaker layout thus can be computed within the replay decoder as part of the installation procedure. However, the weighting functions $\gamma_{p1}$ and $\gamma_{p2}$ can be selected independently, provided the system is stable.

### 5.2 Relationship of System Parameters to Loudspeaker HRTFs

Section 5.1 presented an analysis of decoder parameters where precise HRTF measurement data are known for each loudspeaker position. However, there are some interesting observations and simplifications that can be made, which are considered in this section.

#### 5.2.1 HRTFs Derived Using Linear Interpolation

Consider a pair of nodal loudspeakers within an $n$-array PWC system where the proximity of loudspeakers $r$ and $r + 1$ is such that the image source HRTFs can be approximated by linear interpolation, such that

$$h_{xa} = \beta(m_r)\left\{m_r h_a(r) + (1 - m_r) h_a(r+1)\right\} \quad (30a)$$

$$h_{xb} = \beta(m_r)\left\{m_r h_b(r) + (1 - m_r) h_b(r+1)\right\} \quad (30b)$$

where $m_r$ is the panning variable with a range of 1 to 0, which corresponds to an image pan from loudspeaker $r$ to $r + 1$, and the function $\beta(m_r)$ is a moderator chosen to achieve a constant subjective loudness with variations in $m_r$. By substitution, the positional transfer functions $G_r$ and $G_{r+1}$ then simplify to

$$G_r = m_r \beta(m_r) \quad (31a)$$

$$G_{r+1} = (1 - m_r)\beta(m_r) . \quad (31b)$$

Consequently, for an image source that lies on a radial arc between the two nodal loudspeakers, simple amplitude panning yields the optimum panning algorithm. Effectively, HRTF coding information is derived directly from the loudspeaker locations and therefore is matched precisely to the listener. This assumes that intermediate HRTFs are derived by linear interpolation. It should also be noted that as the image source moves away from the radial arc containing the loudspeaker array, changes in HRTFs occur, making the positional transfer functions complex. Nevertheless, even when more exact HRTF data are available, there remains a strong desensitization to the exact form of the HRTFs when the positional transfer functions are calculated because of the relatively close proximity of loudspeakers in an $n$-array.

Consider next a secondary loudspeaker $p$ that is added to the array, again assuming a linear interpolation model. It is assumed that the secondary loudspeaker is located midway along the same radial arc as the nodal loudspeakers and that its HRTFs $h_a(p)$ and $h_b(p)$ with respect to the listener can be determined by linear interpolation. Thus for the midpoint location,

$$h_a(p) = 0.5h_a(r) + 0.5h_a(r+1) \quad (32a)$$

$$h_b(p) = 0.5h_b(r) + 0.5h_b(r+1) . \quad (32b)$$

From these data the compensation gamma functions $\gamma_{p1}$

and $\gamma_{p2}$ follow,

$$\gamma_{p1} = \gamma_{p2} = 0.5 \tag{33}$$

revealing once more a simple form for this special case. The modified positional transfer functions $GM_r$ $GM_{r+1}$ and $GM_p$ for the respective nodal and secondary loudspeakers $r$, $r + 1$, and $p$ are calculated as

$$GM_p = \lambda_{p1} G_r + \lambda_{p2} G_{r+1} \tag{34a}$$

$$GM_r = \left(1 + \gamma_{p1}\lambda_{p1}\right) G_r + \gamma_{p1}\lambda_{p2} G_{r+1} \tag{34b}$$

$$GM_{r+1} = \gamma_{p2}\lambda_{p1} G_r + G_{r+1}\left(1 + \gamma_{p2}\lambda_{p2}\right). \tag{34c}$$

Assuming symmetry, let $\lambda_{p1} = \lambda_{p2} = 0.5$ and consider the following examples:

1) $G_r = 1$ and $G_{r+1} = 0$, yielding $GM_r = 0.75$, $GM_{r+1} = -0.25$, and $GM_p = 0.5$.
2) $G_r = 0.5$ and $G_{r+1} = 0.5$, yielding $GM_r = 0.25$, $GM_{r+1} = 0.25$, and $GM_p = 0.5$.

For an image located coincident with the secondary loudspeaker there is a 6-dB level difference between secondary and nodal loudspeaker input signals. Also signal processing is simple and uses only real coefficients in the matrices.

