Audio Engineering Society

Convention Paper

Presented at the 124th Convention
2008 May 17–20 Amsterdam, The Netherlands

Bitstream Format for Spatio-Temporal Wave Field Coder

Francisco Pinto[1] and Martin Vetterli[1]

[1] *Ecole Polytechnique Federale de Lausanne, EPFL-IC-LCAV, Station 14, 1015 Lausanne, Switzerland*

Correspondence should be addressed to Francisco Pinto (francisco.pinto@epfl.ch)

ABSTRACT

We present a non-parametric method for compressing multichannel audio data for reproduction through Wave Field Synthesis. The method consists of applying a two-dimensional filterbank to the input multichannel signal, in both time and channel dimensions, and coding the two-dimensional spectra using a spatio-temporal frequency masking model. The coded spectral data is organized into a bitstream together with side information containing scale factors and Huffman codebook information. We demonstrate how this coding method can be applied to any smooth distribution of loudspeakers in space, while obtaining a stable bitrate that is 15% lower compared to coding each channel independently.

1. INTRODUCTION

Reproduction of audio through Wave Field Synthesis (WFS) has gained considerable attention since it was introduced by Berkhout and De Vries [1]. One of the main reasons is the potential for reproducing an acoustic wave field with high accuracy at every location of the listening room. This is not the case in traditional multichannel configurations, such as Stereo and Surround, which are not able to generate the correct spatial impression beyond an optimal location in the room - the sweet spot. With WFS, the sweet spot can be extended to enclose a much larger area, at the expense of an increased number of loudspeakers.

The fact that WFS requires a large amount of audio channels for reproduction presents several challenges related to processing power and data storage. Usually, there is a trade-off between these two criteria: optimally encoded audio data requires more processing power and complexity for decoding, and vice-versa. Parametric schemes [2], for example, consist of gathering information about the acoustic scene, such as sound sources and spatial cues, and then cod-

ing this information through perceptual coding and entropy coding. Some improvements to this technique consist of jointly coding the source signals in order to reduce the required bitrate [3]. The problem associated with these approaches is that there is an assumption regarding the availability the original source signals and rendering algorithms to the end-user. In fact, disclosing a complex and innovative rendering algorithm may not be the ideal marketing solution. In addition, rendering algorithms are very demanding in terms of processing power, which may increase the price dramatically. It would be useful, instead, to provide to the end-users a final masterization ready for playback.

In this paper, we present an alternative coding approach that is non-parametric, has low complexity (relies on basic signal processing operations), does not require any source signal or rendering algorithm for reproduction, and can be applied to any smooth distribution of loudspeakers in space. The encoding method consists of transforming the multichannel audio data into the spatio-temporal frequency domain in a blockwise fashion, *i.e.*, by applying a spatio-temporal window, and then quantizing the spectrum based on a psychoacoustic model derived for spatio-temporal frequencies. The spectral coefficients are then quantized and entropy-coded, and organized into a bitstream. On the decoder side, a spatio-temporal inverse transform recovers the multichannel audio data. In this paper, the coding scheme is referred to as *Wave Field Coding* (WFC).

We evaluate the performance of WFC by feeding the encoder with multichannel audio data generated by two point sources, one in near-field and other in far-field, plus reflections, and observing the required bitrate. The results indicate that WFC achieves a reduction of around 15% in the required bitrate, compared to coding each channel individually. This reduction is obtained through a psychoacoustic model that does not take spatial masking into account, and therefore is suboptimal.

2. WAVE FIELD SYNTHESIS

The WFS technique consists of surrounding the listening area with an arbitrary number of loudspeakers, organized in some selected layout, and using the Huygens-Fresnel principle to calculate the drive signals for the loudspeakers in order to replicate any desired acoustic wave field inside that area. Since an actual wave front is created inside the room, the localization of virtual sources does not depend on the listener's position. Fig. 1 shows two possible WFS configurations.

A typical WFS playback system [4] comprises both the loudspeaker array and a rendering device, which is in charge of generating the drive signals for the loudspeakers in real-time. These signals can be derived from the particle velocity $v(t, \mathbf{r})$ measured at $\mathbf{r_{LS}}$ - where the loudspeakers are located in space - and $v(t, \mathbf{r_{LS}})$ can be calculated from the source signals $s_k(t)$ using the wave equation, plus some desired effects (room impulse response, reverberation, *etc.*). Only the source signals $s_k(t)$ and their positions in space need to be stored; each $s_k(t)$ can be coded using any desired audio format. In such a system, all end-users require a rendering device that generates $v(t, \mathbf{r_{LS}})$ out of $s_k(t)$ in real-time.

Another approach to WFS playback is, instead of providing the source signals $s_k(t)$ to the rendering device, to provide the already generated particle velocity signals $v(t, \mathbf{r_{LS}})$, or even the loudspeaker drive signals, such that the rendering procedure is much more simple and straightforward. In this case, however, all signals $v(t, \mathbf{r_{LS}})$ - one per channel - must be coded for later reproduction.

Both approaches have their own advantages and disadvantages. On the one hand, coding $s_k(t)$ is much more efficient in terms of storage, compared to coding $v(t, \mathbf{r_{LS}})$ for all channels. It is also makes user interaction more flexible, since the source locations and room effects can be manipulated in real-time. On the other hand, rendering $v(t, \mathbf{r_{LS}})$ out of $s_k(t)$ in real-time is much more greedy in terms of processing power, compared to having $v(t, \mathbf{r_{LS}})$ already available. Moreover, from a commercial point of view, it may be undesirable to disclose the source signals $s_k(t)$ or the rendering algorithm, which is a key aspect of WFS and should be protected by their inventors.

In this paper, we propose a new way of coding $v(t, \mathbf{r_{LS}})$ which may lead to a more commercially feasible implementation of the second WFS approach described above. First, we describe the

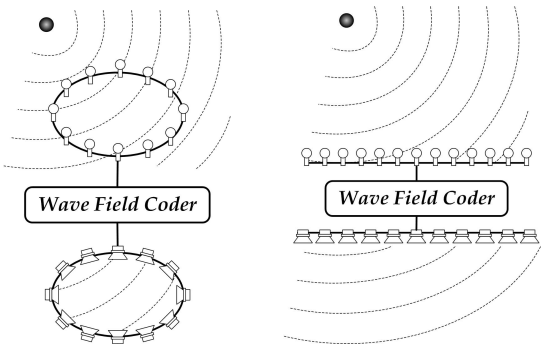


Fig. 1: Typical WFS configuration: (a) on a circle and (b) on a straight line. The acoustic wave field is recorded with microphones, the recorded audio data is coded, and the wave field is reproduced by an array of loudspeakers with a similar positioning in space. The enclosed space in (b) corresponds to the whole upper half space.

mathematical basis of our method, which consists of analyzing the physical aspects of WFS from a signal processing perspective. Then, we present a complete coding approach based of these same mathematical ideas. The resulting compression is relatively efficient, considering the simplicity of the approach.

