A COMPARATIVE STUDY OF 3-D AUDIO ENCODING AND RENDERING TECHNIQUES

JEAN-MARC JOT[1], VERONIQUE LARCHER[2] AND JEAN-MARIE PERNAUX[2]

[1]Joint E-mu/Creative Technology Center, Scotts Valley, CA 95067, USA
jmjot@emu.com
[2]Ircam, 1 place Igor Stravinsky, 75004 Paris, France
larcher@ircam.fr pernaux@ircam.fr

This paper reviews and compares several techniques for multichannel sound panning over loudspeakers, as well as their application to computationally efficient 3-D sound spatialization over headphones. Ambisonic and pairwise amplitude (or intensity) panning approaches are reviewed within a general encoding/decoding framework. The performance of the different techniques is evaluated by use of objective criteria of auditory localization, and in terms of reconstruction of interaural time differences and HRTF spectral cues at the listener's ears. Improved approaches to efficient binaural synthesis are discussed, whereby interaural time differences are explicitly controlled. Specifically, the performance of the "binaural B format" encoding/decoding scheme is evaluated, and practical applications of this technique are reviewed.

INTRODUCTION

Over the past decades, a body of techniques have been invented and refined in order to record or synthesize, transmit, and reproduce the spatial information in natural or artificial sound scenes. Today, the sound engineer or designer, the computer musician or the designer of multimedia systems can consider a variety of multichannel transmission formats and end-user configurations for conveying a 3-D audio scene to a listener [1]. An example is a 5.1 DVD recording intended for reproduction over a standard 3/2-stereo loudspeaker layout. In addition to the practical choice of a multichannel transmission or storage and rendering format, various microphone recording techniques or electronic spatialization methods can be used to encode the directional information in the chosen multichannel format. This paper will focus on the latter problem, and on the associated techniques for reproducing (decoding) the desired directional information over headphones or a number of loudspeakers located at known positions surrounding the listening area.

The present study is restricted more specifically to spatial encoding and reproduction techniques involving a limited number of storage or playback channels, while allowing reproduction of a full 3-D sound scene. These techniques can be classified into three main approaches:

1. Ambisonics and related sound field reconstruction methods, where the intent is to control the acoustical variables of the sound field (pressure, velocity) at or around a reference measuring point in the listening area [2]–[7].
2. Discrete panning techniques, where knowledge of the intended apparent direction of the sound is used to selectively feed the closest loudspeakers in the reproduction system [8]–[10].
3. Head-related stereophony (binaural recording or binaural synthesis), where the intent is to control the acoustic pressure at the ears of the listener, via headphone or loudspeaker playback [11]–[14].

These three approaches yield different tradeoffs between several design criteria, including: (1) fidelity of the directional and timbral reproduction, (2) complexity in terms of number of channels or signal processing, (3) freedom of movement of the listener and size of the listening area. Therefore, the type of application typically determines the choice of one technique over another. For instance, head-related stereophony over 2 or 4 channels is a valid approach in a personal 3-D audio display system, but discrete panning or Ambisonic techniques, with 5 reproduction channels or more, are more effective for home theater applications or in larger auditoria.

In the first part of this paper, we define an encoding/decoding framework within which the relations betwen different multichannel audio techniques are examined, and their respective merits and tradeoffs can be exhibited. In each of the three above approaches, the theory and design of positional audio encoders and decoders are reviewed. More attention is given to the reproduction over headphones or horizontal-only loudspeaker layouts, as these are the most common practical configurations. The presentation of 3-D sound scenes over horizontal-

only layouts is discussed briefly for each technique.

In the second part of the paper, an objective comparison of Ambisonic and discrete panning techniques is presented. A particular reproduction layout is chosen (forming a hexagon with loudspeakers located at $\pm 30°$, $\pm 90°$ and $\pm 150°$ from the forward facing direction in the horizontal plane), in the ideal case of a single, centrally located listener. The performance of the different techniques is evaluated in terms of sound field reconstruction at the "sweet spot" (as in [15]) and in terms of pressure or intensity reconstruction at the ears of the listener (as in [16]). The latter approach is related to a binaural synthesis method based on multichannel panning over "virtual loudspeakers" [17]–[21]. The results presented indicate that the increased computational efficiency of this technique is compensated by a degradation in 3-D reproduction quality.

An improved approach to cost-efficient head-related stereophony is then considered, whereby the two ear signals are decomposed into two separate multichannel signals. One particular embodiment of this approach, initially introduced in [20], adopts an 8-channel "binaural B format" for directional encoding or multichannel recording. It is seen that this technique provides a more accurate reconstruction of the two ear signals when compared to the previous "coincident" multichannel encoding approaches. The design of binaural B format decoders for headphone or loudspeaker playback is reviewed.

1. RELATIONS BETWEEN 3D AUDIO ENCODING AND RENDERING TECHNIQUES

1.1. A general encoding/decoding framework

Figure 1 illustrates the definition adopted in this paper for the directional encoding and decoding operations:

- The **encoder** receives a monophonic signal and produces a multichannel signal conveying information on the direction of incidence of the sound. Mixing of different sources is assumed to be performed in the encoded domain.
- The **decoder** (which may be optional, depending on the technique used), receives the multichannel signal produced after the encoding/mixing stage and derives the audio signals which feed the individual playback channels (loudspeakers). One potential practical benefit of the encoding/decoding model is that awareness of the playback configuration (loudspeaker layout) can be limited to the decoder, while the encoder may use a universal multichannel format. This is a well-known property of the Ambisonic system [3].

We note that, in the audio industry, the terms "format", "encoder" and "decoder" are typically associated with particular loudspeaker layouts, and often imply a data reduction scheme (by matrixing to a smaller number of channels or by means of perceptual low bit rate coding techniques). In this paper, we are not concerned with such schemes as they do not actually realize the directional encoding function. Rather, encoding to these formats is typically applied to already composed multichannel content. This happens **after** the multichannel mixing or recording stage, and therefore leaves the choice of the actual directional encoding technique open (although the choice of the transmission format typically implies limitations).

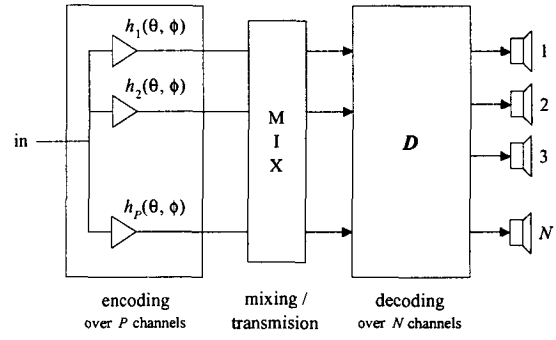

Figure 1: General 3D audio encoding/decoding framework. In the directional encoder (or "panpot"), the blocks h_j may be simple panning coefficients or, more generally, direction-dependent linear filters which can emulate the directivity characteristics of an array of transducers, and include delays in the case of non-coincident recording techniques.

The encoder can take one of the two following forms:

- **Acoustic encoder:** An array of microphones having known directivities, orientations, and positions relative to a reference ("center") point in space. In this case the "source signal" will be defined as the signal which would be captured by an ideal (omnidirectional) pressure microphone located at the center point. Exemples of acoustic directional encoders are traditional 2-channel stereophonic pickup microphones, the Sounfield microphone [22], or a dummy head microphone [11].
- **Electronic encoder** (or **"panpot"**): A bank of linear filters (denoted h_j in Figure 1) which produce the multichannel encoded signal from the monophonic source signal. Since an encoder is necessary for each individual sound source, there is generally a desire to limit the computational complexity of the encoder in any digital mixing system (as well as the number of encoding channels).

Any acoustic encoder can be emulated by an electronic encoder. However, it is not always possible to de-

sign a microphone array emulating an arbitrary electronic encoder (a typical example is a discrete panpot). When a microphone technique exists, an important advantage is that it is possible to make a spatial recording of an existing sound scene and artificially position additional sources within this scene by mixing, without deteriorating the consistence of the overall image. Discrete multichannel panning techniques typically do not offer this advantage.

Another particularity of discrete panning techniques is that the encoder directly produces loudspeaker feeds (no decoder), which means that knowledge of the loudspeaker layout must be assumed at the encoding/mixing stage, and that adaptation of a recording to a different playback configuration will in general not be possible without a substantial degradation in spatial imaging. It will be seen in section 1.4 that discrete panning lends itself to examination within the above encoding/decoding framework, although it involves a "local decoder".

1.2. Objective design criteria for 3D audio encoders and decoders

1.2.1. General assumptions

For a given acoustic 3-D audio encoder (microphone array), an optimal multichannel decoder can be expected to verify the following design criterion: if the same microphone array is placed at the reference listening point during playback (in free field), the recorded signal should be as close as possible to the originally encoded multichannel signal (the input of the decoder). Minimizing the difference according to some sensible error metric is a general approach to designing the 3-D audio decoder. We will refer to this as the "re-encoding principle" for designing 3-D audio decoders.

In order to compare the merits of different multichannel 3-D audio encoding/decoding techniques or optimize their design, it is useful to adopt general perceptually relevant objective criteria allowing to predict the fidelity of the reproduction over loudspeakers, irrespective of the encoding technique. In this paper, this evaluation will be made according to two approaches:

1. Compute the pressure signals at the ears of a centrally located listener (or dummy head). This approach was used in [16] to assess the fidelity of Ambisonic reproduction techniques, and is directly applicable to the optimization of loudspeaker decoders in the context of head-related stereophony [23] [24].
2. Compute the acoustic variables of the sound field (pressure and velocity) at the sweet spot. This is the classical basis for the design of Ambisonic decoders, where some assumptions are made as to the perceptual interpretation of certain acoustic indexes [25]. These indexes are reviewed below, and can be used to evaluate or optimize any multichannel panning technique [26], [15].

We assume that the sound field to be reproduced is a plane wave of arbitrary direction (we will not attempt here to evaluate the ability of the different techniques to reproduce a near-field sound source). This 'target direction' is defined by the unit-magnitude vector $\mathbf{s} = [x\ y\ z]^T$. The multichannel reproduction system consists of a set of N identical ideal loudspeakers. The position of loudspeaker i is defined by the unit-magnitude vector $\mathbf{s}_i = [x_i\ y_i\ z_i]^T$. As shown in Figure 2, the x, y and z axes point respectively forward, leftward, and upward. All loudspeakers are fed with the same source signal, weighted by the set of amplitude gains g_i.

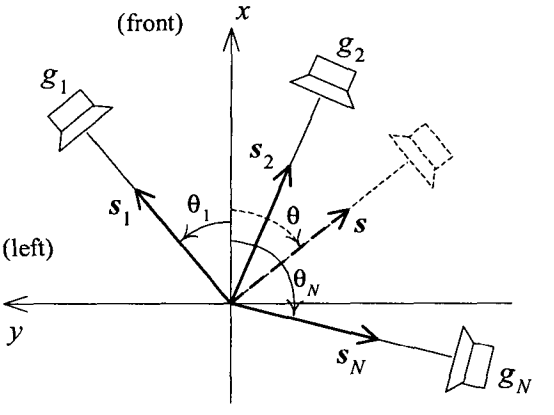

Figure 2: Conventions for coordinates and angles in the multichannel reproduction system (view from top: the z axis, not shown, points upward).

1.2.2. Low-frequency (velocity-based) localization criteria

At frequencies where the wavelength is substantially larger than the radius of the head, the knowledge of the sound pressure and velocity at the center point is sufficient to determine the acoustic pressure at the ears. This yields localization indexes which are considered to be perceptually valid for frequencies below roughly 700 Hz, where diffraction of sound by the head is negligible and the interaural time delay (ITD) is the principal auditory localization cue [27] [11] [25] [16].