### 5.2.2 Compensation and Incorporation of Exact HRTF Data

It is instructive to compare three other options for incorporating secondary loudspeaker and image HRTF coordinates. In each case correction functions $\lambda_{p1}$ and $\lambda_{p2}$ are calculated to match the selected HRTFs and the corresponding transfer functions for system input-to-loudspeaker inputs evaluated for loudspeakers $r$, $p$, and $r + 1$ located at 20°, 30°, and 40°, respectively, with an image at 30°. The four cases are as follows.

*Case 1*:

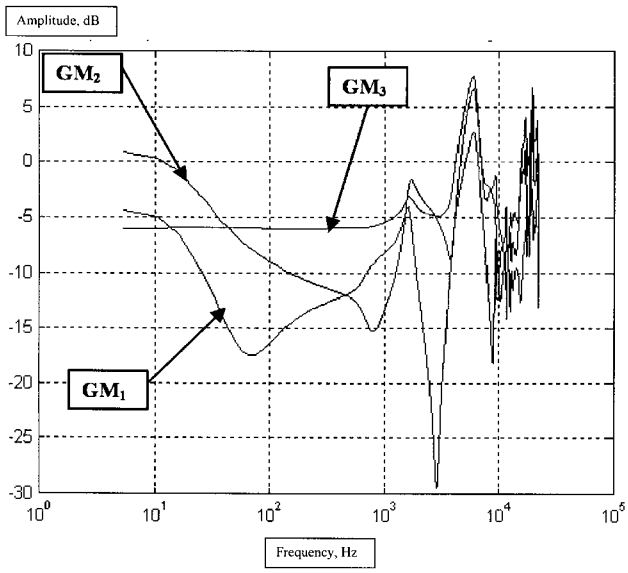
| | |
|---|---|
| Image | Measured HRTF data |
| Secondary loudspeaker | Measured HRTF data |
| Results | See Fig. 17(a) |

*Case 2*:

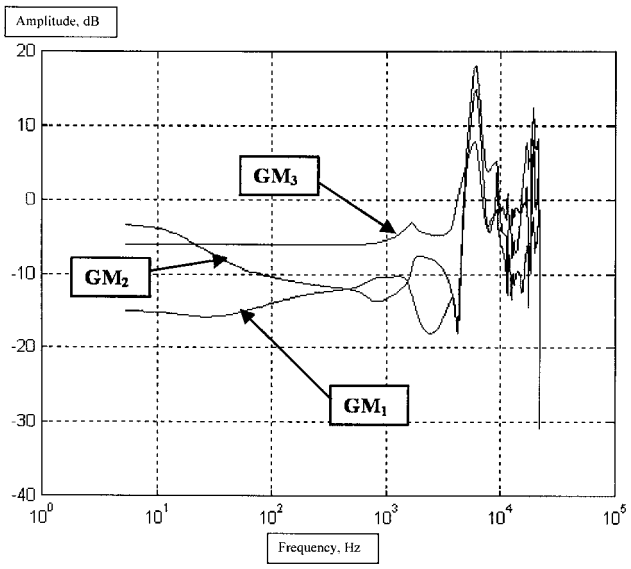
| | |
|---|---|
| Image | Measured HRTF data |
| Secondary loudspeaker | HRTF derived by linear interpolation |
| Results | See Fig. 17(b) |

*Case 3*:

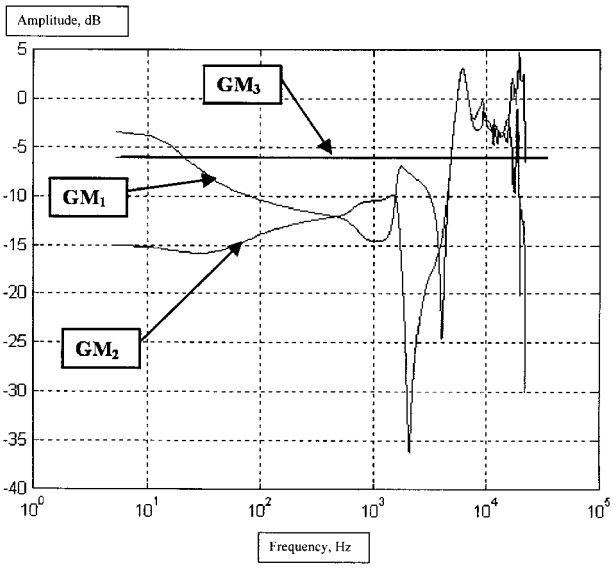
| | |
|---|---|
| Image | HRTF derived by linear interpolation |
| Secondary loudspeaker | Measured HRTF data |
| Results | See Fig. 17(c) |




(b)




(a)




(c)

Fig. 17. Transfer functions linking input to three-loudspeaker feeds at 20°, 30°, and 40°, image at 20°. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4.