3. CONTINUOUS SPACETIME DOMAIN

3.1. Spacetime Representation

The acoustic wave field can be modeled as a superposition of point sources in the three-dimensional space of coordinates (x, y, z). If the point sources are located at $z = 0$, as is usually the case in a WFS scenario, the three dimensional space can be reduced to the horizontal xy-plane. Under this assumption, let $p(t, \mathbf{r})$ be the sound pressure[1] at $\mathbf{r} = (x, y)$ generated by a point source located at $\mathbf{r_s} = (x_s, y_s)$, as shown in Fig. 2. The theory of acoustic wave propagation [5] states that

$$p(t, \mathbf{r}) = \frac{1}{\|\mathbf{r} - \mathbf{r_s}\|} s\left(t - \frac{\|\mathbf{r} - \mathbf{r_s}\|}{c}\right), \quad (1)$$

where $s(t)$ is the temporal signal driving the point source, and c is the speed of sound. Accordingly, given a wave field generated by an arbitrary number of point sources, $s_0, s_1, \ldots, s_{S-1}$, located at $\mathbf{r_0}, \mathbf{r_1}, \ldots, \mathbf{r_{S-1}}$, the superposition principle implies that

$$p(t, \mathbf{r}) = \sum_{k=0}^{S-1} \frac{1}{\|\mathbf{r} - \mathbf{r_k}\|} s_i\left(t - \frac{\|\mathbf{r} - \mathbf{r_k}\|}{c}\right). \quad (2)$$

If $p(t, \mathbf{r})$ is measured on a straight line, *p.e.*, the x-axis, (2) becomes

$$p(t, x) = \sum_{k=0}^{S-1} \frac{1}{\|x - \mathbf{r_k}\|} s_k\left(t - \frac{\|x - \mathbf{r_k}\|}{c}\right), \quad (3)$$

which we call the *continuous-spacetime signal*, with temporal dimension t and spatial dimension x. In particular, if $\|\mathbf{r_k}\| \gg \|\mathbf{r}\|$ for all k, then all point sources are located in far-field, and thus

$$p(t, x) \approx \sum_{k=0}^{S-1} \frac{1}{\|\mathbf{r_k}\|} s_k\left(t + \frac{\cos\alpha_k}{c} x - \frac{\|\mathbf{r_k}\|}{c}\right), \quad (4)$$

since $\|x - \mathbf{r_k}\| \approx \|\mathbf{r_k}\| - x\cos\alpha_k$, where α_k is the angle of arrival of the plane wave-front k. If (4) is normalized and the initial delay discarded, the terms $\|\mathbf{r_k}\|^{-1}$ and $\frac{1}{c}\|\mathbf{r_k}\|$ can be removed.

3.2. Frequency Representation

The spacetime signal $p(t, x)$ can be represented as a linear combination of complex exponentials with temporal frequency Ω and spatial frequency Φ, by applying a spatio-temporal version of the Fourier transform:

$$P(\Omega, \Phi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(t, x)\, e^{-j(\Omega t + \Phi x)} dt dx, \quad (5)$$

[1] In this analysis, we focus only on the sound pressure $p(t, \mathbf{r})$ and not on the particle velocity $v(t, \mathbf{r})$. In fact, $p(t, \mathbf{r})$ and $v(t, \mathbf{r})$ are very similar signals, and both can be efficiently coded with our WFC.

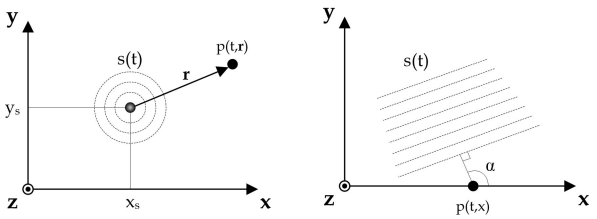


Fig. 2: The acoustic wave field is modeled by a superposition of point sources located in: (a) near-field and (b) far-field. The sound pressure in near-field can be measured at any point in space using the equation that governs the spherical wave propagation, and in far-field using an approximated result that depends only on the angle of arrival of the wave front.

which we call the *continuous-spacetime spectrum.*

Consider the spacetime signal $p(t,x)$ generated by a point source located in far-field, and driven by $s(t)$. According to (4),

$$p(t,x) = s\left(t + \frac{cos\alpha}{c}x\right), \tag{6}$$

where, for simplicity, the amplitude was normalized and the initial delay discarded. The Fourier transform is then

$$P(\Omega,\Phi) = S(\Omega)\,\delta\left(\Phi - \frac{cos\alpha}{c}\Omega\right), \tag{7}$$

which represents, in the spacetime frequency domain, a wall-shaped Dirac function[2] with slope $\frac{c}{cos\alpha}$ and weighted by the one-dimensional spectrum of $s(t)$. In particular, if $s(t) = e^{j\Omega_o t}$,

$$P(\Omega,\Phi) = \delta(\Omega - \Omega_o)\,\delta\left(\Phi - \frac{cos\alpha}{c}\Omega_o\right), \tag{8}$$

which represents a single spatio-temporal frequency centered at $\left(\Omega_o, \frac{cos\alpha}{c}\Omega_o\right)$, as shown in Fig. 3a. Also, if $s(t) = \delta(t)$, then

$$P(\Omega,\Phi) = \delta\left(\Phi - \frac{cos\alpha}{c}\Omega\right), \tag{9}$$

as shown in Fig. 3b.

If the point source is not far enough from the x-axis to be considered in far-field, (1) must be used, such that

$$p(t,x) = \frac{1}{\|x - \mathbf{r_s}\|}\delta\left(t - \frac{\|x - \mathbf{r_s}\|}{c}\right), \tag{10}$$

for which the spacetime spectrum can be shown [6] to be

$$P(\Omega,\Phi) = -j\pi e^{-j\Phi x_s} H_o^{(1)\star}\left(y_s\sqrt{\left(\frac{\Omega}{c}\right)^2 - \Phi^2}\right), \tag{11}$$

where $H_o^{(1)\star}$ represents the complex conjugate of the zero-order Hankel function of the first kind. In Fig. 3c, it can be seen that $P(\Omega,\Phi)$ has most of its energy concentrated inside a triangular region satisfying $|\Phi| \leq \frac{|\Omega|}{c}$, and some residual energy on the outside.

Note that the spacetime signal $p(t,x)$ generated by a source signal $s(t) = \delta(t)$ is in fact a Green's solution for the wave equation [5] measured on the x-axis. This means that (9) and (11) act as a transfer function between $p(t,\mathbf{r_s})$ and $p(t,x)$, depending on how far the source is away from the x-axis. Furthermore, the transition from (11) to (9) is smooth, in the sense that, as the source moves away from the x-axis, the dispersed energy in the spectrum of Fig. 3c slowly collapses into the Dirac function of Fig. 3b. In Section 4.4, we present another interpretation for this phenomenon, in which the near-field wave front is represented as a linear combination of plane waves, and therefore a linear combination of Dirac functions in the spectral domain.