The first assumption in the low-frequency localization theory is that the direction of the perceived sound event is perpendicular to the resulting wavefront [28]. This implies that the sum of the acoustic velocity vectors generated by the different loudspeakers should point in the target direction $\mathbf{s}$. The apparent direction of the

sound event for low frequencies is therefore given by:

$$r_v.\mathbf{s}_v = \frac{\mathbf{S.g}}{\sum g} \quad \textbf{velocity vector,} \tag{1}$$

where
- the matrix $\mathbf{S} = [\mathbf{s}_1\ \mathbf{s}_2 ... \mathbf{s}_N]^T$ characterizes the reproduction layout,
- the vector $\mathbf{g} = [g_1\ g_2 ... g_N]^T$ contains the corresponding amplitude gains,
- the scalar $\sum g$ denotes the sum of the amplitude gains g_i,
- the magnitude r_v of the velocity vector will be called the **velocity factor**.

For correct directional reproduction, we need to ensure $\mathbf{s}_v = \mathbf{s}$. The total gain $\sum g$ is introduced in (1) for normalization purposes (so that r_v is 1 if all louspeakers are located in the same direction).

The second localization criterion relates to the **apparent velocity** of the composite sound wave, which should be the same as for a natural plane wave [28] [25] [16]. This condition relates the composite pressure (proportionnal to $\sum g$) to the magnitude of the unnormalized composite velocity vector (given by $r_v . \sum g$). It implies that the velocity factor should verify:

$$r_v = 1 \tag{2}$$

It is straightforward to establish that r_v is always less than 1 if all the weights g_i are forced to be positive, unless the source signal is fed to only one loudspeaker (or several loudspeakers located in the same direction).

Third, to ensure that the resulting pressure at the center point is the same as when the source signal is simply panned to one of the loudspeakers, the sum of the amplitude weights should be scaled to unity:

$$\sum g = 1 \quad \text{(amplitude normalization)} \tag{3}$$

1.2.3. High-frequency (intensity-based) localization criteria

For mid-range and higher frequencies (above roughly 700 Hz), the knowledge of the composite presure and velocity at the center point no longer allows predicting the pressure signals at the listener's ears without additional knowledge of the composition of the sound field. To address this frequency range, Gerzon developped a localization model analogous to the above low-frequency model, although pressure is replaced by energy density and the velocity vector $\mathbf{s}_v$ is replaced by the energy (or intensity) vector $\mathbf{s}_e$ [25].

In this high-frequency model, the apparent direction of the sound event is determined by summing the intensity vectors generated by the different loudspeakers:

$$r_e.\mathbf{s}_e = \frac{\mathbf{S.q}}{\sum q} \quad \textbf{intensity vector,} \tag{4}$$

where
- the vector $\mathbf{q} = [g_1^2\ g_2^2 ... g_N^2]^T$ contains the, power gains
- the scalar $\sum q$ is the sum of the power gains,
- the magnitude r_e of the intensity vector will be called the **intensity factor** .

The natural relationship between the energy density and the magnitude of the intensity vector for a plane wave implies that the intensity factor should verify:

$$r_e = 1 \tag{5}$$

Since the elements of $\mathbf{q}$ are all non negative, r_e is always strictly less than 1 when the source signal is distributed to more than one loudspeaker. In order to optimize the quality of the reproduction, Gerzon recommends maximizing r_e (as close to 1 as possible). This is equivalent to maximizing the directional concentration of energy over the loudspeakers [16].

Finally, to ensure that the resulting energy density at the center point is the same as when the source signal is panned to one of the loudspeakers, the sum of the power gains should be scaled to unity:

$$\sum q = 1 \quad \text{(power normalization)} \tag{6}$$

1.3. Ambisonic encoding and decoding

1.3.1. First-order and second-order Ambisonic encoding equations

For a plane wave coming from direction $\mathbf{s} = [x\ y\ z]^T$ carrying the pressure signal S, the first-order Ambisonic encoder produces the four signals:

$$\begin{array}{lllll} W & = & S & & \\ X & = & S.x & = & S.\cos\theta.\cos\phi \\ Y & = & S.y & = & S.\sin\theta.\cos\phi \\ Z & = & S.z & = & S.\sin\phi \end{array} \tag{7}$$

which are respectively proportional to the pressure and the three components of the velocity at the center point. The angles θ and ϕ respectively denote the conventional azimuth and elevation angles.

The corresponding acoustic encoder is known as the Soundfield Microphone [22], which is equivalent to the coincident association of an omnidirectional capsule with three bidirectional capsules pointing towards perpendicular directions. The directional sensitivities of these four microphone capsules match the first-order spherical harmonics. The multichannel format $[W\ \sqrt{2}.X\ \sqrt{2}.Y\ \sqrt{2}.Z]$, where the velocity signals are amplified by 3 dB, is called the 'B format' [3]. In this paper, we simplify the notations by ignoring this 3 dB boost (which must be compensated for in the decoder).

For reproduction over a horizontal loudspeaker layout, an Ambisonic decoder cannot make use of the Z

component. To accommodate this situation, an Ambisonic panpot can be designed to incorporate the energy of Z in the W component [15]:

$$W' = S.\sqrt{1+\sin^2(\phi)}. \quad (8)$$

This modification of the encoding equations can be used to produce "fly-over" effects over a horizontal loudspeaker setup with Ambisonic techniques (but it does not resolve the problem of positionning sounds at negative elevations).

These encoding equations can also be extended to include the second-order harmonic components [29] [16]. Only two additional components are necessary for a horizontal-only second-order Ambisonic decoder:

$$\begin{aligned} U &= S.\cos(2\theta).\cos\phi, \quad (9)\\ V &= S.\sin(2\theta).\cos\phi. \end{aligned}$$

1.3.2. General Ambisonic decoder design

In this section, we derive a general solution for low-frequency Ambisonic decoders of any order, by applying the "re-encoding principle" exposed in section 1.2.1. This general method is outlined for first-order decoders, and then generalized to any order.

For a single plane wave with direction $\mathbf{s} = [x\,y\,z]^T$, the encoding equations (7) produce a B-format signal given by:

$$[W\ X\ Y\ Z]^T = [1\ x\ y\ z]^T.S = \mathbf{b}.S \quad (10)$$

During playback via N loudspeakers respectively producing the pressure signals $g_i.S$ at the center point, the resulting B-format signal is:

$$[W\ X\ Y\ Z]^T = \sum_{i=1}^{N} \mathbf{b}_i.g_i.S = \mathbf{B}.\mathbf{g}.S \quad (11)$$

where the matrix $\mathbf{B} = [\mathbf{b}_1\ \mathbf{b}_2 ... \mathbf{b}_N]^T$ characterizes the reproduction layout and its column vectors are defined by $\mathbf{b}_i = [1\,x_i\,y_i\,z_i]^T$. Equating (10) and (11), we obtain, in matrix form, the set of equations to solve for the amplitude weights g_i:

$$\mathbf{B}.\mathbf{g} = \mathbf{b} \quad (12)$$

The gains g_i are obtained by applying a decoding matrix $\mathbf{D}$ to the vector $\mathbf{b}$ (equivalently, the loudspeaker feeds are produced by applying $\mathbf{D}$ to the B-format signal):

$$\mathbf{g} = \mathbf{D}.\mathbf{b}, \quad (13)$$

where $\mathbf{D}$ verifies $\mathbf{D}.\mathbf{B} = \mathbf{Id}$, the identity matrix. A general solution for $\mathbf{D}$ is the pseudo-inverse of the matrix $\mathbf{B}$:

$$\mathbf{D} = \mathbf{B}^T.(\mathbf{B}.\mathbf{B}^T)^{-1}, \quad (14)$$

which has the favorable property of minimizing the L_2 norm of $\mathbf{g}$, that is $\sum q$ (and, consequently, the total power radiated by the loudspeakers).

We note that equation (12) can also be derived by replacing equations (2) and (3) in the velocity vector condition (1), and imposing amplitude normalization (3) as an additional condition. Furthermore, the vectors $\mathbf{b}$ and $\mathbf{b}_i$ in equations (10) and (11) can be extended to include the second-order components U and V, defined in equation (9), and any additional harmonic components of second or higher order. We conclude that equation (12) extends to the design of 2-D or 3-D Ambisonic decoders of any order, and that the decoder solutions automaticly satisfy the low-frequency localization criteria reviewed in section 1.2.2.

Moorer [6] obtained a set of equations equivalent to (12) via Fourier series decomposition of the directivity function of the plane wave –a method related to the spatial sampling approach developped in [2]. The matrix equation (12) is also obtained by truncating the harmonic (Fourier-Bessel) series expansion of a plane wave in the vicinity of the center point to a given order [29] [5] [16] [7]. It is worth noting that these recent studies tend to confirm the approximate value of 700 Hz as the upper frequency limit for correct pressure reconstruction at the ears with first-order Ambisonics. They also tend to confirm the intuitive conjecture that, for a harmonic decomposition truncated at order M, this upper frequency limit is multiplied by M (or, for a fixed frequency range, the size of the listening area is multiplied by M).

For higher frequencies, the solutions of (12) are in general not optimal with regards to the localization criteria (4) to (6) of section 1.2.3. Furthermore, even at low frequencies, it may be desirable to impose power normalization (6) rather than amplitude normalization. Unfortunately, these criteria lead to non-linear equations in the unknowns g_i for which no general closed-form solution is known. Ambisonic decoders for high frequencies must typically be optimized by numerical methods, except in the case of "regular" loudspeaker layouts.

1.3.3. Ambisonic decoders for regular polygons

A trivial solution for the decoder $\mathbf{D}$ in (14) arises when the matrix $\mathbf{B}.\mathbf{B}^T$ is diagonal due to the geometry of the loudspeaker Layout [16]. Gerzon [25] discusses several particular cases of first-order decoders for low and high frequencies, and Daniel [16] extends these solutions to the second order. In this section, we review the particular case of the regular horizontal polygon. For this configuration, denoting θ_i the azimuth of loudspeaker i and S_i the corresponding signal, the decoding equations become [25] [30]:

$$S_i = g_i\,S = 0.5\,[k_0 W + k_1 X\cos\theta_i + k_1 Y\sin\theta_i] \quad (15)$$

Gerzon established that this class of Ambisonic decoders satisfies both equations (1) and (4) irrespective of k_0 and k_1, which implies that the perceived azimuth is correct at both low and high frequencies [25] [16]. This property determines the "conventional" design of Ambisonic decoders (Figure 3), where the decoding matrix is preceded by a bank of phase-matched shelving filters [3]. The frequency-dependent coefficients k_0 and k_1 can be adjusted in order to satisfy several additional criteria, according to the frequency range and the listening conditions:

- ensuring $r_v = 1$ or maximizing r_e (eq. (2) or (5)),
- amplitude or power normalization (eq. (3) or (6)),
- the intended size of the listening area (which reduces the upper frequency limit for validity of the low-frequency localization criteria),
- ensuring that all weights g_i be non-negative (maintaining in-phase loudspeaker feeds is desirable to avoid noticeable artifacts in large auditoria [31]).