*Case 4*:

| | |
|---|---|
| Image | HRTF derived by linear interpolation |
| Secondary loudspeaker | HRTF derived by linear interpolation |
| Results | See Fig. 17(d) and discussion in Section 5.2.1 |

Fig. 17(c) shows that when the HRTF coordinates for the image are derived by linear interpolation, the center channel response is constant with frequency, even though the secondary loudspeaker HRTF coordinates are derived by measurement. Also, when both sets of coordinates are derived by linear interpolation, all three responses are constant, as demonstrated in Section 5.2.1. However, for cases 1 and 2 all responses vary with frequency, although case 2 maintains a greater high-frequency content in the center channel response, which is desirable for an image located coincident with the secondary loudspeaker.

### *5.2.3 Compensation for Encoder–Decoder Loudspeaker Displacement Error*

The encoder assumes a nominal loudspeaker array location when determining the positional transfer functions. If the nodal loudspeakers are placed in the listening space in equivalent positions, then no additional processing is required at the decoder. However, in circumstances where the loudspeaker locations are displaced, positional compensation is required. It is important that positional compensation be independent of source coding and can be applied at the decoder without knowledge of the encoding algorithm. A pairwise positional correction scheme is shown in Fig. 18, where the compensation functions $GC_{11}(r)$, $GC_{12}(r)$, $GC_{11}(r+1)$, and $GC_{12}(r+2)$ are used to derive modified loudspeaker feeds. The form of this compensation is not unique, as correction signals can be applied to other loudspeakers in the array via appropriate filters. However, it is suggested that the loudspeaker selected to process the correction signal be the one closest to the loudspeaker that has been displaced. Hence by way of example, consider the scheme shown in Fig. 18. When the only active primary fed is $LS_r$, and equating the ear signals for both optimum and displaced loudspeaker locations, the positional compensation filters are as follows:

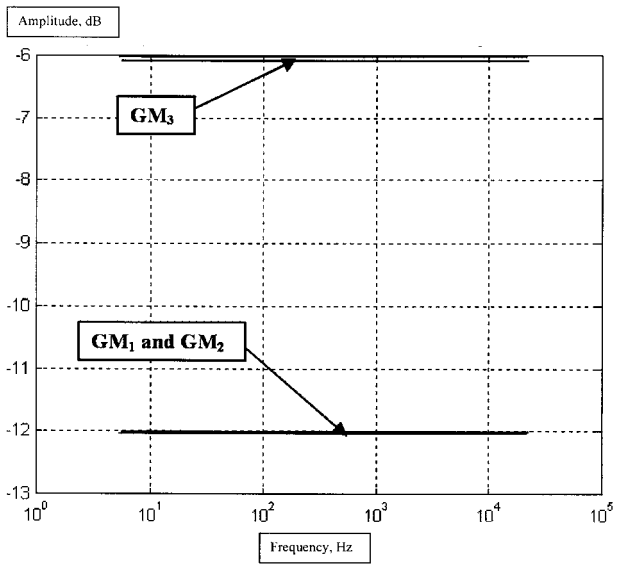
$$\begin{bmatrix} GC_{11}(r) \\ GC_{12}(r) \end{bmatrix} = \begin{bmatrix} h'_a(r) & h'_a(r+1) \\ h'_b(r) & h'_b(r+1) \end{bmatrix}^{-1} \begin{bmatrix} h_a(r) \\ h_b(r) \end{bmatrix}. \quad (35)$$

Similarly, when only $LS_{r+1}$ is active, then

$$\begin{bmatrix} GC_{11}(r+1) \\ GC_{12}(r+1) \end{bmatrix} = \begin{bmatrix} h'_a(r+1) & h'_a(r) \\ h'_b(r+1) & h'_b(r) \end{bmatrix}^{-1} \begin{bmatrix} h_a(r+1) \\ h_b(r+1) \end{bmatrix}. \quad (36)$$