3.3. Short-Spacetime Analysis

Consider an enclosed space E with a smooth boundary on the xy-plane, as depicted in Fig. 4a. Outside this space, an arbitrary number of point sources in far-field generate an acoustic wave field that equals

[2] When viewed in three dimensions, the Dirac function resembles a wall of infinite height placed on the line $\Phi - \frac{cos\alpha}{c}\Omega = 0$.

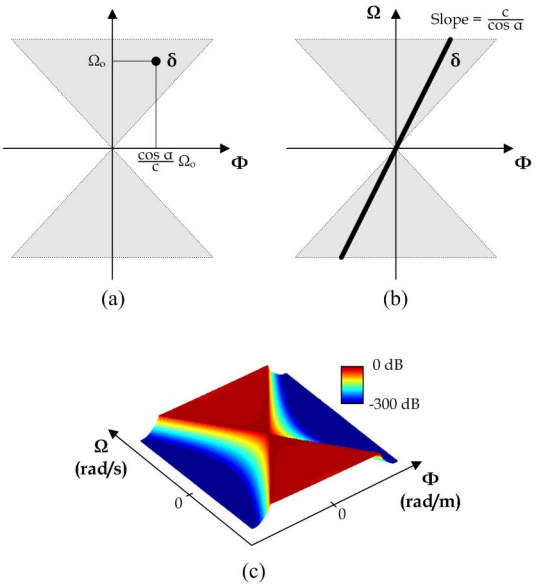


Fig. 3: Spacetime spectrum of: (a) a complex exponential $s(t) = e^{j\Omega_o t}$ in far-field, (b) a Dirac pulse $s(t) = \delta(t)$ in far-field, and (c) a Dirac pulse $s(t) = \delta(t)$ in near-field. Although (b) and (c) result from the same source signal, the differences in curvature of the wave field result in a different energy dispersion in the spectrum.

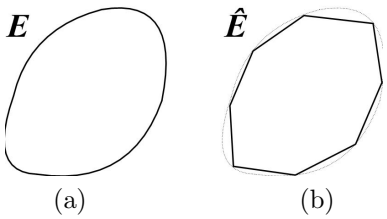


Fig. 4: Approximation of an enclosed space E by a K-sided polygon. The accuracy of the approximation increases with the number of sides in the polygon.

$p(t, \mathbf{r})$ on the boundary of E according to (2). If the boundary is smooth enough, it can be approximated by a K-sided polygon, as depicted in Fig. 4b. Consider that x goes around the boundary of the polygon as if it were stretched into a straight line. Then, (4) can be written as

$$p(t,x) = \sum_{l=0}^{K_l-1} w_l(x) \sum_{k=0}^{S-1} s_k\left(t + \frac{\cos\alpha_{kl}}{c}x\right) \quad (12)$$

$$= \sum_{l=0}^{K_l-1} w_l(x)\, p_l(t,x)\ , \quad (13)$$

where α_{kl} is the angle of arrival of the wave-front k to the polygon's side l (see Fig. 5), in a total of K_l sides, and $w_l(x)$ is a rectangular window of amplitude 1 within the boundaries of side l and zero otherwise (see Section 3.4). The windowed partition $w_l(x)\, p_l(t,x)$ is called a spatial block, and is analogous to the temporal block $w(t)\, s(t)$ known from traditional signal processing.

In the frequency domain,

$$P_l(\Omega,\Phi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_l(x)\, p_l(t,x)\, e^{-j(\Omega t+\Phi x)}dtdx\,, \quad (14)$$

$$l = 0,\ldots,K_l-1\,,$$

which we call the short-space Fourier transform. If a window $w_g(t)$ is also applied to the time domain, the Fourier transform is performed in spatio-temporal blocks, $w_g(t)\, w_l(x)\, p_{g,l}(t,x)$, and thus

$$P_{g,l}(\Omega,\Phi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_g(x)\, w_l(x)\cdot p_{g,l}(t,x)\, e^{-j(\Omega t+\Phi x)}dtdx \quad (15)$$

$$g = 0,\ldots,K_g-1\,,\ \ l = 0,\ldots,K_l-1\,,$$

where $P_{g,l}(\Omega,\Phi)$ is the short-spacetime Fourier transform of block g,l, in a total of $K_g \times K_l$ blocks.

3.4. Spacetime Windowing

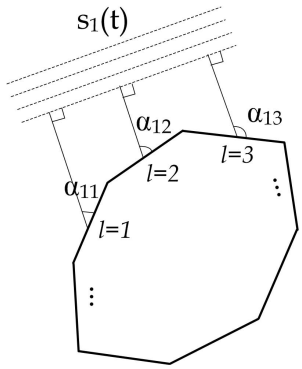


Fig. 5: Short-space analysis of the acoustic wave field. Each side of the polygon "sees" the plane wave from a different angle α.

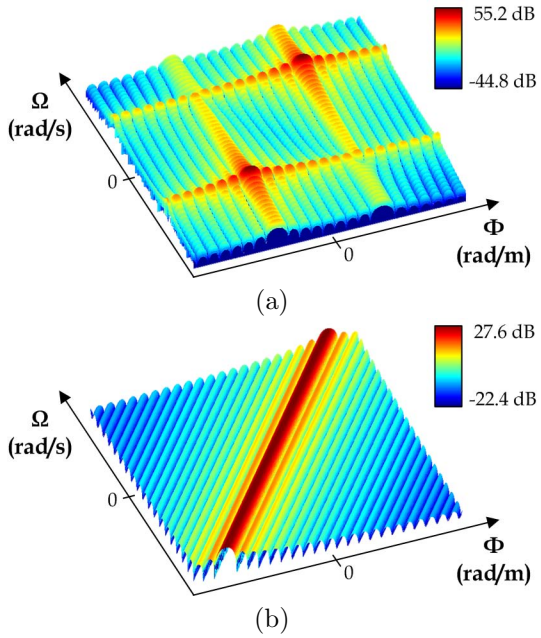


Fig. 6: Effects of windowing in the spacetime spectrum for: (a) a complex exponential $s(t) = e^{j\Omega_o t}$ in far-field and (b) a Dirac pulse $s(t) = \delta(t)$ in far-field. The parameters for both cases are: $\alpha = 0\,\text{rad}$, $\Omega_S = 2\pi \cdot 44100\,\text{rad/s}$, $\Phi_S = 2\pi \cdot 167\,\text{rad/s}$, $L_t = 48$ for (a), $L_x = 24$. In the case of the Dirac pulse, we used a larger temporal window, with $L_t = 512\,\text{samples}$, to make the approximation more accurate.

The short-space analysis of the acoustic wave field is similar to its time domain counterpart, and therefore exhibits the same issues. For instance, the length L_x of the spatial window controls the x/Φ resolution trade-off: a larger window generates a sharper spectrum, whereas a smaller window exploits better the curvature variations along x (see Fig. 3). The window type also has an influence on the spectral shaping, including the trade-off between amplitude decay and width of the main lobe in each frequency component. Furthermore, it is beneficial to have overlapping between adjacent blocks, to avoid discontinuities after reconstruction. Our WFC approach comprises all these aspects in a spatio-temporal filterbank (see Section 4.3).