These different requirements can be combined into four main types of decoders, summarized in Table 1 (see also [30] and [16] for 3-D and second-order decoders).

size	freq.	optim.	norm.	k_0^2	k_1^2	k_1/k_0
small	low	$r_v = 1$	ampl.	$\frac{4}{N^2}$	$\frac{16}{N^2}$	2
small	low	$r_v = 1$	power	$\frac{4}{3N}$	$\frac{16}{3N}$	2
small	high	max. r_e	power	$\frac{2}{N}$	$\frac{4}{N}$	$\sqrt{2}$
large	all	in phase	power	$\frac{8}{3N}$	$\frac{8}{3N}$	1

Table 1: 1st-order Ambisonic decoder coefficients according to different criteria for regular horizontal polygon layout.

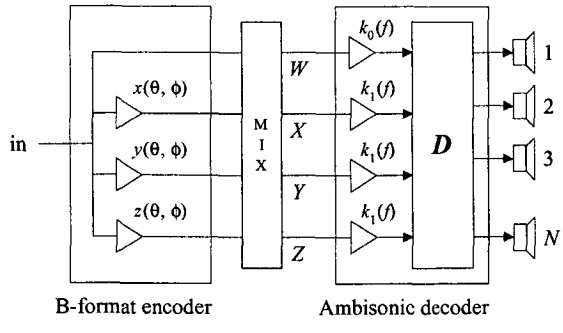

Figure 3: B-format encoder and "conventional" first-order Ambisonic decoder for regular loudspeaker layouts.

Replacing the encoding equations (7) in (15), we obtain the resulting gain panning law $g_i(\theta)$ for each loudspeaker. For a virtual sound source in the horizontal plane ($\phi = 0$):

$$g_i(\theta) = 0.5\,[k_0 + k_1 . \cos(\theta - \theta_i)] \qquad (16)$$

For $k_1/k_0 = 1$ (large-area decoder), each loudspeaker feed is equal to the signal captured by a cardoid microphone pointing in the direction θ_i. For any target azimuth θ, there is typically a contribution from all loudspeakers. For $k_1/k_0 > 1$, the pattern is more directive, but this results in diametrally opposite loudspeakers radiating out-of-phase signals. This effect is accentuated with low-frequency/small-area decoders, where the optimum directivity pattern becomes a hypercardioid [27] [2]. In the next section, we discuss discrete panning techniques which, unlike Ambisonic techniques, ensure non-negative panning gains and maximum directional concentration of the energy over the loudspeakers.

1.4. Discrete amplitude or intensity panning

1.4.1. Optimal panning laws for low frequencies

The application of the localization criteria (1) to (6) to the design of 2-D discrete multichannel panpots is discussed in [26]. The low-frequency criteria (1) to (3) lead to a set of four Linear equations (12), which can always be resolved by weighting four loudspeakers located in different directions (or only three loudspeakers in the 2-D case). Therefore, for any 3-D (resp. 2-D) loudspeaker reproduction setup, it is always possible to design a discrete multichannel panpot that is optimal with respect to the low-frequency localization criteria (1) to (3). As illustrated in Figure 4, this involves selecting the quadruplet (resp. triplet) of closest loudspeakers according to the target direction (θ, ϕ), and performing a "local" Ambisonic decoding operation over this set of loudspeakers. Selecting the closest loudspeakers actually has no effect on the fidelity of the directional reproduction at low frequencies according to criteria (1) to (3). However, it reduces the occurrence of out-of-phase panning gains and allows improving the localization at high frequencies.

The obvious drawback of this "discrete Ambisonic panning technique" is that it is not possible to use a common decoder for a multiplicity of virtual sound sources positionned at arbitrary target directions. The assumptions of the encoding/decoding model of Figure 1 are not satified, since, in general, mixing multiple sound sources can only be performed **after** decoding. One therefore loses two essential advantages of the Ambisonic encoding/decoding model [3] [4]: (a) the possibility of mixing a 3-D sound scene without a priori knowledge of the geometry of the reproduction setup, and (b) the possibility of applying global tranformations to the sound scene (ro-

tations, symmetries...) without remixing the individual source signals.

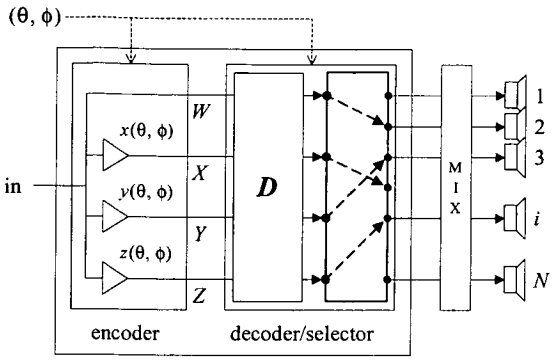

Figure 4: Discrete amplitude panning in the encoding/decoding framework.

The conventional approach to discrete panning in the horizontal plane consists of feeding pairs of adjacent loudspeakers [8] [9]. In this case, it is not possible to satisfy all the low frequency localization criteria (1) to (3). A natural compromise is to maintain the correct perceived direction (1) and impose amplitude or power normalization, (3) or (6), while dropping the requirement (2) on the velocity factor r_v. We establish here that this leads to the vector-base amplitude panning technique (VBAP) proposed by Pulkki [10].

When panning between two loudspeakers located at azimuths θ_1 and θ_2, with amplitude gains g_1 and g_2, equation (1) can be written:

$$\begin{pmatrix} \cos\theta_1 & \cos\theta_2 \\ \sin\theta_1 & \sin\theta_2 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \qquad (17)$$

with $g_i = a_i . r_v . \sum g$.

This is equivalent to the vector base equation posed by Pulkki. In [10], the amplitude panning gains g_i are obtained by resolving (17) to obtain the non-normalized gains a_i, and then applying power normalization:

$$g_1 = \frac{a_1}{\sqrt{a_1^2 + a_2^2}} \;\; ; \;\; g_2 = \frac{a_2}{\sqrt{a_1^2 + a_2^2}} \qquad (18)$$

Alternatively, if we opt for amplitude normalization (valid for very low frequencies and centrally located listener):

$$g_1 = \frac{a_1}{a_1 + a_2} \;\; ; \;\; g_2 = \frac{a_2}{a_1 + a_2} \qquad (19)$$

This amplitude-preserving VBAP law (19) is plotted on Figure 5 (dashed curve) for a pair of loudspeakers at azimuths $\pm 30°$. It is almost identical to the dotted curve, which represents linear panning vs. target azimuth angle, or "angle-based amplitude panning" (ABAP):

$$g_1 = \frac{\theta - \theta_2}{\theta_1 - \theta_2} \;\; ; \;\; g_2 = \frac{\theta - \theta_1}{\theta_2 - \theta_1} \qquad (20)$$

Figure 5 shows that, for a 60° separation between the loudspeakers, interpolating over angles (ABAP method) is almost equivalent to interpolating over direction vectors (VBAP method). The difference in azimuth angle is less than ± 1 degree. However, this error becomes larger for larger aperture angles, with only VBAP remaining accurate according to the low-frequency localization theory. Indeed, we have shown that VBAP is analogous to "local Ambisonic decoding" (Figure 4), although the "VBAP decoder" ignores the W component and must be followed by an amplitude (or power) normalization stage. This analogy also applies to the 3-D implementation of VBAP, where the local decoding operation is performed over three adjacent loudspeakers [10].

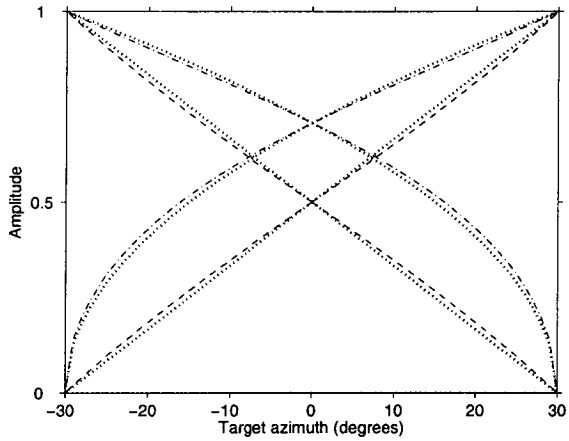

Figure 5: Pairwise amplitude and intensity panning laws for aperture 60°. Dashed: VBAP (dotted: ABAP). Dash-dotted: VBIP (dotted: ABIP).

1.4.2. Optimal panning law for high frequencies

We can apply the same approach as above in order to optimize discrete panning laws for frequencies higher than 700 Hz. In the case of panning between two loudspeakers located at azimuths θ_1 and θ_2, with amplitude gains g_1 and g_2, equation (4) can be written:

$$\begin{pmatrix} \cos\theta_1 & \cos\theta_2 \\ \sin\theta_1 & \sin\theta_2 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \qquad (21)$$

with $g_i^2 = a_i . r_e . \sum q$.

The optimal panning gains g_i are obtained by inverting the same matrix as in the low-frequency case (17) to obtain the non-normalized power gains a_i, and then applying the power-normalization condition (6). The solutions a_1 and a_2 of (21) or (17) are both non-negative provided that the target direction points between the two loudspeakers, which is the natural rule for selecting the adequate loudspeaker pair in a 2-D multichannel system.

The optimal panning law for high frequencies is simply the square root of the optimal low-frequency panning law given by equations (17) and (19):

$$g_1 = \sqrt{\frac{a_1}{a_1 + a_2}} \; ; \; g_2 = \sqrt{\frac{a_2}{a_1 + a_2}} \qquad (22)$$

We called this panning law "vector-based intensity panning" (VBIP) in [15]. It is plotted on Figure 5 (dash-dotted curve) for a pair of loudspeakers at azimuths $\pm 30°$, and compared to "angle-based intensity panning", or ABIP, (dotted curve) which is defined by:

$$g_1 = \sqrt{\frac{\theta - \theta_2}{\theta_1 - \theta_2}} \; ; \; g_2 = \sqrt{\frac{\theta - \theta_1}{\theta_2 - \theta_1}} \qquad (23)$$

Since the intensity panning laws are obtained by taking the square root of the amplitude panning laws, the angular reproduction error between ABIP and VBIP is obviously the same as between ABAP and VBAP. For reasonable aperture angles, angle-based panning (ABP) yields nearly optimal angular localization at low frequencies or high frequencies. A practical advantage of ABP over VBP is that the panning laws are essentially independent of the position of the loudspeakers (requiring only linear mapping applied to the target azimuth prior to applying the panning laws).

Like VBAP, VBIP naturally extends to sound spatialization over 3-D loudspeaker arrays, by selecting the three closest loudspeakers. The matrix **S** of equations (1) or (4) is then a 3×3 matrix whose inversion yields non-negative amplitude or power panning weights if the dot products of the target direction vector with the loudspeaker direction vectors are all positive. This 3-D panning technique can be used to produce fly-over effects on a 2-D horizontal loudspeaker array, by introducing a "virtual loudspeaker" located at the zenith position, as proposed in [15]. For a centrally located listener, the illusion of a sound located above the head can be created by distributing the amplitude -or power- over the real loudspeakers so that the velocity and intensity vectors (1) and (4) have zero magnitude.

1.5. Binaural techniques

1.5.1. Binaural encoding and decoding

The general encoding/decoding model of Figure 1 applies in the context of head-related stereophony. The essential difference with Ambisonic or sound field related techniques is that the encoding method now takes into account the presence of the head. The role of the 3-D audio encoder, in this case, is to capture (or synthesize) the pressure signals at the two ears for a plane wave of direction (θ, ϕ) [11] [13]. The acoustic version of the encoder is typically a dummy head. The electronic version of the encoder uses a pair of filters (left and right) which model the head-related transfer functions (HRTFs) measured on a dummy head or a human subject [12] [13] [32]–[34]. Because the binaural signals encode acoustic transformations due to the diffraction of the sound wave by the head and torso, the encoded signal depends on the morphology of the head. The general role of the decoder is to produce, from the left and right signals, the proper signals to feed a pair of headphones or two or more loudspeakers. The design of decoders for reproduction over loudspeakers is discussed in section 1.5.3 below.