### 5.3 Standardization of HRTF Grid and Nodal Loudspeaker Locations

The techniques described in the preceding require knowledge of the HRTF coordinates for each nodal loudspeaker. Establishing a standardized encoding grid where each grid point is assigned nominal HRTF coordinates and where nodal loudspeakers are assigned nominal grid locations can satisfy this requirement. All encoder and decoder users then universally know this information. An example grid proposal, as shown in Fig. 19, is based on a 60° subtended angle for nodal loudspeakers with two additional




(d)

Fig. 17. *Continued*

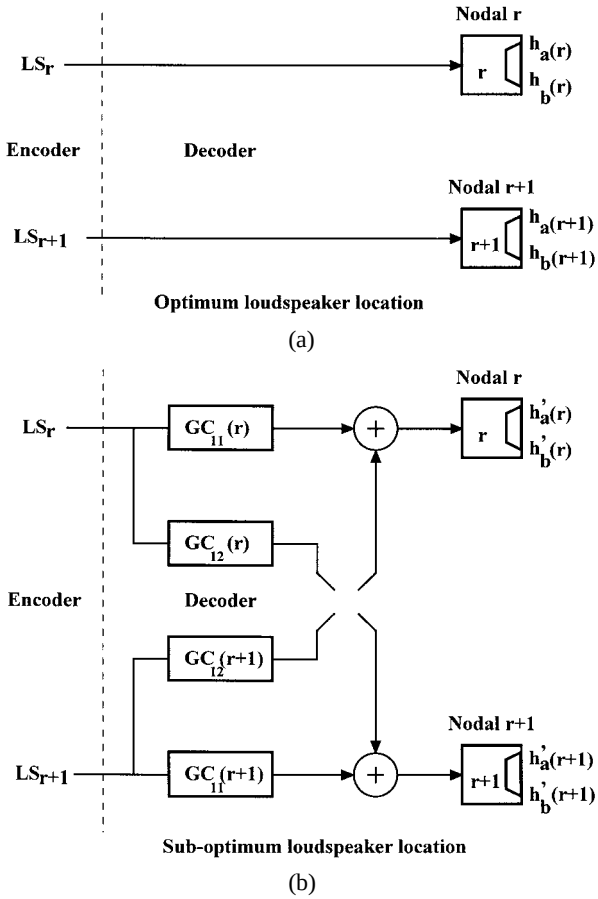



Fig. 18. Compensation process for nodal loudspeaker positional errors. (a) Optimum loudspeaker location. (b) Suboptimum loudspeaker location.

secondary loudspeakers. Three layers of nodes are suggested at radii of 1.5 m, 3 m, and 6 m. It is recognized that HRTFs are not unique, being listener specific, but when a multiloudspeaker array is formed using a set of HRTFs that are shared with image synthesis, errors are reduced.

HRTF coordinates that are noncoincident with the nodal points can be inferred by interpolation. For example, assume that an image $X$ is located at the cylindrical coordinates $\{r, \theta\}$, where the four nearest nodes are $\{r_1, \theta_1\}$, $\{r_1, \theta_2\}$, $\{r_2, \theta_1\}$, and $\{r_2, \theta_2\}$. The interpolated HRTFs $h_a(r, \theta)$ and $h_b(r, \theta)$ are then

$$h_a(r,\theta) = m_r\left[m_\theta h_a(r_1, \theta_1) + (1 - m_\theta) h_a(r_1, \theta_2)\right] + (1 - m_r)\left[m_\theta h_a(r_2, \theta_1) + (1 - m_\theta) h_a(r_2, \theta_2)\right] \quad (37a)$$

$$h_b(r,\theta) = m_r\left[m_\theta h_b(r_1, \theta_1) + (1 - m_\theta) h_b(r_1, \theta_2)\right] + (1 - m_r)\left[m_\theta h_b(r_2, \theta_1) + (1 - m_\theta) h_b(r_2, \theta_2)\right] \quad (37b)$$

where $m_\theta$ and $m_r$ are the angular and radial linear interpolation parameters defining the image $X$. For images that lie either within the inner radius or beyond the outer radius, angular interpolation is performed first, followed by an appropriate adjustment to the amplitude and time delays based on the radial distance from the head.