The windowing operation in the spacetime domain consists of multiplying $p(t,x)$ both by a temporal window $w_t(t)$ and a spatial window $w_x(x)$, in a separable fashion. The lengths L_t and L_x of each window determine the temporal and spatial frequency resolutions.

Consider the plane wave examples of Section 3.2, and let $w_t(t)$ and $w_x(x)$ be two rectangular windows such that

$$w_t(t) = \sqcap\left(\frac{t}{L_t}\right) = \begin{cases} 1 & , \quad |t| < \frac{L_t}{2} \\ 0 & , \quad |t| > \frac{L_t}{2} \end{cases} , \tag{16}$$

and the same for $w_x(x)$. In the spectral domain,

$$W_t(\Omega) = L_t\, sinc\left(\frac{L_t \Omega}{2\pi}\right) . \tag{17}$$

For the first case, where $s(t) = e^{j\omega_o t}$,

$$p(t,x) = e^{j\omega_o\left(t + \frac{cos\alpha}{c}x\right)} w_t(t)\, w_x(x) , \tag{18}$$

and thus

$$P(\Omega,\Phi) = W_t(\Omega - \Omega_o) \cdot W_x\left(\Phi - \tfrac{cos\alpha}{c}\Omega_o\right) \tag{19}$$

$$= L_t\, sinc\left(\tfrac{L_t}{2\pi}(\Omega - \Omega_o)\right) \cdot L_x\, sinc\left(\tfrac{L_x}{2\pi}\left(\Phi - \tfrac{cos\alpha}{c}\Omega_o\right)\right) \tag{20}$$

For the second case, where $s(t) = \delta(t)$,

$$p(t,x) = \delta\left(t + \frac{cos\alpha}{c}x\right) w_t(t)\, w_x(x)\ , \qquad (21)$$

and thus

$$P(\Omega,\Phi) = \frac{c}{|cos\alpha|} W_t\left(\frac{c}{cos\alpha}\Phi\right) \star_\Phi W_x\left(\Phi - \frac{cos\alpha}{c}\Omega\right) \qquad (22)$$

$$= \frac{c}{|cos\alpha|} L_t\, sinc\left(\frac{L_t}{2\pi}\cdot\frac{c}{cos\alpha}\Phi\right) \star_\Phi L_x\, sinc\left(\frac{L_x}{2\pi}\left(\Phi - \frac{cos\alpha}{c}\Omega\right)\right) \qquad (23)$$

where $\star_\Phi$ denotes convolution in Φ. Using the property $\lim_{a\to\infty} a\, sinc(ax) = \delta(x)$, (23) is simplified to

$$P(\Omega,\Phi) \approx 2\pi\delta(\Phi) \star_\Phi L_x\, sinc\left(\frac{L_x}{2\pi}\left(\Phi - \frac{cos\alpha}{c}\Omega\right)\right) \qquad (24)$$

$$= 2\pi L_x\, sinc\left(\frac{L_x}{2\pi}\left(\Phi - \frac{cos\alpha}{c}\Omega\right)\right) \qquad (25)$$

The results in (20) and (25) are shown in Fig. 6.

4. WAVE FIELD CODER

4.1. Overview

The WFC scheme, as illustrated in Fig. 7, can be interpreted as a spatio-temporal extension of a traditional perceptual mono coder. The sampled multichannel signal, or spacetime signal, is transformed into the frequency domain by applying an MDCT filterbank to both temporal and spatial dimensions. In the spectral domain, the two-dimensional coefficients are quantized according to a psychoacoustic model derived for spatio-temporal frequencies, and then converted to binary base through entropy coding. Finally, the binary data is organized into a bitstream, together with side information necessary to decode it. On the decoder side, the bitstream is parsed, and the binary data converted back to spectral coefficients, from which the inverse MDCT recovers the multichannel signal. These steps are described in detail in the next sections.

4.2. Sampling and Reconstruction

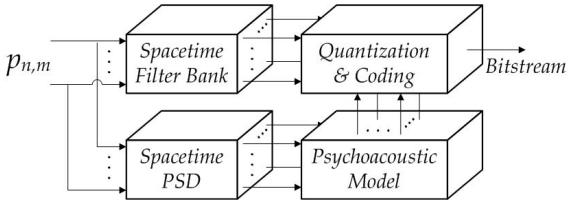


Fig. 7: Block diagram of the encoder.

In practice, $p(t,x)$ can only be measured on discrete points along the x-axis. A typical scenario is when the wave field is measured with microphones, where each microphone represents one spatial sample. If $s_k(t)$ and $\mathbf{r_k}$ are known, $p(t,x)$ may also be computed through (3). In either case, the goal is to obtain (and code) only the spacetime signal $p(t,x)$, and not the original source signals $s_k(t)$ (see Section 2).

The *discrete-spacetime signal* $p_{n,m}$, with temporal index n and spatial index m, is defined as

$$p_{n,m} = p\left(n\frac{2\pi}{\Omega_S}, m\frac{2\pi}{\Phi_S}\right)\ , \qquad (26)$$

where Ω_S and Φ_S are the temporal and spatial sampling frequencies. We assume that both temporal and spatial samples are equally spaced. The sampling operation generates periodic repetitions of $P(\Omega,\Phi)$ in multiples of Ω_S and Φ_S, as illustrated in Fig. 8. Perfect reconstruction of $p(t,x)$ requires that $\Omega_S \geq 2\Omega_{max}$ and $\Phi_S \geq 2\Phi_{max} = \frac{2\Omega_{max}}{c}$, which happens only if $P(\Omega,\Phi)$ is bandlimited in both Ω and Φ. While this may be the case for mono signals, in the case of spacetime signals there is no way to avoid spatial aliasing, unless the wave field is composed solely by far-field components. For an extensive analysis on spatio-temporal sampling and interpolation, the reader may consult Ajdler and Vetterli [6].

4.3. Spacetime-Frequency Mapping

In our WFC approach, the actual coding occurs in the frequency domain, where each frequency pair (Ω,Φ) is quantized and coded, and then stored in the bitstream. The transformation to the frequency domain is performed by a two-dimensional filterbank that represents a spatio-temporal lapped block

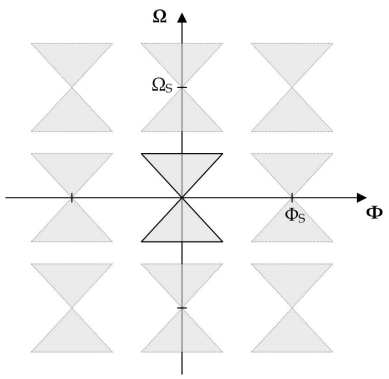


Fig. 8: Spectral repetitions that result from sampling the continuous-spacetime signal, centered on multiples of Ω_S and Φ_S.

transform. For simplicity, we assume that the transformation is separable, *i.e.*, the individual temporal and spatial transforms can be cascaded and interchanged. In this analysis, we assume that the temporal transform is performed first.