Stricly speaking, in the case of headphone reproduction, satisfying the "re-encoding principle" only requires a spectral correction to compensate for the effect of the headphones [13]. The goal of this correction is to ensure that, if the ear signals are recorded during playback with the head used for encoding and with the microphones at the same positions in the ear canal, the same binaural signal is obtained. However, this is not exactly sufficient to ensure perfect reproduction for any listener. Perfect reproduction means that the signals at the ears of the listener during playback are the same as if this same listener were present in the original scene. This requires, ideally, that the same head be used for encoding and for listening.

The decoder can therefore be used to adapt the reproduced signals to the particular pair of headphones used, but also, as much as possible, to the morphology of the listener. Since this correction cannot depend on the direction (θ, ϕ), the adaptation of a binaural recording to the individual characteristics of the listener cannot be perfectly achieved for all directions unless the encoding and listening heads are the same. The design goal must therefore be reduced to achieving correct reproduction in a reference sound field. The two mostly used conventions are free-field equalization (correct reproduction ensured for a frontal plane wave), and diffuse-field equalization (correct reproduction for a superposition of uncorrelated planes waves uniformly distributed in direction) [11]. Methods for diffuse-field equalization of the binaurally encoded signals are discussed in [35], where the results show that it is more robust and yields better repeatability than free-field equalization.

1.5.2. Binaural synthesis

Figure 6 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 into its minimum-phase and all-pass components, so that interaural time differences (ITD) and spectral transformations are modelled and rendered independently [36], [37]. According to this factorization, the left and right HRTF filters are decomposed as follows:

$$\begin{aligned} L(\theta, \phi, f) &= \tau_L(\theta, \phi) \cdot \underline{L}(\theta, \phi, f) \\ R(\theta, \phi, f) &= \tau_R(\theta, \phi) \cdot \underline{R}(\theta, \phi, f) \end{aligned} \qquad (24)$$

where $\tau(\theta, \phi)$ represents a frequency-independent delay and $\underline{L}$ represents the minimum-phase filter having the same magnitude frequency response as L. For the purpose of binaural synthesis using minimum-phase HRTF models, the ITD should be derived from the interaural excess-phase difference [37] [36]. For each measured HRTF, the excess phase is defined as in [38], by subtracting the minimum phase from the original phase. The minimum phase is derived from the Hilbert transform of the log-magnitude frequency response:

$$\mathrm{Arg}(\underline{L}(f)) = \mathrm{Re}(\mathrm{Hilbert}(-\log(|L(f)|)) \qquad (25)$$

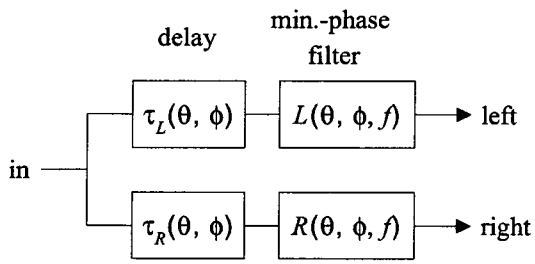

Figure 6: Binaural synthesis for one source.

Figure 14 (thick solid curve) shows the interaural excess-phase difference computed at azimuth 60° in the horizontal plane (0° elevation) from the KEMAR dummy head mesurements collected by Gardner and Martin [34]. It can be seen that the excess-phase difference is linear up to about 8 kHz, and tends to follow the same slope across the whole frequency range. This appears to be a quasi-systematic behaviour for any head and any direction. Departure from perfect linearity typically takes the form of a phase rotation of 2π corresponding to a deep notch in the HRTF magnitude for one of the two ears. Approximating the excess-phase difference by a pure delay appears to have no perceptual consequence.

Figures 15 and 16 (thick solid curve) show the computed ITD vs. azimuth in the horizontal plane, at low and high frequencies –indicating a slight frequency dependence of ITD for azimuths close to $\pm 100°$. Ignoring this particular phenomenon, we find that the variation of ITD with direction is well approximated by a spherical head model with diametrally opposite ears [37]:

$$ITD = \frac{d}{2c}(\arcsin(\cos\phi.\sin\theta) + \cos\phi.\sin\theta) \qquad (26)$$

where d denotes the distance between the two ears and c is the speed of sound.

The diffuse-field equalized models of the minimum-phase HRTFs $\underline{L}$ and $\underline{R}$ are derived from the equalized magnitude frequency responses. A variety of digital filter design techniques can be used to obtain FIR or IIR models of these transfer functions [39] [36]. IIR filters of order 12 to 16 can retain most of the perceptually relevant features of the HRTFs. In order to produce dynamic variations of the apparent direction of the sound without audible artifacts, the delay lines and the filters in Figure 6 must be made variable. This can be achieved by use of fractional delay lines [40] and cross-fading techniques using paired filters [36] [20]. As a result, the total estimated cost of the binaural encoding module is in the order of 100 operations per input sample, i. e. about 5 MIPS (million instructions per second) for a sample rate of 48 kHz [20] [36].

The incremental cost per additional sound source is at least ten times larger with binaural synthesis than with Ambisonic or pairwise panning techniques. This suggest applying these techniques to binaural synthesis (i. e. designing multichannel decoders for headphones). In the second part of this paper, we will present an objective performance evaluation of this approach, and compare it with an alternative method proposed in [20], based on the "binaural B format" encoding/decoding scheme.

1.5.3. Decoding binaural signals over loudspeakers

Applying the re-encoding principle to the design of loudspeaker decoders for binaural signals leads to specifying the decoding matrix **D** (in Figure 1) as the inverse of the acoustic transfer function matrix (loudspeakers to ears). In free-field listening conditions (ignoring the reflections and reverberation in the listening environment), the entries in this acoustic transfer matrix are free-field HRTFs –which should be diffuse-field equalized if this convention is adopted for the incoming binaurally encoded signal. An individualized decoder can be designed from HRTFs measured on the listener.

The simplest case is the reproduction over two lousspeakers for a single listener, which has been widely studied in the litterature (see, e. g., [38] [13]). The technique is often called "cross-talk cancellation" and is extensively reviewed and evaluated, both objectively and subjectively, in [14]. In practice, it is unrealistic to expect accurate reproduction at the ears of the listener for frequencies higher than about 2 kHz, since that would require positionning the listener's ears within less than a centimeter from their ideal position [14]. This limits the fidelity of the reproduction for rear and elevated sounds, and suggests employing headtracking technology to dynamicly adapt the decoding matrix **D** to the position and orientation of the listener [14]. Since the diffuse-field equalized HRTFs magnitude spectra of different listeners for a given direction tend to be very similar up to 4 or 5 kHz [35], the 2 kHz limit also implies that listener-specific adaptation of the decoder is essentially useless except for large differences in head size (which affect the ITD).

This approach can be generalized to accomodate an arbitrary number of loudspeakers N, and an arbitrary number of "ears" P [23] [24]. The multichannel acoustic transfer function is then represented by a $N \times P$ matrix. The pseudo-inverse of this matrix provides a generally valid solution to the design of the decoder [23]. When the problem is underdetermined ($N > P$), this will amount to selecting the solution of minimal power among the infinity of possible solutions. When the problem is overdetermined ($N < P$), it will provide the least squares solution. A general optimization framework is reviewed in [24], addressing situations where the acoustic transfer functions are not minimum-phase (which can happen particularly if reflections in the listening environment are taken into account). Configurations with $P > 2$ (more than two ears) arise in the reproduction for several listeners located at constrained listening positions [23]. Alternatively, one can consider a modified binaural encoding technique where the sound field is sampled at multiple points in the vicinity of each ear (with or without the head present), in order to reduce the sensitivity of the reproduced 3-D audio image to movements of the listener [41]. One such approach is the binaural B format recording technique defined in section 2.3.3.

1.6. Summary and concluding notes

In the first part of this paper, we have reviewed the fundamental principles and the implementation of Ambisonic encoding and decoding techniques. A formal connection between Ambisonics and discrete multichannel panning techniques is established by showing that "optimal" discrete panning laws (with regards to the velocity-based low-frequency localization theory) can be interpreted as the result of "local Ambisonic decoding" over a subset of loudpeakers selected according to the target direction of the virtual sound source. Optimal pairwise amplitude (resp. intensity) panning can be interpreted as reconstruction of the velocity (resp. intensity) vector for any target direction by means of a **local interpolation** technique. For low-frequency signals, this coincides with the vector base amplitude panning (VBAP) method proposed by Pulkki [10]. For high-frequenccy signals, we define a variant of this technique, called "vector base intensity panning" (VBIP) [15]. Ambisonic decoding, on the other hand, achieves this vector reconstruction via a **global interpolation** technique (involving all the loudspeakers). In the second part of this paper, we will compare these two approaches by use of objective localization performance criteria, at the ideal listening position.

We have drawn another parallel between Ambisonic and binaural techniques within the general encoding/decoding framework of Figure 1. In both approaches, the encoder samples the sound field at a set of points in space, and the decoding matrix can be designed according to a "re-encoding principle": decoder design is a linear inversion problem involving the acoustic tranfer function matrix from the loudspeakers to an array of microphones representing the "acoustic" version of the encoder (as opposed to the "electronic" version which is used for spatializing monophonic source signals). The fundamental difference is that, for binaural techniques, the sound field is sampled in the presence of the head and at the two ears, while, in Ambisonics, the head is absent and the sound field is sampled at the position of its center. Both techniques can be extended by increasing the number of transducers used to sample the sound field in the vicinity of these reference points (equivalent to increasing the order M in the case of Ambisonics). In both cases, the anticipated result is to extend the frequency range and/or enlarge the listening area where correct reconstruction of the acoustic pressure signals at the ears is achieved.

Theoretically, when the order of complexity is increased to the point that the size of the reproduction area becomes larger than the size of the head, Ambisonic and binaural techniques become equivalent in performance. High-order Ambisonics then becomes analogous to sampling the sound field over a larger volume or area, since, for a fixed frequency range, increasing the order of the harmonics is analogous to sampling the sound field in a larger zone around the center point [5]. The head-related approach then has the practical disadvantage of encoding the morphology-dependent diffraction effects. At this stage, a connection can be envisionned between Ambisonic techniques and Wave Field Synthesis (WFS), since WFS achieves sound field reconstruction within a volume (or area) from sampling along its frontier [42] [43]. This analogy is worked out and discussed in [7], and suggests a unified approach to soundfield reconstruction using multichannel systems. High-order sound field reconstruction techniques seem a promising approach to extending the size of the listening area at low frequencies, but remain impractical at higher frequencies (because of the need to limit the number of channels). At higher frequencies, one could consider relying on the intensity-based criteria which are used to optimize Ambisonic decoders.

2. OBJECTIVE PERFORMANCE EVALUATION AND EFFICIENT METHODS FOR BINAURAL SYNTHESIS

In this second part, we present an objective comparison of the localization performance of four different positional audio reproduction techniques, previously reviewed in sections 1.3 and 1.4:

- Two "low-frequency" techniques:
  - First-order low-frequency Ambisonics: equation (16) with $k_0 = 2/N$ and $k_1 = 4/N$.

- Vector-based amplitude panning (VBAP): equations (17) and (19).

- Two "high-frequency" techniques:
  - First-order high-frequency Ambisonics: equation (16) with $k_0 = \sqrt{2/N}$ and $k_1 = \sqrt{4/N}$.
  - Vector-based intensity panning (VBIP): equations (17) and (22).