## 6 PERCEPTUALLY BASED CODING EXPLOITING EMBEDDED CODE IN PRIMARY SIGNALS TO ENHANCE SPATIAL RESOLUTION

The techniques described in Section 5 can be extended to a system with any number of nodal and secondary loudspeakers and thus can be matched to a wide variety of multichannel system configurations. However, inevitably there is a limit to spatial resolution arising from the use of matrixing only, which imposes crosstalk between nodal and secondary loudspeaker feeds. Some advantage may be gained by using nonlinear decoding with dynamic parameterization, although for high-resolution music reproduction linear decoding should be retained.

Because DVD-A is capable of six channels at 24 bit 96 kHz, some of the lower bits in the LPCM stream can be sacrificed [12] while still retaining an exemplary dynamic range by using standard methods of psychoacoustically motivated noise shaping and equalization [16]. The least significant bits in the LPCM streams together with a randomization function can then be used to encode additional audio channels using perceptual coders such as AC-3,[9] DTS,[10] or MPEG.[11] For example, 4 bit per sample per LPCM channel

[9] Proprietary perceptual coding developed by Dolby Laboratories.
[10] Digital Theatre Systems.

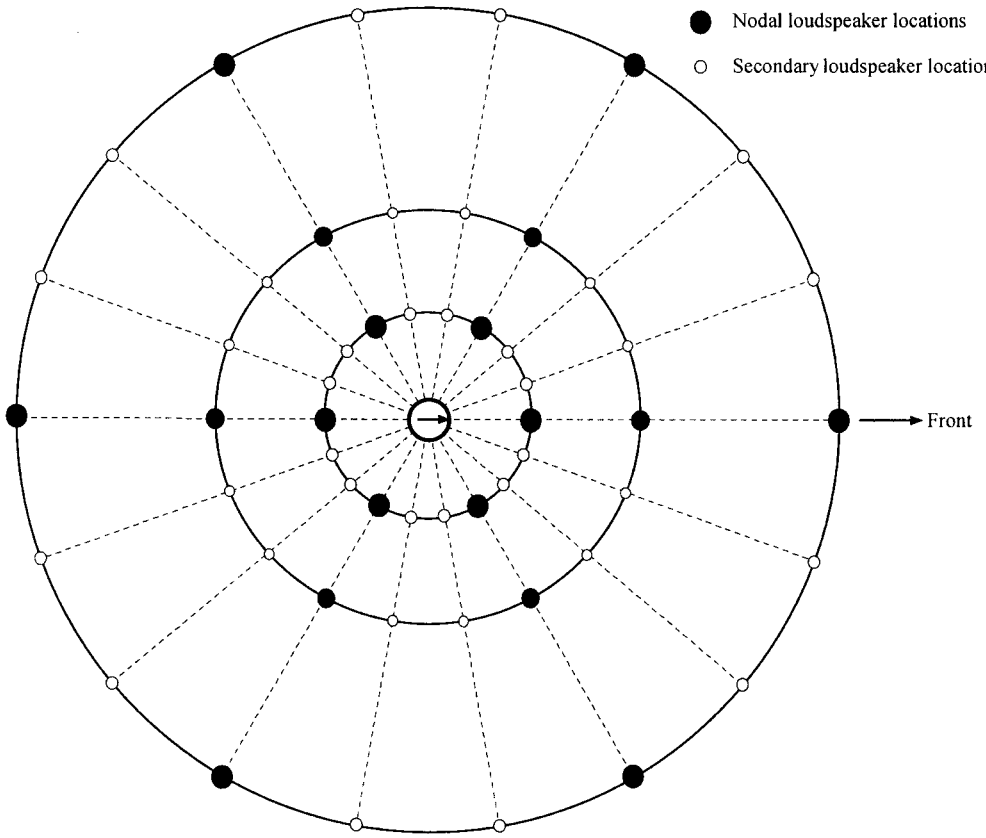



Fig. 19. Proposed 18-segment constellation map to define a standard set of HRTFs.

at 96 kHz yields a serial bit rate of 384 kbit/s.