Let $p_{n,m}$ be represented in a matrix notation,

$$\mathbf{P} = \begin{bmatrix} p_{0,0} & p_{0,1} & \cdots & p_{0,M-1} \\ p_{1,0} & p_{1,1} & \cdots & p_{1,M-1} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N-1,0} & p_{N-1,1} & \cdots & p_{N-1,M-1} \end{bmatrix}, \quad (27)$$

where N and M are the total number of temporal and spatial samples, respectively. If the measurements are performed with microphones, then M is the number of microphones and N is the length of the temporal signal received in each microphone. Let also $\tilde{\mathbf{\Psi}}$ and $\tilde{\mathbf{\Upsilon}}$ be two generic transformation matrices of size $N \times N$ and $M \times M$, respectively, that generate the temporal and spatio-temporal spectral matrices $\mathbf{X}$ and $\mathbf{Y}$. The matrix operations that define the spacetime-frequency mapping can be organized as follows:

	Temporal	Spatial
Direct transform:	$\mathbf{X} = \tilde{\mathbf{\Psi}}^{\mathbf{T}}\mathbf{P}$	$\mathbf{Y} = \mathbf{X}\tilde{\mathbf{\Upsilon}}$
Inverse transform:	$\hat{\mathbf{P}} = \tilde{\mathbf{\Psi}}\hat{\mathbf{X}}$	$\hat{\mathbf{X}} = \hat{\mathbf{Y}}\tilde{\mathbf{\Upsilon}}^{\mathbf{T}}$

The matrices $\hat{\mathbf{X}}$,$\hat{\mathbf{Y}}$, and $\hat{\mathbf{P}}$ are the estimations of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{P}$, and have size $N \times M$. Combining all transformation steps in the table yields $\hat{\mathbf{P}} = \tilde{\mathbf{\Psi}}\tilde{\mathbf{\Psi}}^{\mathbf{T}} \cdot \mathbf{P} \cdot \tilde{\mathbf{\Upsilon}}\tilde{\mathbf{\Upsilon}}^{\mathbf{T}}$, and thus perfect reconstruction is achieved if $\tilde{\mathbf{\Psi}}\tilde{\mathbf{\Psi}}^{\mathbf{T}} = \mathbf{I}$ and $\tilde{\mathbf{\Upsilon}}\tilde{\mathbf{\Upsilon}}^{\mathbf{T}} = \mathbf{I}$, *i.e.*, if the transformation matrices are orthonormal.

For the WFC scheme, we have chosen a well known orthonormal transformation matrix called the Modified Discrete Cosine Transform (MDCT) [7], which is applied to both temporal and spatial dimensions. The transformation matrix $\tilde{\mathbf{\Psi}}$ (or $\tilde{\mathbf{\Upsilon}}$ for space) is defined by

$$\tilde{\mathbf{\Psi}} = \begin{bmatrix} \mathbf{\Psi_1} & & \\ \mathbf{\Psi_0} & \mathbf{\Psi_1} & \\ & \mathbf{\Psi_0} & \ddots \\ & & \ddots \end{bmatrix}, \quad (28)$$

and has size $N \times N$ (or $M \times M$). The matrices $\mathbf{\Psi_0}$ and $\mathbf{\Psi_1}$ are the lower and upper halves[3] of the transpose of the basis matrix $\mathbf{\Psi}$, which is given by

$$\psi_{b_n,2B-1-n} = w_n \sqrt{\tfrac{2}{B_n}} \cos\left[\tfrac{\pi}{B_n} \cdot \left(n + \tfrac{B_n+1}{2}\right)\left(b_n + \tfrac{1}{2}\right)\right] \quad (29)$$

$$b_n = 0, 1, \ldots, B_n - 1 \; ; \; n = 0, 1, \ldots, 2B_n - 1 \,,$$

where n (or m) is the signal sample index, b_n (or b_m) is the frequency band index, B_n (or B_m) is the number of spectral samples in each block, and w_n (or w_m) is the window sequence. For perfect reconstruction, the window sequence must satisfy the Princen-Bradley conditions [7],

$$w_n = w_{2B_n-1-n} \text{ and } w_n^2 + w_{n+B_n}^2 = 1 \,.$$

Note that the spatio-temporal MDCT generates a transform block of size $B_n \times B_m$ out of a signal block of size $2B_n \times 2B_m$, whereas the inverse spatio-temporal MDCT restores the signal block of size $2B_n \times 2B_m$ out of the transform block of size $B_n \times B_m$. Each reconstructed block suffers both from time-domain aliasing and spatial-domain aliasing, due to the downsampled spectrum. For the

[3] Note that $\mathbf{\Psi_0}$ and $\mathbf{\Psi_1}$ are overlapped in the transformation matrix $\tilde{\mathbf{\Psi}}$.

aliasing to be canceled in reconstruction, adjacent blocks need to be overlapped in both time and space. However, if the spatial window is large enough to cover all spatial samples, a DCT of Type IV with a rectangular window is used instead.

One last important note is that, when using the spatio-temporal MDCT, if the signal is zero-padded, the spatial axis requires $K_l B_m + 2B_m$ spatial samples to generate $K_l B_m$ spectral coefficients. While this may not seem much in the temporal domain, it is actually very significant in the spatial domain because $2B_m$ spatial samples correspond to $2B_m$ more channels, and thus $2B_m N$ more spacetime samples. For this reason, the signal is mirrored in both domains, instead of zero-padded, so that no additional samples are required.

4.4. Psychoacoustic Model

The psychoacoustic model for spatio-temporal frequencies is a key aspect of the WFC, and an open subject that requires further research. It requires the knowledge of both temporal-frequency masking and spatial-frequency masking, and these may be combined in a separable or non-separable way. The advantage of using a separable model is that the temporal and spatial contributions can be derived from existing models that are used in state-of-art audio coders. On the other hand, a non-separable model would be capable of estimating the dome-shaped masking effect produced by each individual spatio-temporal frequency over the surrounding frequencies. These two possibilities are illustrated in Fig. 9.

A non-separable masking model does not exist to date, and it would most likely be difficult to develop. For this reason, we chose to focus on the separable combination of temporal-frequency masking and spatial-frequency masking, although in this paper the described model is based on temporal-frequency masking only, whereas spatial-frequency masking is discarded.

The goal of the psychoacoustic model is to estimate, for each spatio-temporal spectral block of size $B_n \times B_m$, a matrix $\mathbf{M}$ of equal size that contains the maximum quantization noise power that each spatio-temporal frequency can sustain without causing perceivable artifacts. Throughout the development of the WFC, we tested three different methods for estimating $\mathbf{M}$, which are described next.

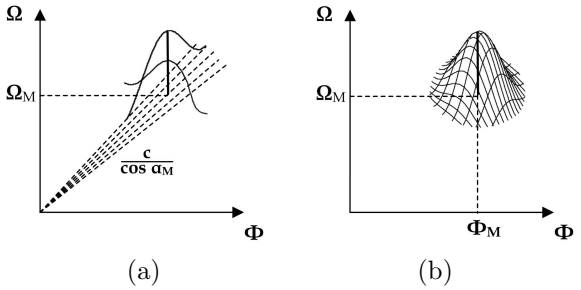


Fig. 9: Two methods for estimating the spatio-temporal masking surface induced by a given masker: (a) separable combination of temporal masking and spatial masking models and (b) a non-separable masking effect estimated from an hypothetical spatio-temporal psychoacoustic model.