The comparisons assume an ideally located listener at the center of a hexagonal reproduction system with loudspeakers at azimuths $\pm 30°$, $\pm 90°$ and $\pm 150°$ in the horizontal plane. We first present the performance evaluated according to the localization criteria reviewed in sections 1.2.2 and 1.2.3. The same techniques are then compared in terms of reproduction of the HRTF spectral cues and ITD cues at the ears of the listener. In the light of these results, we discuss the application of these techniques to binaural synthesis, and review a computationally efficient alternative approach decoupling ITD and spectral cues [20]. This second comparison includes two additional techniques:

- Second-order low-frequency Ambisonics: encoding to $[W\ X\ Y\ U\ V]$, with decoder designed by the pseudo-inverse technique, equation (14).
- Binaural synthesis using the recently proposed "binaural B format" scheme [20], which uses six channels for horizontal-only encoding.

2.1. Sound field reconstruction at the sweet spot

2.1.1. Angular localization

First-order Ambisonic decoders for regular polygons, eq. (15), produce the desired direction for both the velocity and the intensity vectors (whether the decoder is optimized for low frequencies or high frequencies). As illustrated in Figure 7, pairwise discrete amplitude or intensity panning techniques do not verify this desirable property.

VBAP ensures the correct direction for the velocity vector, i. e. correct angular reproduction of low-frequency sounds. However, the direction of the intensity vector tends to pull towards the loudspeaker closest to the target direction. The angular error, for high-frequency sounds, reaches 8° for a target azimuth of 15°. This effect also blurs the localization of wide-band sounds, as high frequencies tend to "stick to the loudspeakers". A comparable effect is seen for VBIP, which yields the correct localization angle at high frequencies but tends to pull low-frequency sounds towards the middle direction. We note that the choice of amplitude or power normalization for VBAP, eq. (18) or (19), does not affect these results, since it cancels out in equation (1).

Similar results would be obtained for the "angle-based" pairwise panning techniques (ABAP, ABIP) presented in section 1.4. These results indicate that intensity-based discrete panning techniques (VBIP or ABIP) are of more general interest than their amplitude-based (velocity-based) counterparts for reproduction over multichannel loudspeaker systems, particularly in installations designed to accommodate medium to large audiences. For personal or small-scale playback systems, this study suggests replacing the frequency-independent panning coefficients (h_j in Figure 1) by variable shelving filters, in order to implement a hybrid discrete panning technique matching VBAP at low frequencies and VBIP at high frequencies.

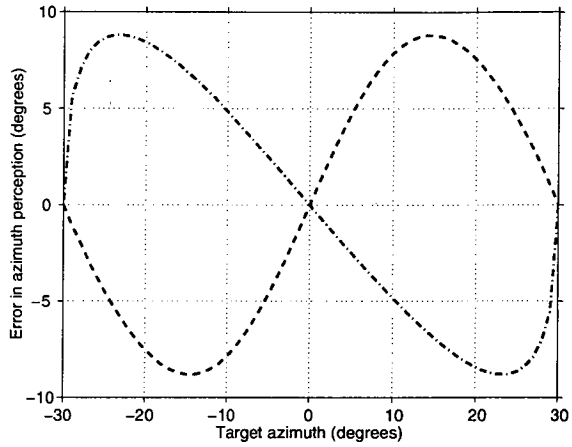

Figure 7: Angle reconstruction error for VBAP at high frequencies (dashed) and VBIP at low frequencies (dash-dotted), for a 60° separation between the loudspeakers.

2.1.2. Stability of the spatial image

The velocity and intensity factors r_v and r_e defined in sections 1.2.2 and 1.2.3 can be interpreted as measures of the stability of the spatial image under head rotation [28] [25] [16]. As shown in [28] or [16], r_v is approximately equal to the ratio of the reproduced ITD vs. the target ITD at low frequencies: $r_v = 1$ indicates correct lateralization of the reproduced sounds. If $r_v < 1$, the localization of sound events is pulled towards the median plane of the head, and the apparent position of sound sources therefore varies with head orientation. Gerzon describes a similar effect for high frequencies when $r_e < 1$, and mentions that the angular displacements of the sound image under head rotation are roughly proportional to $1 - r_e$.

A property of Ambisonic decoders for regular layouts is that r_v and r_e are independent of the target azimuth and the number of loudspeakers N [16]. Low-frequency Ambisonic decoders ensure $r_v = 1$ by design. High-frequency decoders maximize r_e, yielding $r_e = 0.707$ for first-order decoders, and $r_e = 0.866$ for second-order decoders. On the contrary, with discrete panning techniques, the values of r_v and r_e vary with the target az-

imuth, as shown in Figure 8. Unlike Ambisonics, discrete panning guarantees $r_v = r_e = 1$ in the direction of a loudspeaker. The minimum value (maximum instability) is obtained for median positions, and is lower for a wider separation of the loudspeakers.

Replacing eq. (19) or (22) in eq. (1) or (4) leads to the following expressions:

- for VBAP:

$$r_v.\mathbf{s}_v = \frac{\mathbf{s}_1.a_1 + \mathbf{s}_2.a_2}{a_1 + a_2}, \qquad (27)$$
$$r_e.\mathbf{s}_e = \frac{\mathbf{s}_1.a_1^2 + \mathbf{s}_2.a_2^2}{a_1^2 + a_2^2},$$

- for VBIP:

$$r_v.\mathbf{s}_v = \frac{\mathbf{s}_1.\sqrt{a_1} + \mathbf{s}_2.\sqrt{a_2}}{\sqrt{a_1} + \sqrt{a_2}}, \qquad (28)$$
$$r_e.\mathbf{s}_e = \frac{\mathbf{s}_1.a_1 + \mathbf{s}_2.a_2}{a_1 + a_2},$$

where a_1 and a_2 are functions of the target azimuth θ according to equation (17). This implies the following relations, valid for any target azimuth and any loudspeaker layout:

- $r_v(\text{VBAP}) = r_e(\text{VBIP})$,
- $r_v(\text{VBAP}) > r_v(\text{VBIP})$,
- $r_e(\text{VBAP}) > r_e(\text{VBIP})$.

These relations are verified in Figure 8 and in similar plots shown in [15], where Ambisonics and vector-based discrete panning are compared over a non regular 3/2-stereo ("5.1") layout. Although these relations might suggest that amplitude panning systematically yields better image stability than intensity panning, it can be shown that equations (27) and (28) all yield identical plots of r_e or r_v vs. **perceived** azimuth. This comforts the conclusions of the previous section in favor of intensity panning in discrete multichannel reproduction systems.

These comparisons, based on theoretical localization criteria, illustrate the benefits of Ambisonics: consistent localization across the whole frequency range according to both velocity and intensity theories, and uniform performance (image stability) irrespective of the target azimuth. Discrete panning, on the contrary, tends to reveal the positions of the loudspeakers to the listener and reproduce a less coherent virtual sound scene. Still, Figure 8 indicates a poor performance of first-order Ambisonics in terms of image stability at mid and high frequencies. However, it shoud be noted that the two approaches, as compared in this section, do not make use of the same number of encoding channels (3 vs. 6). A second-order Ambisonics system (using 5 encoding channels) would yield values of r_e more comparable to those obtained with the 6-channel discrete panpot (although still inferior).

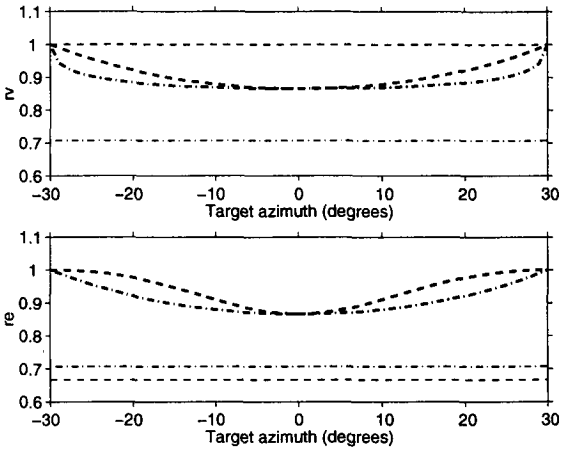

Figure 8: Stability of sound event vs. target azimuth. Top: velocity factor r_v (low frequencies). Bottom: intensity factor r_e (high frequencies). Thick dashed curve: VBAP. Thick dash-dotted curve: VBIP. Thin dashed curve: LF Ambisonics. Thin dashdotted curve: HF Ambisonics.

2.2. Reconstruction of HRTF cues at the ears

By computing the composite pressure signals at the two ears, we hope to verify some of the assumptions and results of the previous section, particularly: (a) the concordance between the computed ITD cues and the theoretical localization criteria at low-frequencies, (b) the interpretation of the velocity and intensity factors in terms of lateralization, and (c) the respective valid frequency ranges for low- and high-frequency-optimized techniques. We also hope to gain more insight into: (d) the reproduction of high-frequency ITD cues (revealed by transient sounds [11]), and (e) the potential coloration of the source timbre.

The two ear signals can be computed via binaural synthesis, by superposing the contributions of the individual loudspeakers weighted by their respective gain factors g_i, as in [16]. This process in effect produces a "binaural downmix" of the multichannel signal which can be used for reproduction over heaphones or "reformatting" to a different loudspeaker layout [17]–[21] [24]. Within the encoding/decoding framework of Figure 1, this "virtual loudspeaker" paradigm (Figure 9) is a general approach to designing decoders for adapting any encoding format to playback over headphones or any multichannel loudspeaker layout. However, it is subject to the specific limitations of binaural techniques, discussed earlier in sections 1.5 and 1.6.

This method can also be regarded as a cost-efficient approach to binaural synthesis, where a common bank of static HRTF filters allows reproducing multiple moving sound sources [17]–[20]. It can also be interpreted as

an interpolation method for reconstructing HRTFs corresponding to any direction on the basis of a set of directions [20]. In this context, the HRTF filters reconstructed by a discrete panning technique are the result of a "local interpolation", while Ambisonic techniques realize a "global interpolation" based on spherical harmonic decomposition and involving weighted contributions from all loudspeakers in the system.

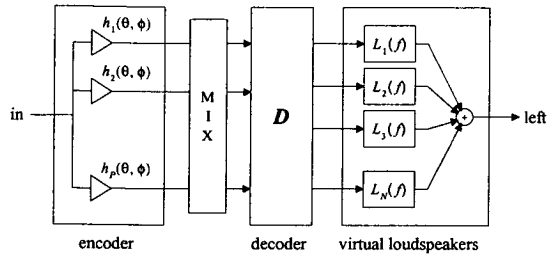

Figure 9: Binaural synthesis using "virtual loudspeakers".

2.2.1. *Reconstruction of HRTF spectra*

The incoming magnitude spectra at the ears are estimated according to the respective design assumptions of the different techniques: pressure summation for "low-frequency" (amplitude- or velocity-based) techniques, power summation for "high-frequency" (intensity-based) techniques. The HRTF spectra are therefore derived as follows (for the left ear):

- for "low-frequency" (amplitude-based) techniques:

$$L'(\theta, f) = \sum_{i=1}^{6} g_i(\theta).L(\theta_i, f) \qquad (29)$$

- for "high-frequency" (intensity-based) techniques:

$$|L'(\theta, f)| = \sqrt{\sum_{i=1}^{6} g_i^2(\theta).|L(\theta_i, f)|^2} \qquad (30)$$

The original HRTFs $L(\theta, f)$ are derived from the KEMAR dummy head mesurements collected by Gardner and Martin [34], post-processed for diffuse-field equalization. For the two intensity-based techniques (VBIP and first-order high-frequency Ambisonics), the phase spectra of the HRTFs $L(\theta_i, f)$ are ignored in the estimation of the resulting magnitude spectrum $|L'(\theta, f)|$. For low-frequency techniques, the spectrum computed by eq. (29) will exhibit "comb-filtering" effects (notches) due to the combination of signals incurring different delays. In order to isolate these effects in our computed results, we choose to assign the ITD to the contralateral HRTFs and use a minimum-phase model of the ipsilateral HRTFs. This implies that, in the summation (29), only the loudspeakers located on the contralateral side of the median plane will contribute delayed signals possibly causing comb filtering effects.