The proposal retains the primary signals in high-resolution LPCM and uses the matrix methods described in Section 5 to estimate the secondary loudspeaker feeds. Spatially related difference signals are then calculated from the discrete secondary loudspeaker signals available at the encoder and the matrix-derived signals. Also, because of close spatial clustering of the additional channels there is a high degree of interchannel correlation with both the primary and the secondary loudspeaker signals, a factor that bodes well for accurate perceptual coding. Close clustering also implies that perceptual coding errors are not widely dispersed in space, yielding an improved masking performance. Given that DVD-A already supports six LPCM channels, it is suggested that an extra two encoded signals per primary signal is a realistic compromise, yielding a total of 18 channels, as proposed in the standardized HRTF constellation illustrated in Fig. 18. In the grand plan there would be $n$ perceptual coders in operation, one per nodal loudspeaker feed. In such a scheme further gains are possible by integrating dynamic bit allocation across all coders as well as using a perceptual model designed specifically for multichannel stereo encoding. In difficult encoding situations dynamic spatial blending can be used to reduce the difference signals prior to perceptual encryption. Fig. 20 shows the basic encoder architecture where the error signals $D_1$ and $D_2$ are indicated. In Fig. 21 a decoder is shown where identical estimates are made of the secondary loudspeaker input signals, but with the addition of the difference signals to yield discrete loudspeaker feeds. Of course, if the embedded perceptually coded difference signals are not used, estimates can still be made for the secondary loudspeakers as shown in Fig. 16. Alternatively, for a basic scheme an array of nodal loudspeakers only can be used.

[11] Perceptual based audio coding proposed by the Motion Picture Expert Group.

[12] Registered trademark of Company name, New Transducers plc, UK.

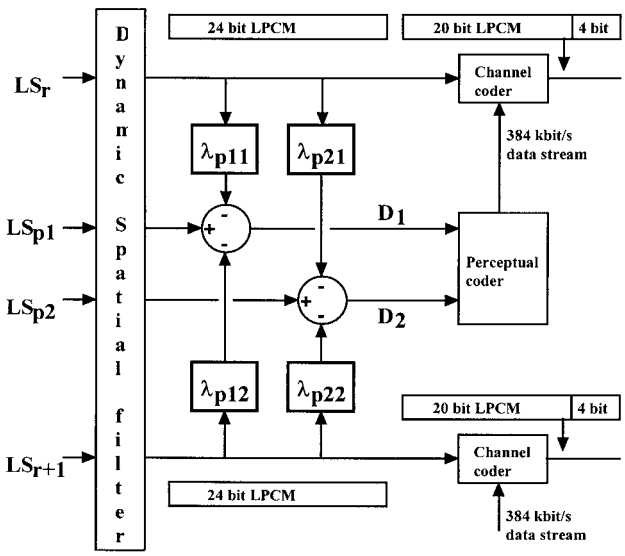



Fig. 20. High-level processor architecture for encoding discrete secondary loudspeaker feeds.

## 7 CONCLUSIONS

This paper has presented a method for multichannel audio that is fully compatible with DVD (DVD-A and SACD) multichannel formats, which have the capability of six high-resolution signals. The key to this technology is the exploitation of spatial coding using HRTF data to enhance the positional representation of sound sources. As such it forms a link between two-channel transaural techniques and conventional multichannel audio using many loudspeakers. This technique has already been demonstrated in telepresence and teleconferencing applications [9], [10] to be effective in representing spatial audio. However, the methods described here show how a particular loudspeaker array can be configured where issues of positional calibration were discussed for loudspeakers displaced from those locations assumed during coding.

The method is scalable and fully backward compatible. In a simple system there is no additional processing at the decoder where, for example, the outputs of a DVD player are routed directly to an array of loudspeakers. However, if additional loudspeakers are used, as might be envisaged with tiled walls of flat-panel NXT[12] loudspeakers (see, for example, Fig. 22), then formal methods exist, enabling the correct ear signals at the listening position to be maintained. Also for DVD-A, a method was suggested where perceptually coded information is embedded within the LPCM code to enable discrete loudspeaker signals to be derived. It was proposed that an upper limit of 18 channels should be accommodated, although full compatibility with systems down to the basic array is maintained.

An interesting observation for systems using a large number of closely spaced loudspeakers is that image positioning on the arc of the array can use simple linear amplitude panning applied between pairs of adjacent loud-