4.4.1. Average based estimation

The simplest and fastest way of obtaining a rough estimation of $\mathbf{M}$ is to first compute the masking curve produced by the signal in each channel independently, and then use the same average masking curve in all spatial frequencies.

Let $x_{n,m}$ be the spatio-temporal signal block of size $2B_n \times 2B_m$ for which $\mathbf{M}$ is to be estimated. The temporal signals for the channels m are $x_{n,0}, ..., x_{n,B_m-1}$ Suppose that $\mathbb{M}[\cdot]$ is the operator that computes a masking curve, with index b_n and length B_n, for a temporal signal or spectrum. Then,

$$\mathbf{M} = \left[\begin{array}{ccc} \overline{\mathbf{mask}} & \cdots & \overline{\mathbf{mask}} \end{array} \right] , \qquad (30)$$

where,

$$\overline{\mathbf{mask}} = \frac{1}{B_m} \sum_{m=0}^{B_m-1} \mathbb{M}[x_n]_m \qquad (31)$$

$$= \frac{1}{B_m} \sum_{m=0}^{B_m-1} \mathbf{mask}_m \, . \qquad (32)$$

4.4.2. Spatial-frequency based estimation

Another way of estimating $\mathbf{M}$ is to compute one masking curve per spatial frequency. This way, the

triangular energy distribution in the spectral block $\mathbf{Y}$ is better exploited.

Let $x_{n,m}$ be the spatio-temporal signal block of size $2B_n \times 2B_m$, and Y_{b_n,b_m} the respective spectral block. Then,

$$\mathbf{M} = \left[\begin{array}{ccc} \mathbf{mask}_0 & \cdots & \mathbf{mask}_{B_m-1} \end{array} \right] , \qquad (33)$$

where

$$\mathbf{mask}_{b_m} = \mathbb{M}\left[Y_{b_n}\right]_{b_m} . \qquad (34)$$

Note that $\mathbf{mask}_{b_m}$ is actually computed from a Power Spectral Density (PSD), as shown in Fig. 7, but for the sake of simplicity we represent it with the same notation of the MDCT.

One interesting remark about this method is that, since the masking curves are estimated from vertical lines along the Ω-axis, this is actually equivalent to coding each channel separately after decorrelation through a DCT. In Section 5, we show that this method gives a worst estimation of $\mathbf{M}$ than the plane-wave method, which is the most optimal without spatial masking consideration.

4.4.3. Plane-wave based estimation

The most accurate way we found for estimating $\mathbf{M}$ was by decomposing the spacetime signal $p(t,x)$ into plane-wave components, and estimating the masking curve for each component. The theory of wave propagation states that any acoustic wave field can be decomposed into a linear combination of plane waves and evanescent waves traveling in all directions. In the spacetime spectrum, plane waves constitute the energy inside the triangular region $|\Phi| \leq \frac{|\Omega|}{c}$, whereas evanescent waves constitute the energy outside this region [6]. Since the energy outside the triangle is residual, we can discard evanescent waves and represent the wave field solely by a linear combination of plane waves, which have the elegant property described next.

As derived in (7), the spacetime spectrum $P(\Omega,\Phi)$ generated by a plane wave with angle of arrival α is given by

$$P(\Omega,\Phi) = S(\Omega)\,\delta\left(\Phi - \frac{\cos\alpha}{c}\Omega\right) , \qquad (35)$$

where $S(\Omega)$ is the temporal-frequency spectrum of the source signal $s(t)$. Consider that $p(t,x)$ has F plane-wave components, $p_0(t,x), \ldots, p_{F-1}(t,x)$, such that

$$p(t,x) = \sum_{k=0}^{F-1} p_k(t,x) . \qquad (36)$$

The linearity of the Fourier transform implies that

$$P(\Omega,\Phi) = \sum_{k=0}^{F-1} S_k(\Omega)\,\delta\left(\Phi - \frac{\cos\alpha_k}{c}\Omega\right) . \qquad (37)$$

Note that, according to (37), the higher the number of plane-wave components, the more dispersed the energy is in the spacetime spectrum. This provides good intuition on why a source in near-field generates a spectrum with more dispersed energy then a source in far-field (see Section 3.2): in near-field, the curvature is more stressed, and therefore has more plane-wave components.

As mentioned before, we are discarding spatial-frequency masking effects in this analysis, *i.e.*, we are assuming there is total separation of the plane waves by the auditory system. Under this assumption,

$$M(\Omega,\Phi) = \sum_{k=0}^{F-1} \mathbb{M}\left[S_k(\Omega)\right]\delta\left(\Phi - \frac{\cos\alpha_k}{c}\Omega\right) , \qquad (38)$$

or, in discrete-spacetime,

$$\mathbf{M} = \sum_{k=0}^{F-1} \mathbb{M}\left[S_{b_n}\right]\delta_{b_n,\frac{c}{\cos\alpha_k}b_m} . \qquad (39)$$

If $p(t,x)$ has an infinite number of plane-wave components, which is usually the case, the masking curves can be estimated for a finite number of components, and then interpolated to obtain $\mathbf{M}$. This method is illustrated in Fig. 10.

4.5. Quantization

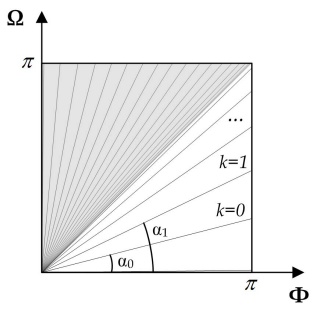


Fig. 10: Estimation of the spatio-temporal masking surface, by interpolation of the contributions of a selected number of plane waves. Note that the region outside the shaded triangle is not supposed to have plane-wave energy, but we use these profiles to help in the interpolation process.

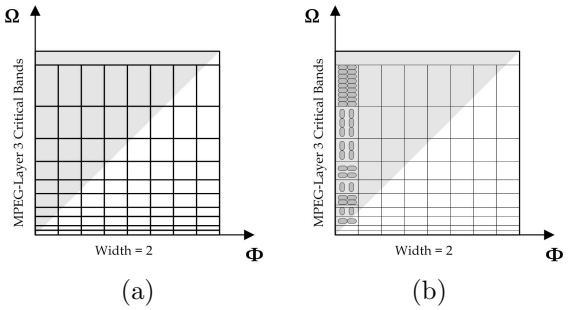


Fig. 11: (a) Spatio-temporal distribution of critical bands: 21 non-uniform bands in Ω; $\frac{1}{2}B_m$ uniform bands in Φ. (b) Pair-wise Huffman coding. Vertical and horizontal pairs are selected according to the one that requires less bits do encode.

The main purpose of the psychoacoustic model, and the matrix $\mathbf{M}$, is to determine the quantization step Δ_{b_n,b_m} required for quantizing each spectral coefficient Y_{b_n,b_m}, so that the quantization noise is lower than M_{b_n,b_m}. If the bitrate decreases, the quantization noise may increase beyond $\mathbf{M}$ to compensate for the reduced number of available bits. In this paper, we assume that $p_{n,m}$ is encoded with maximum quality, which means that the quantization noise is strictly bellow $\mathbf{M}$.