Figure 10 shows the reconstructed ipsilateral and contralateral HRTF spectra at azimuth 30° for the first-order Ambisonic decoders, compared with the target HRTFs for that azimuth (thick solid curve). Since this corresponds to the direction of a loudspeaker, the discrete panning techniques are not represented, as they yield a perfect match in this case. The two Ambisonic decoders, on the other hand, do not provide perfect reconstruction, due to the combined contributions of mutiple loudspeakers. Although the ipsilateral HRTF is reasonably well matched, we note that the high-frequency decoder (dash-dotted curve) reduces the interaural level difference above 700 Hz. The other noticeable effect is the deep notch near 1 kHz on the contralateral ear (dashed curve). The other notches on this curve are less relevant since the assumption of pressure summation becomes invalid at higher frequencies.

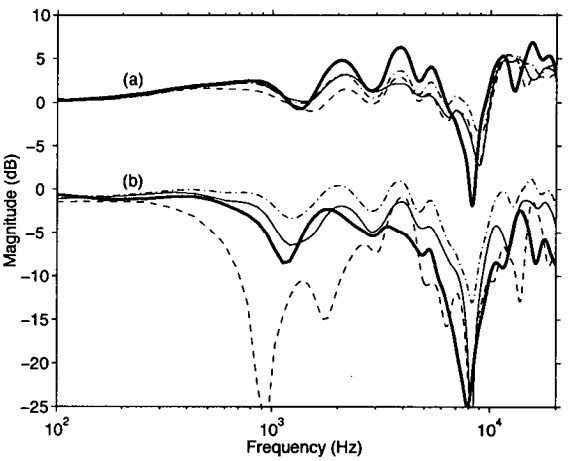

Figure 10: Reconstructed HRTF spectra for $\theta = 30°$, with half-tone resolution. (a): ipsilateral. (b): contralateral. Thick solid curve: original diffuse-field equalized HRTF. Dashed curve: 1st-order LF Ambisonics. Dash-dotted curve: 1st-order HF Ambisonics. Thin solid curve: "binaural-B format" encoding/decoding scheme.

Figures 11 and 12 show similar results obtained for azimuth 60°. They also include the second-order low-frequency ambisonic decoder (top thin solid curve) and, on the bottom graph, the discrete panning techniques VBAP (dashed) and VBIP (dash-dotted). Again, the different techniques perform similarly (but none of them perfectly) on the ipsilateral ear, while the largest errors appear on the contralateral ear due to the comb filtering effects, starting at about 500 Hz. The second-order Ambisonic decoder appears to perform better than the first-order de-

coder, and comparably to VBAP. For power spectrum reconstruction, VBIP performs somewhat better than the first-order Ambisonic decoder above 3 kHz, but both reduce the interaural level differences. Figure 13 shows the L_2-average of the dB error over all azimuths, which tends to confirm this comparative evaluation of the different methods.

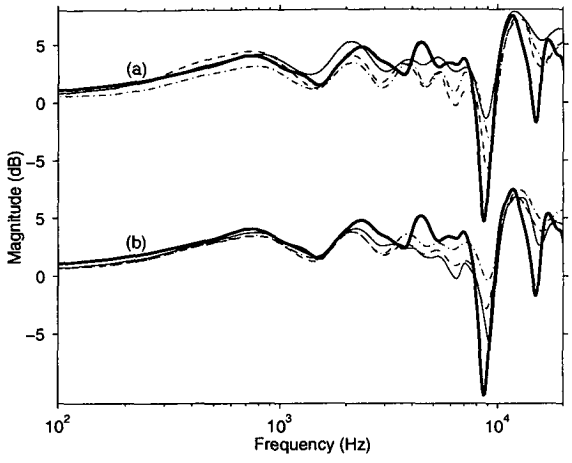

Figure 11: Reconstructed ipsilateral HRTF spectrum for $\theta = 60°$. Thick solid curve: original. Dashed (a): 1st-order LF Ambisonics; (b): VBAP. Dash-dotted (a): 1st-order HF Ambisonics; (b): VBIP. Thin solid curve (a): 2nd order LF Ambisonics; (b): binaural-B format. The same layout and curve types are used in Figures 12 to 16.

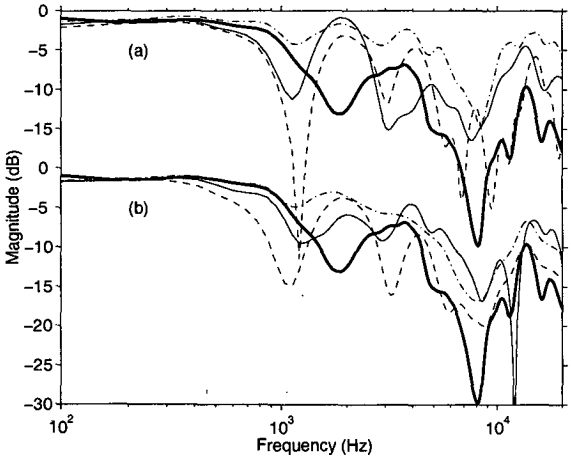

Figure 12: Reconstructed contralateral HRTF spectrum for $\theta = 60°$. Layout and curve types as in Fig. 11.

2.2.2. *Reconstruction of interaural time differences*

The ITD is derived, for all techniques, by the method described in section 1.5.2, applied to the left and right HRTFs reconstructed according to equation (29). Figure 14 shows the interaural excess-phase difference computed for target azimuth 60°. The curves obtained for Ambisonic or discrete panning techniques differ substantially from the original above 1 kHz. The noticeable feature, on any of these curves, is the reduced slope at higher frequencies, indicating a lower interaural group delay than in the original HRTFs. A similar behaviour is observed in [16], for various Ambisonic decoders, on plots of the interaural phase difference (containing the minimum-phase component).

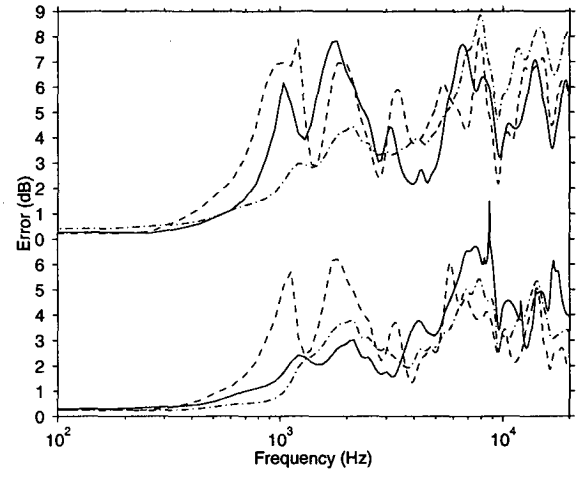

Figure 13: L_2-averaged dB error over all azimuths. Layout and curve types as in Fig. 11.

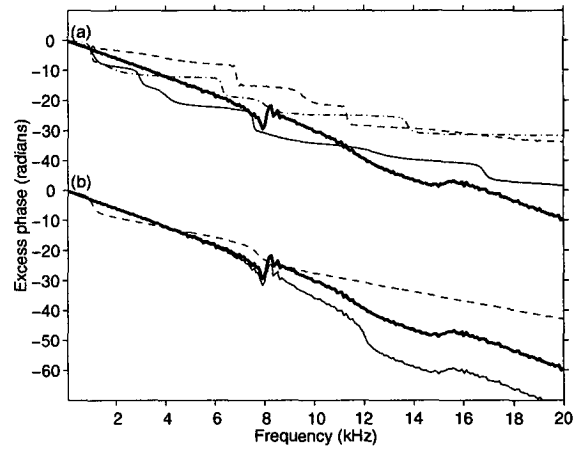

Figure 14: Reconstructed interaural excess-phase difference for $\theta = 60°$. Layout and curve types as in Fig. 11.

Estimated values for the ITD at low-frequencies and high-frequencies were derived from each of these curves over all azimuths, in 5° steps. The estimated ITD is obtained by researching, within the considered frequency range, the largest frequency interval in which the excess-phase is a quasi-linear function of frequency. The results

are plotted in Figures 15 and 16 over all azimuths.

At low-frequencies, the ITD is reasonnably well matched by all methods (except for the high-frequency Ambisonic decoder), although it is systematically inferior to the natural ITD. Increasing the order of the Ambisonic decoder appears to significantly improve the reproduction of low-frequency ITD cues –which was not predicted by the theory. With pairwise panning techniques, the low-frequency ITD curves appear to conform to the predictions of sections 2.1.1 and 2.1.2: degradation of performance for intermediate positions between two loudspeakers, and better performance for VBAP. The low-frequency ITD varies smoothly vs. azimuth with Ambisonic techniques, while, with pairwise panning, the lateralization tends to increase in successive steps from a loudspeaker to the next. This latter effect is emphasized at high frequencies, where the performance is dramatically degraded (and worse for Ambisonics than for discrete panning).

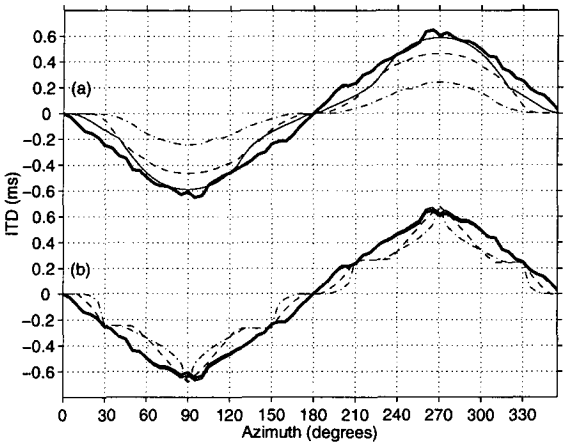

Figure 15: Estimated ITD vs. azimuth for frequencies below 700 Hz. Layout and curve types as in Fig. 11.

As a general conclusion, the computed HRTF cues indicate a reduced lateralization of sound events: the reproduced ITD is systematically inferior to the natural value, and so are the interaural level differences. This indicates, as noted in [16], a tendency for sound events to be localized above the horizontal plane, especially with first-order Ambisonics. The most notable other effects are: (a) the spectral distortions in the range 500 Hz – 2 kHz, potentially detrimental to both the localization and the reproduced timbre, and (b) the inability, for all techniques considered so far, to faithfully reproduce ITD cues at higher frequencies (except, of course, in the direction of the loudspeakers with discrete panning techniques). Noticeable improvements are obtained when raising Ambisonics from first-order to second order, with an average performance comparable to that of the 6-channel pairwise

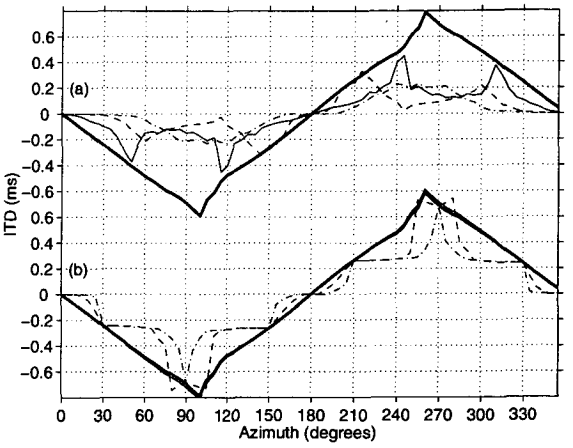

Figure 16: Estimated ITD vs. azimuth for frequencies between 1 kHz and 6 kHz. Layout and curve types as in Fig. 11.

panpot –but more uniform across all azimuths.