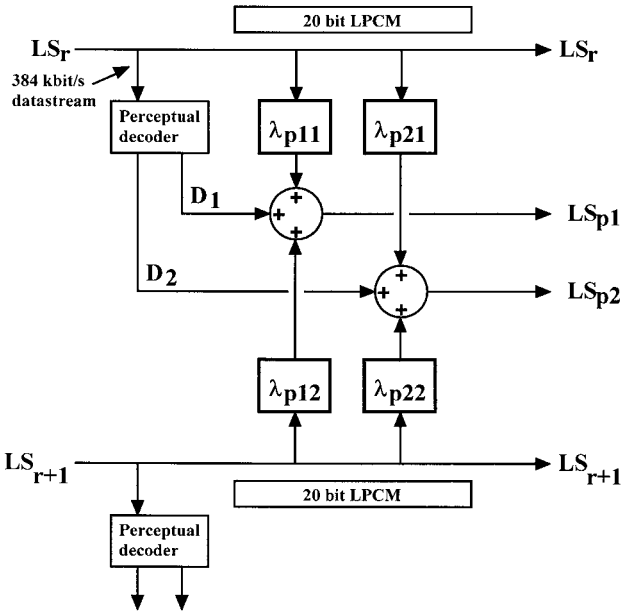



Fig. 21. High-level decoder architecture to derive discrete secondary loudspeaker feeds.

speakers. This approximation assumes that the image HRTF coordinates can be estimated to sufficient accuracy using linear interpolation between adjacent loudspeaker HRTF coordinates. This expedience effectively embeds HRTF data matched exactly to the listener simply because of the physical location of the loudspeakers. However, as the loudspeaker spacing increases, this approximation fails, requiring then the use of more accurate HRTF image coordinates together with transaural PWC, as described. This is particularly important where an image is located away from the arc of the loudspeaker array and where reflections are to be rendered to craft a more accurate virtual acoustic.

This work is also targeted at new communication formats for virtual reality, telepresence, and video conferencing [17], [18], where future research should investigate its application. Such schemes are not constrained by the normal paradigms of multichannel stereophonic reproduction, nor is compatibility necessarily sought. The approach is to form an optimum methodology for constructing phantom images and to consider coding paradigms appropriate for communication. For example, one possible communication format assigns a discrete channel to each phantom image. The channel then conveys the auditory signals together with the spatial coordinates updated at a rate compatible with motion tracking of the sound source. At the receiver, a processor carries a downloaded program with knowledge of the positional data and source acoustics from which the required reflections and reverberation are computed. These data would then be formatted to match the selected loudspeaker array. Such a scheme has great flexibility and can allow many mono sources to contribute to the final soundscape.

In conclusion, the techniques presented describe a means by which spatial resolution and image coding performance can transcend the six-channel limitation of the current DVD formats, yet without requiring additional storage capacity. Also, by basing signal processing on a perceptual model of hearing, it is revealed how sound images can be rendered and, in particular, how interaural amplitude differences and interaural time differences can be accommodated without seeking tradeoffs between time and amplitude clues. Essentially the work has presented a scalable and reverse compatible solution to multichannel audio that is particularly well matched to an LPCM format on DVD-A.

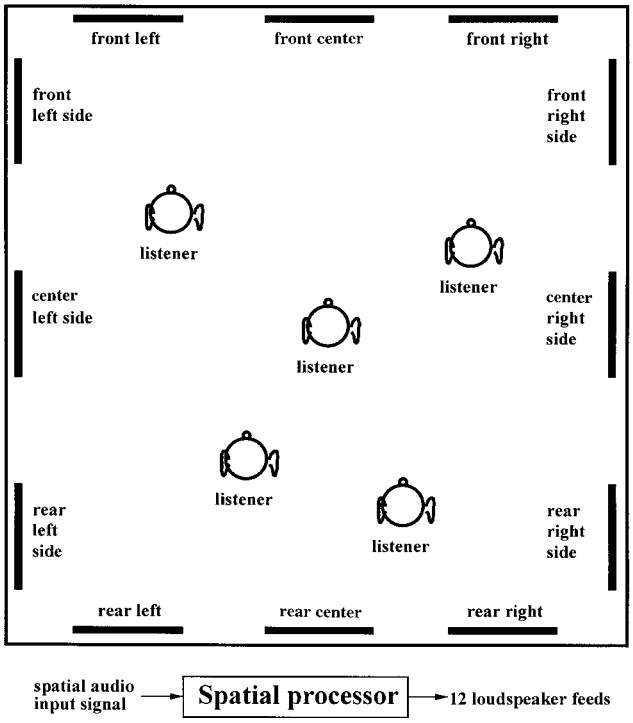



Fig. 22. Multichannel configuration using array of wall-mounted flat-panel diffuse loudspeakers.

## 8 REFERENCES

[1] *HFN/RR*, "Digital Frontiers," vol. 40, pp. 58–59, 106 (1995 Feb.).