Another way of controlling the quantization noise, which we adopted for the WFC, is by setting $\Delta_{b_n,b_m} = 1$ for all b_n and b_m, and scaling the coefficients Y_{b_n,b_m} by a scale factor SF_{b_n,b_m}, such that $SF_{b_n,b_m}Y_{b_n,b_m}$ falls into the desired integer. In this case, given that the quantization noise power equals $\frac{\Delta^2}{12}$,

$$SF_{b_n,b_m} = \sqrt{12M_{b_n,b_m}}\,. \tag{40}$$

The quantized spectral coefficient $Y^Q_{b_n,b_m}$ is then

$$Y^Q_{b_n,b_m} = \operatorname{sign}\left(Y_{b_n,b_m}\right) \cdot \left\lfloor \left(SF_{b_n,b_m} \cdot |Y_{b_n,b_m}|\right)^{\frac{3}{4}} \right\rfloor , \tag{41}$$

where the factor $\frac{3}{4}$ is used to increase the accuracy at lower amplitudes. Conversely,

$$Y_{b_n,b_m} = \operatorname{sign}\left(Y^Q_{b_n,b_m}\right) \cdot \left(\frac{1}{SF_{b_n,b_m}} \cdot \left|Y^Q_{b_n,b_m}\right|^{\frac{4}{3}}\right) . \tag{42}$$

In state-of-art audio coders, it is not possible to have one scale factor per coefficient. Instead, a scale factor is assigned to one critical band, such that all coefficients within the same critical band are quantized with the same scale factor. In WFC, the critical bands are two-dimensional, and the scale factor matrix $\mathbf{SF}$ is approximated by a piecewise constant surface. The spatio-temporal critical bands are organized as shown in Fig. 11a, with uniform bandwidths in bark scale [8] for the Ω-axis, and uniform bandwidths in linear scale for the Φ-axis. In the Ω-axis, the bandwidths are the ones used in MPEG-Layer 3 [9], whereas in the Φ-axis the critical bands have width 2. The bandwidth distribution of the critical bands along the Φ-axis is still an open subject, and requires further research. Some preliminary results have suggested that the bandwidth can be larger for higher spatial frequencies, just like in the case of temporal frequencies.

4.6. Huffman Coding

After quantization, the spectral coefficients are converted into binary base using entropy coding. A Huffman codebook with a certain range is assigned to each spatio-temporal critical band, and all coefficients in that band are coded with the same codebook.

The use of entropy coding is possible because the MDCT has a different probability of generating certain values. An MDCT occurrence histogram, for different signal samples, clearly shows that small absolute values are more likely than large absolute values, and that most of the values fall within the range of -20 to 20. For this reason, state-of-art audio coders use a predefined set of Huffman codebooks that cover all ranges up to a certain value r. If any coefficient is bigger than r or smaller than $-r$, it is encoded with a fixed number of bits using Pulse Code Modulation (PCM). In addition, adjacent values (Y_{b_n}, Y_{b_n+1}) are coded in pairs, instead of individually. Each Huffman codebook covers all combinations of values from $(Y_{b_n}, Y_{b_n+1}) = (-r, -r)$ up to $(Y_{b_n}, Y_{b_n+1}) = (r, r)$.

In the WFC, a set of 7 Huffman codebooks covering all ranges up to $[-7, 7]$ is generated according to the following probability model. Consider a pair of spectral coefficients $\mathbf{y} = (Y_0, Y_1)$, adjacent in the Ω-axis. For a codebook of range r, we define a probability measure $\mathbb{P}[\mathbf{y}]$ such that

$$\mathbb{P}[\mathbf{y}] = \frac{\mathbb{W}[\mathbf{y}]}{\sum_{Y_0=-r}^{r} \sum_{Y_1=-r}^{r} \mathbb{W}[\mathbf{y}]}, \tag{43}$$

where

$$\mathbb{W}[\mathbf{y}] = \frac{1}{\mathbb{E}[|\mathbf{y}|] + \mathbb{V}[|\mathbf{y}|] + 1}. \tag{44}$$

The weight of $\mathbf{y}$, $\mathbb{W}[\mathbf{y}]$, is inversely proportional to the average $\mathbb{E}[|\mathbf{y}|]$ and the variance $\mathbb{V}[|\mathbf{y}|]$, where $|\mathbf{y}| = (|Y_0|, |Y_1|)$. This comes from the assumption that $\mathbf{y}$ is more likely to have both values Y_0 and Y_1 within a small amplitude range, and that $\mathbf{y}$ has no sharp variations between Y_0 and Y_1.

When performing the actual coding of the spectral block $\mathbf{Y}$, the appropriate Huffman codebook is selected for each critical band according to the maximum amplitude value Y_{b_n,b_m} within that band, which is then represented by r. In addition, the selection of coefficient pairs is performed vertically in the Ω-axis or horizontally in the Φ-axis (see Fig. 11b), according to the one that produces the minimum overall weight $\mathbb{W}[\mathbf{y}]$. Hence, if $\mathbf{v} = (Y_{b_n,b_m}, Y_{b_n+1,b_m})$ is a vertical pair and $\mathbf{h} = (Y_{b_n,b_m}, Y_{b_n,b_m+1})$ is an horizontal pair, then the selection is performed according to

$$\min_{\mathbf{v},\mathbf{h}} \left\{ \sum_{b_n,b_m} \mathbb{W}[\mathbf{v}], \sum_{b_n,b_m} \mathbb{W}[\mathbf{h}] \right\}.$$

If any of the coefficients in $\mathbf{y}$ is greater than 7 in absolute value, the Huffman codebook of range 7 is selected, and the exceeding coefficient Y_{b_n,b_m} is encoded with the sequence corresponding to 7 (or -7 if the value is negative) followed by the PCM code corresponding to the difference $Y_{b_n,b_m} - 7$.

4.7. Bitstream Format

The final step of the WFC is to organize all binary data into a time series of bits, called the *bitstream*, in a way that the decoder can parse the data and use it reconstruct the multichannel signal $p(t, x)$. The basic components of the bitstream are the main header, and the frames that contain the coded spectral data for each block (see Fig. 12). The frames themselves have a small header with side information necessary to decode the spectral data.

The main header is located at the beginning of the bitstream, and contains information about the sampling frequencies Ω_S and Φ_S (see Section 4.2), the window type and the size $B_n \times B_m$ of spatio-temporal MDCT (see Section 4.3), and any parameters that remain fixed for the whole duration of the multichannel audio signal.

The frame format is repeated for each spectral block $\mathbf{Y}_{g,l}$, and organized in the following order:

$$\mathbf{Y}_{0,0} \ldots \mathbf{Y}_{0,K_l-1} \mathbf{Y}_{K_g-1,0} \ldots \mathbf{Y}_{K_g-1,K_l-1},$$

such that, for each time instance, all spatial blocks are consecutive. Each block $\mathbf{Y}_{g,l}$ is encapsulated in a frame, with a header that contains the scale factors used by $\mathbf{Y}_{g,l}$ (see Section 4.5) and the Huffman codebook identifiers (see Section 4.6).