2.3. Efficient binaural synthesis preserving interaural time differences

The above results illustrate the limitations of the virtual loudspeaker paradigm for accurate binaural synthesis. Whenever an attempt is made to reconstruct mixed-phase HRTFs (containing the delay component) by combining other mixed-phase HRTFs containing different delays, this is likely to result in incorrect ITD cues and spectral distortions due to phase cancellations. A closer approximation of the original HRTFs may only be obtained by increasing the number of encoding channels. For discrete panning techniques, this implies a larger number of virtual loudspeakers (to reduce the time differences between adjacent speakers). For Ambisonics, it implies raising the order of spherical harmonics (and also more loudspeakers).

An alternative approach is to encode the left and right ear signals separately in a multichannel format, and therefore synthesize ITD cues explicitly for each individual source [20]. Since the decoder then involves the linear combination of "coincident" signals, this eliminates spectral distortions due to comb filter effects, and naturally ensures correct reproduction of ITD cues over the whole frequency range.

2.3.1. General approach

This approach relies on a linear decomposition of the minimum-phase models of the HRTFs, denoted $\underline{L}$ and $\underline{R}$ in equation (24), with the purpose of achieving a separa-

tion of the direction and frequency variables:

$$\begin{aligned} \underline{L}(\theta,\phi,f) &= \sum_{j=1}^{P} h_j(\theta,\phi).L_j(f), \qquad (31) \\ \underline{R}(\theta,\phi,f) &= \underline{L}(2\pi-\theta,\phi,f), \end{aligned}$$

where it is assumed, for simplicity, that the left and right ear directivity characteristics are symmetric with respect to the median plane ($\theta = 0$). Figure 17 shows the implementation of the encoding and decoding modules when the spatial functions $h_j(\theta,\phi)$ are the spherical harmonics up to first order. The decomposition (31) can be considered in two ways: decomposition over a set of spectral functions, or decompositon over a set of spatial functions. These two approaches are discussed below.

Principal Component Analysis (PCA) has been used in several studies for representing HRTF spectra, measured on a population of subjects, by decomposition over a set of orthogonal spectral functions [44]–[46]. The initial studies were applied to logarithm magnitude frequency responses (dB spectra), and demonstrated the possibility of significant data reduction (as few as five principal components). Their results, however, are not directly applicable to the linear decomposition (31). The application of PCA to a linear representation of the HRTFs (complex frequency spectrum or impulse response) is proposed in [47]–[49]. In [49], the propagation delay is extracted from the impulse responses before applying the PCA (which approximates a minimum-phase reconstruction), and it is found that this modification allows reconstructing the HRTFs from a limited number of spectral basis functions.

The result of a PCA applied to HRTF data can be viewed indifferently as a decomposition over spectral functions or over spatial functions. Neither set of basis functions is known in advance: they are determined by an optimization process, which aims at explaining the variations observed in the data. Alternatively, the set of spatial functions $h_j(\theta,\phi)$ could be chosen arbitrarily. If this set is orthogonal, the corresponding set of spectral functions $L_j(f)$ can be obtained by projecting the set of measured HRTFs over the different spatial basis functions. This approach is applied in [50] to the analysis of mixed-phase HRTFs, using spherical harmonics as the spatial basis functions. In [20], we propose to achieve the decomposition (31) via spherical harmonic analysis of the minimum-phase HRTFs. This particular choice of the spatial functions may not optimize the reconstruction performance for a given number of components in the least squares sense. However, it has the practical advantage of relying on a mathematically tractable set of spatial functions which does not depend on the original HRTF data set (see section 2.3.3).

2.3.2. Binaural synthesis using the "binaural B format" encoding/decoding scheme

In this method, introduced in [20], the minimum phase HRTFs are decomposed over the spherical harmonics up to first order. Since the left and right ear signals are processed separately, 8 channels are required for directional encoding and mixing, and 8 filters are used to decode the binaural B format into a 2-channel binaural signal, as illustrated in Figure 17.

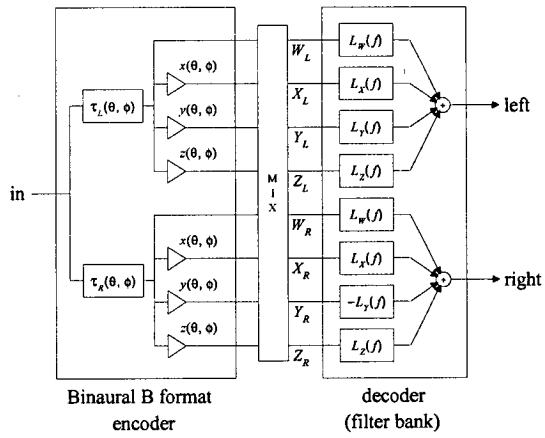

Figure 17: Binaural synthesis using the binaural B format.

The method is applied here to the KEMAR data from [34], corrected for diffuse-field equalization and minimum-phase reconstruction (see section 1.5). The decomposition (31) is achieved by projecting, at each discrete frequency f, the discrete set of complex spectra $\underline{L}(\theta,\phi,f)$ over the real-valued spherical harmonic functions $w(\theta,\phi)$, $x(\theta,\phi)$, $y(\theta,\phi)$ and $z(\theta,\phi)$ defined in equation (7). Each dot product yields a complex coefficient at each frequency for each spherical harmonic, which results in the four filters L_W, L_X, L_Y and L_Z of Figure 17.

The practical implementation of the projection requires the calculation of a discrete dot product between sampled spatial functions. If the azimuthal sampling of HRTFs is uniform at each elevation, as is the case in [34], then:

$$< A|B > = \int_S A(\theta,\phi).\overline{B}(\theta,\phi).\mathrm{d}S(\theta,\phi) \qquad (32)$$

is approximated by:

$$< A|B > = \sum_{\phi=-90°}^{\phi=90°} \mathrm{d}S(\phi) \sum_{\theta(\phi)} A(\theta,\phi).\overline{B}(\theta,\phi) \qquad (33)$$

where $\overline{B}$ denotes the conjugate of B.

A difficulty resides in the choice for the surface element $\mathrm{d}S$ at each Elevation. In principle, it should weigh each HRTF proportionally to the solid angle it covers.

However, the spherical harmonics may not constitute an orthogonal basis with respect to the resulting dot product. To eliminate potential reconstruction errors, the Gram-Schmidt orthonormalization procedure is applied to the discrete approximations of the spherical harmonics in the calculation of the projection.

In order to optimize the HRTF reconstruction in or near the horizontal plane (which is where most sound sources lie in usual applications), the projection was computed in two steps: (a) 2-D projection of the horizontal HRTFs over harmonics w, x, and y, yielding the first three basis filters L_W, L_X, L_Y, and (b) 3-D projection of the complex reconstruction error on the last harmonic, z, yielding the last filter, L_Z.

We find that all basis filters are essentially minimum phase, except for L_Z. The magnitude frequency responses of the four reconstruction filters obtained by this method are plotted in Figure 18 for the two pinnae measured in [34]. The two sets of filters are very similar except at high frequencies, which is consistent with the expected frequency range of pinna effects [11] [12]. The W component is the only significant one up to about 1 kHz, which indicates that the ear is essentially omnidirectional in this frequency range. Above 6 kHz, the W and Y responses are very similar, and 10 dB above the X component, which indicates that the directivity pattern of the ear (truncated to first order) is approximately a cardioid pointing out towards the side.

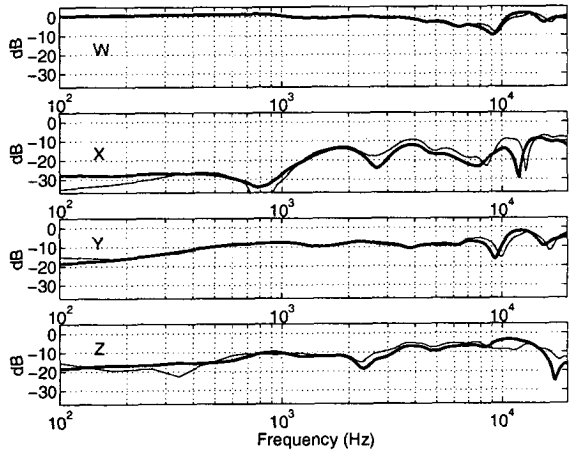

Figure 18: Magnitude spectra of the binaural B format decoding filters (HRTF reconstruction filters) for the MIT KEMAR data [34]. Thick curve: "small" pinna. Thin curve: "large red" pinna.

Figures 10–16 (bottom thin solid curve) illustrate the performance of the binaural B format encoding/decoding scheme in the horizontal plane, using 6 encoding channels (only W, X and Y –"small" pinna). The reconstruction of HRTF spectra is free of comb filter effects, which results in a substantial improvement in the range 700 Hz – 4 kHz. Although the spectra are obtained by amplitude summation, the reconstruction is comparable to the power superposition computed for "high-frequency" techniques (which ignores the phase information). This is due to the fact that, in the binaural B format, the signal components are "coincident" for each ear. Compared to the "high-frequency" techniques (VBIP or the high-frequency Ambisonic decoder), the main difference lies in the dramatic improvement of ITD cues, visible in Figures 14–16: perfect match with the original HRTFs on Figures 15 and 16 for all azimuths, and up to 6 kHz for the interaural excess-phase difference in Figure 14.

These results confirm the improved reproduction of both ITD and spectral HRTF cues over headphones. Like the "virtual loudspeaker" technique, the binaural B format encoding/decoding scheme is an efficient method for spatializing mutiple sound sources over headphones, since it only requires a bank of static HRTF filters in the decoder. The price to pay for the improved performance is a moderate increase in the complexity of the encoder, needed for synthesizing the ITD.

2.3.3. Applications of the binaural B format

We have reviewed a general approach for representing HRTFs, comprising the two following steps:

1. The decoupling of ITD and spectral cues, where the latter are represented by a minimum-phase model of the HRTF (equations (24) and (25)), which can be equalized to the diffuse field in order to eliminate all direction-independent features.
2. The linear decomposition of the minimum-phase HRTFs over a basis of spectral and spatial functions (equation (31). This can be achieved by use of statistical analysis techniques yielding an optimal reconstruction, in the least squares sense, from a minimal number of components.

This provides a method for interpolating between measured HRTFs to reproduce any direction, and for the efficient implementation of 3-D audio display systems involving multiple moving sound sources, within the encoding/decoding framework of Figure 1. As proposed in [37], the ITD can be approximated by a closed form expression, eq. (26), using an estimate of the head radius which can be derived from measured HRTFs. This representation is also useful for studying interindividual variations of HRTFs, since each "head" is entirely represented by its radius, a limited number of spectral functions, and a set of corresponding spatial functions.

The binaural B format encoding/decoding scheme (Figure 17) represents an particular choice of the set of spatial functions, for which the corresponding set of spectral functions can be derived by orthogonal projection.