[2] M. O. J. Hawksford, "High-Definition Digital Audio in 3-Dimensional Sound Reproduction," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 1016 (1997 Nov.), preprint 4560.

[3] A. J. Berkout, D. de Vries, and P. Vogel, "Acoustic Control by Wavefield Synthesis," *J. Acoust. Soc. Am.*, vol. 93, pp. 2764–2778 (1993).

[4] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* (AP Professional, 1994).

[5] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10 (1973 Jan./Feb.).

[6] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *J. Audio Eng. Soc.*, vol. 33, pp. 859–871 (1985 Nov.).

[7] M. A. Gerzon, "Hierarchical Transmission Systems for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, pp. 692–705 (1992 Sept.).

[8] K. C. K. Foo and M. O. J. Hawksford, "HRTF Sensitivity Analysis for Three-Dimensional Spatial Audio Using the Pairwise Loudspeaker Association Paradigm," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 1018 (1997 Nov.), preprint 4572.

[9] K. C. K. Foo, M. O. J. Hawksford, and M. P. Hollier, "Three-Dimensional Sound Localization with Multiple Loudspeakers Using a Pairwise Association Paradigm and Embedded HRTFs," presented at the 104th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 46, p. 572 (1998 June), preprint 4745.

[10] K. C. K. Foo, M. O. J. Hawksford, and M. P. Hollier, "Pairwise Loudspeaker Paradigms for Multichannel Audio in Home Theatre and Virtual Reality," presented at the 105th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 46, p. 1035 (1998 Nov.), preprint 4796.

[11] M. Gerzon, "Practical Periphony: The Reproduction of Full-Sphere Sound," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 364 (1980 May), preprint 1571.

[12] M. A. Gerzon and P. G. Craven, "A High-Rate

Buried Data Channel for Audio CD," *J. Audio Eng. Soc.*, vol. 43, pp. 3–22 (1995 Jan./Feb.).

[13] J. Blauert, *Spatial Hearing*, rev. ed. (MIT Press, Cambridge, MA, 1997).

[14] J. Bauck and D. H. Cooper, "Generalized Transaural Stereo and Applications," *J. Audio Eng. Soc.*, vol. 44, pp. 683–705 (1996 Sept.).

[15] M. A. Gerzon, "Optimum Reproduction Matrices for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, pp. 571–589 (1992 July/Aug.).

[16] J. R. Stuart and R. J. Wilson, "Dynamic Range Enhancement Using Noise-Shaped Dither Applied to Signals with and without Preemphasis," presented at the 96th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 400 (1994 May), preprint 3871.

[17] "Telepresence Theme," *BT Technol. J.*, vol. 15 (1997 Oct.).

[18] D. M. Burraston, M. P. Hollier, and M. O. J. Hawksford, "Limitations of Dynamically Controlling the Listening Position in a 3-D Ambisonic Environment," presented at the 102nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 413 (1997 May), preprint 4460.

## THE AUTHOR

Malcolm Hawksford received a B.Sc. degree with First Class Honors in 1968 and a Ph.D. degree in 1972, both from the University of Aston in Birmingham, UK. His Ph.D. research program was sponsored by a BBC Research Scholarship and investigated delta modulation and sigma–delta modulation (SDM, now known as bitstream coding) for color television and produced a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

Dr. Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at Essex University, where his research and teaching interests include audio engineering, electronic circuit design, and signal processing. His research encompasses both analog and digital systems with a strong emphasis on audio systems including loudspeaker technology. Since 1982, research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. A first in 1986 was for a prototype system to be demonstrated at the Canon Research Centre in Tokyo, work sponsored by a research contract from Canon. Much of this work has appeared in the *JAES*, together with a substantial number of contributions at AES conventions.

His research has also encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with a special emphasis on SDM. Other research has included the linearization of PWM encoders, diffuse loudspeaker technology, and three-dimensional spatial audio and telepresence including multichannel sound reproduction.

Dr. Hawksford is a recipient of the 1997/1998 AES Publications Award for his paper, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover." He is a chartered engineer as well as a fellow of the AES, IEE, and IOA. He is currently chair of the AES Technical Committee on High-Resolution Audio and is a founder member of the Acoustic Renaissance for Audio (ARA). He is also a technical consultant for NXT, UK and LFD Audio, UK.