Although one can expect the resulting decomposition to be less efficient in terms of reconstruction performance and/or number of components, it offers a number of practical advantages, due to the universality and mathematical tractability of the spherical harmonics [20]:

- A universal 8-channel encoding format is defined, along with a practical multichannel recording technique using two Soundfield microphones. The distance separation between the microphones can be adjusted to provide an acceptable reproduction of the ITD over all directions [20]. Obviously, the first four channels can be used as a conventional B-format signal to feed an Ambisonic decoder.
- The directional encoding functions do not involve individual features. A 3-D audio display system (over headphones or loudspeakers) can be "individualized" by modifying only the reconstruction filters in the decoder. A more compact representation is obtained for characterizing a particular head (the spatial basis functions are not needed).
- The design of multichannel loudspeaker decoders can exploit the multichannel nature of the encoding format. From a general point of view, this involves the (pseudo-) inversion of a $N \times 8$ matrix, as discussed in section 1.5.3. In a conventional 4-channel layout, one particular solution, illustrated in Figure 19, consists of coupling two crosstalk cancellers and combining the encoded signals to discrimate rear sounds from frontal sounds, in a manner similar to [51].

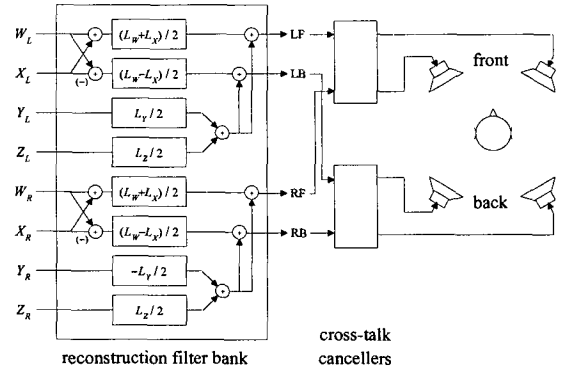

Figure 19: Decoder for transaural reproduction over 4 speakers using the "binaural B format" [20].

CONCLUSION

In this paper, we have reviewed localization theories and practical 3-D audio techniques for multichannel reproduction over loudspeakers or headphones. Connections between "optimal" multichannel discrete panning methods and Ambisonic techniques were established, and the relations between Ambisonic and binaural techniques were discussed. An objective comparative evaluation of these different techniques was carried out, assuming reproduction over a particular layout for a centrally located listener, or decoding over heaphones via the "virtual loudspeaker" paradigm. The results showed a deterioration in both the spectral cues and the interaural time difference (ITD) cues at the ears of the listener.

A general approach was proposed for improving the fidelity of the reproduction in personal playback systems over headphones or loudspeakers: it is based on decoupling ITD and spectral cues by multichannel encoding of the two ear signal separately, and related to a linear decomposition of minimum-phase HRTFs over a set of spatial encoding functions and a set of spectral reconstruction functions. The particular choice of first-order spherical harmonics for the spatial encoding functions, which defines the 8-channel "binaural B format", provides an efficient panning and mixing scheme for binaural synthesis of 3-D sound scenes containing multiple sources. Furthermore, it allows compatible mixing with a recording using a pair of non-coincident Soundfield microphones, and enables improved decoding over multichannel loudspeaker setups.

ACKNOWLEDGEMENTS

The authors wish to thank Olivier Warusfel for his useful suggestions regarding the implementation of the spherical harmonic analysis of HRTF data. Some methods disclosed in this paper are the subject of pending patent applications.

REFERENCES

[1] Steinke G. 1996. Surround sound - the next phase: an overview. In Proc. 100th Conv. Audio Engineering Society (preprint 4286).

[2] Cooper D. H., Shiga T. 1972. Discrete matrix multichannel stereo. Journal of the Audio Engineering Society, 20, 5, pp. 346-360.

[3] Gerzon M. A. 1985. Ambisonics in multichannel broadcasting and video. Journal of the Audio Engineering Society, 33, 11.

[4] Malham D. G., Myatt A. 1995. 3-D sound spatialization using ambisonic techniques. Computer Music Journal, 19, 4.

[5] Poletti M. 1996. The design of encoding functions for stereophonic and polyphonic sound systems. Journal of the Audio Engineering Society, 44, 9, pp. 948-963.

[6] Moorer J. A., Vad J. H. 1998. Towards a rational basis for multi-channel music recording. In Proc.

104th Conv. Audio Engineering Society (preprint 4680).

[7] Nicol R., Emerit M. 1999. 3-D sound reproduction over an extensive listening area: a hybrid method derived from Holophony and Ambisonics. In Proc. 16th Conf. Audio Engineering Society, Rovaniemi.

[8] Chowning J. 1971. The simulation of moving sound sources. Journal of the Audio Engineering Society, 19, 1.

[9] Theile G., Plenge G. 1977. Localization of lateral phantom sources. Journal of the Audio Engineering Society, 25, 4.

[10] Pulkki V. 1997. Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society, 45, 6, pp. 456-465.

[11] Blauert J. 1983. Spatial Hearing: the Psychophysics of Human Sound Localization. MIT Press.

[12] Begault D. 1994. 3-D Sound for Virtual Reality and Multimedia. Academic Press.

[13] Moller H. 1992. Fundamentals of binaural technology. Applied Acoustics, 36, pp. 171-217.

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

[15] Pernaux J.-M., Boussard P., Jot J.-M. 1998. Virtual sound source positioning and mixing in 5.1 - Implementation on the real-time system Genesis. In Proc. 1998 Digital Audio Effects Workshop (DAFX'98), Barcelona.

[16] Daniel, J., Rault J.-B., Polack J.-D. 1998. Ambisonics encoding of other audio formats for multiple listening conditions. In Proc. 105th Conv. Audio Engineering Society (preprint 4795).

[17] Kendall G., Martens W., Freed D., Ludwig D., Karstens R. 1986. Image-model reverberation from recirculating delays. In Proc. 81st Conv. Audio Engineering Society (preprint 2408).

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

[19] Travis C. 1996. A virtual reality perspective on headphone audio. In Proc. Audio Engineering Society U.K. Conference 'Audio For New Media'.

[20] Jot J.-M., Wardle S. A., Larcher V. 1998. Approaches to Binaural Synthesis. In Proc. 105th Conv. Audio Engineering Society (preprint 4861).

[21] Sincaglia N., Rebaud S. 1999. Product design considerations for implementing 3D audio algorithms within existing multichannel formats. In Proc. 16th Conf. Audio Engineering Society.

[22] Farrah K. 1979. The Soundfield Microphone. Wireless World, 85, pp 48-50; 99-102.

[23] Bauck J., Cooper D. H. 1996. Generalized transaural stereo and applications. Journal of the Audio Engineering Society, 44, 9, pp. 683-705.

[24] Nelson P. A., Orduna-Bustamante F., Hamada H. 1996. Multichannel signal processing techniques in the reproduction of sound. Journal of the Audio Engineering Society, 44, 11, pp. 973-989.

[25] Gerzon M. A. 1992. General metatheory of auditory localization. In Proc. 92nd Conv. Audio Engineering Society (preprint 3306).

[26] Gerzon M. A. 1992. Panpot laws for multispeaker stereo. In Proc. 92nd Conv. Audio Engineering Society (preprint 3309).

[27] Bernfeld B. 1975. Simple equations for multichannel stereophonic sound localization. Journal of the Audio Engineering Society, 23, 7, pp. 553-557.

[28] Makita Y. 1962. On the directional localization of sound in the stereophonic sound field. E.B.U. Review, Part A, 73, pp. 1536-1539.

[29] Bamford J., Vanderkooy J. 1995. Ambisonic sound for us. In Proc. 99th Conv. Audio Engineering Society (preprint 4138).

[30] Farina A. 1998. Software implementation of B-format encoding and decoding. In Proc. 104th Conv. Audio Engineering Society (preprint 4691).

[31] Malham D. G. 1992. Experience with large area 3-D Ambisonic sound systems. In Proc. Institute of Acoustics Autumn Conference on Reproduced Sound, Windermere, 8, pp. 209-216.

[32] Wenzel E. M., Whightman F. L, Foster S. H. 1988. A virtual display system for conveying three-dimensional acoustic information. In Proc. Human Factors Society 32nd Annual Meeting, pp. 86-90.

[33] Foster S., Wenzel E. M., Taylor R. M. 1991. Real-time synthesis of complex acoustic environments. In Proc. 1991 IEEE Workshop on Applications of Digital Signal Processing to Audio and Acoustics.

[34] Gardner W. G., Martin K. 1994. HRTF measurements of a KEMAR dummy-head microphone. Technical report 280, Massachussets Institute of

Technology - Media Lab Perceptual Computing Group.

[35] Larcher V., Jot J.-M., Vandernoot G. 1998. Equalization methods in binaural technology. In Proc. 105th Conv. Audio Engineering Society (preprint 4858).

[36] Jot J.-M., Larcher V., Warusfel O. 1995. Digital signal processing issues in the context of binaural and transaural stereophony. In Proc. 98th Conv. Audio Engineering Society (preprint 3980).

[37] Larcher V., Jot J.-M. 1997. Techniques d'interpolation de filtres audio-numeriques - Application a la reproduction spatiale des sons sur ecouteurs. In Proc. 4th French Congress on Acoustics, pp 97-100.

[38] Cooper D. H., Bauck J. L. 1989. Prospects for transaural recording. Journal of the Audio Engineering Society, 37, 1/2, pp. 3-19.

[39] Huopaniemi J., Zacharov N., Karjalainen, M. 1998. Objective and subjective evaluation of head-related transfer function filter design. In Proc. 105th Conv. Audio Engineering Society (preprint 4805).

[40] Valimaki 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.

[41] Abe K., Asano F, Suzuki Y., Sone T. 1995. Sound pressure control at multiple points for sound reproduction. In Proc. 15th International Congress on Acoustics, pp 349-352.

[42] Berkhout A. J., de Vries D., Vogel P. 1993. Acoustic control by wave field synthesis. Journal of the Acoustical Society of America, 93, pp. 2764-2778.

[43] Boone M. M., Verheijen E. N. G., Van Tol P. F. 1995. Spatial sound field reproduction by wave-field synthesis. Journal of the Audio Engineering Society, 43, 12, pp. 1003-1012.

[44] Martens W.L. 1987. Principal components analysis and resynthesis of spectral cues to perceived direction. In Proc. 1987 International Computer Music Conference.

[45] Kistler D., 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), pp. 1637-1647.

[46] Middlebrooks J. C. 1992. Observations on a principal components analysis of head-related transfer functions. J. Acoust. Soc. Am. 92(1). pp. 597-599.

[47] Chen J., Van Veen B. D., Hecox K. E. 1995. A spatial feature extraction and regularization model for the head-related transfer function. Journal of the Acoustical Society of America, 97, 1, pp. 439-452.

[48] Marolt M. 1996. A new approach to HRTF audio spatialization. In Proc. 1996 International Computer Music Conference.

[49] Ahn C.-Y., Pang H.-S., Sung K.-M. 1997. Model of HRTF based on complex-valued PCA considering group delay. In Proc. International Symposium on Simulation, Visualization and Auralization for Acoustic Research and Education (ASVA'97), Tokyo, pp. 365-372.

[50] Evans M., Angus J., Tew A. 1997. Spherical harmonic spectra of head-related transfer functions. In Proc. 103rd Conv. Audio Engineering Society (preprint 4571).

[51] Bruck J. 1996. The KFM 360 Surround - A purist approach. In Proc. 103rd Conv. Audio Engineering Society (preprint 4637).