The scale factors can be encoded in logarithmic scale using 5 bits. The number of scale factors depends on the size B_m of the spatial MDCT, and the size of the critical bands. Since the width along the Φ-axis is 2 and the number of bands in the Ω-axis is

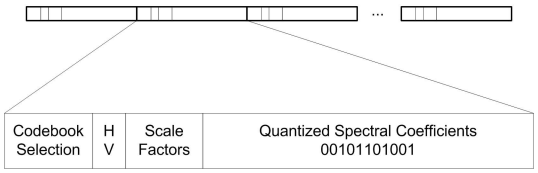


Fig. 12: Bitstream format. Each frame contains coded spectral data plus a header with side information.

21, the number of scale factors is $21 \cdot \frac{B_m}{2}$, and the number of bits required is $\left(21 \cdot \frac{B_m}{2}\right) \cdot 5$. The Huffman codebook selection can be done with 3 bits per band, plus 1 bit for switching between vertical and horizontal coefficient-pair selection. Hence, the side information for Huffman coding requires $\left(21 \cdot \frac{B_m}{2}\right) \cdot 4$ bits per frame. Overall, each frame contains at least $\left(21 \cdot \frac{B_m}{2}\right) \cdot 9$ bits of side information.

4.8. Decoding

The decoding stage of the WFC comprises three steps: decoding, re-scaling, and inverse filterbank. The decoding is controlled by a state machine representing the Huffman codebook assigned to each critical band. Since Huffman encoding generates prefix-free binary sequences, the decoder knows immediately how to parse the coded spectral coefficients. Once the coefficients are decoded, the amplitudes are re-scaled using (42) and the scale factor associated to each critical band. Finally, the inverse MDCT is applied to the spectral blocks, and the recombination of the signal blocks is obtained through overlap-and-add in both temporal and spatial domains.

The decoded multichannel signal $p_{n,m}$ can be interpolated into $p(t, x)$, without loss of information, as long as the anti-aliasing conditions are satisfied (see Section 4.2). The interpolation can be useful when the number of loudspeakers in the playback setup does not match the number of channels in $p_{n,m}$.

5. EXPERIMENTAL RESULTS

To test the WFC scheme in Matlab, we generated a spacetime signal $\mathbf{P}$ in two acoustic scenarios, depicted in Fig.13. In both cases, there is a source in near-field and one in far-field, and the two are uncorrelated. Each source has two mirror reflections with attenuation factor 0.75 and one double reflection with attenuation 0.75^2. In the first acoustic scenario, the listening area is defined by a circle with 96 loudspeakers, whereas in the second one the listening area is the half plane bellow a line of 96 loudspeakers. We also performed a test with $\mathbf{P}$ downsampled in space by a factor of 4.

B_m	**96**	**48**	**24**	**12**
Entropy (bit/sample)	3.0	2.8	2.8	2.7
Data Bitrate (kbit/s/ch)	147	149	149	151
Side Info. Bitrate (kbit/s/ch)	7	7	7	7

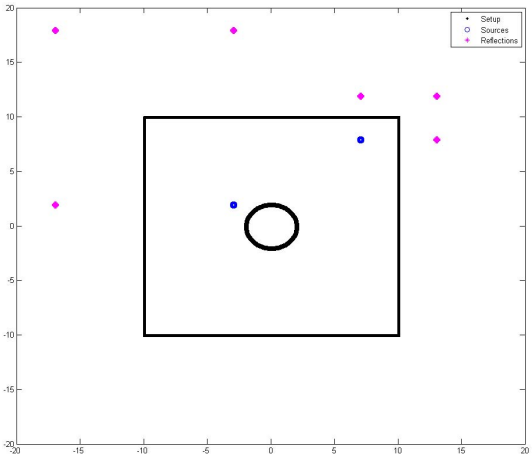
Table 1: Circle setup with 96 channels. Plane-wave method.

B_m	**96**	**48**	**24**	**12**
Entropy (bit/sample)	3.1	2.8	2.7	2.7
Data Bitrate (kbit/s/ch)	144	146	146	150
Side Info. Bitrate (kbit/s/ch)	7	7	7	7

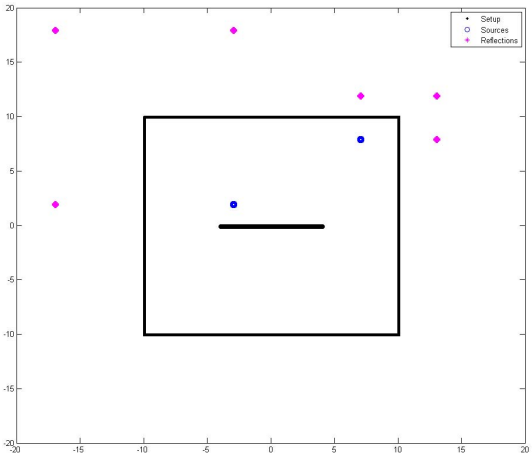
Table 2: Line setup with 96 channels. Plane-wave method.

The sampling frequencies are $\Omega_S = 2\pi \cdot 44100\, s^{-1}$ for both cases, and $\Phi_S = 2\pi \cdot 7.64\, m^{-1}$ for the circle setup and $\Phi_S = 2\pi \cdot 11.9\, m^{-1}$ for the line setup. The spatio-temporal MDCT has parameters $B_n = 576$ and B_M according to the tables. The quantization and Huffman encoding are performed according to Sections 4.5 and 4.6. We show results for two psychoacoustic models: plane-wave based estimation (Tables 1 and 2) and spatial-frequency based estimation (Tables 3 and 4). In the last two tables, the method used is the plane-wave based estimation. After decoding, we informally confirmed that the decoded spacetime signal $\hat{\mathbf{P}}$ had no audible artifacts. The achieved bitrates per channel are shown in the following tables.

The entropies are computed using Shannon's formula, $-\sum p \log_2 p$, where the probabilities p are estimated through an histogram of quantized values $\mathbf{Y}^{\mathbf{Q}}$. The side information bitrate accounts only



(a)



(b)

Fig. 13: Two loudspeaker setups: (a) circle and (b) straight line. In both acoustic scenes, there is one source in near-field and one source in far-field, both with four mirror reflections.

B_m	24	12
Entropy (bit/sample)	3.7	3.7
Data Bitrate (kbit/s/ch)	184	185
Side Info. Bitrate (kbit/s/ch)	7	7

Table 3: Circle setup with 96 channels. Spatial-frequency method.

B_m	24	12
Entropy (bit/sample)	3.7	3.7
Data Bitrate (kbit/s/ch)	180	184
Side Info. Bitrate (kbit/s/ch)	7	7

Table 4: Line setup with 96 channels. Spatial-frequency method.

B_m	24	12
Entropy (bit/sample)	3.0	3.0
Data Bitrate (kbit/s/ch)	150	156
Side Info. Bitrate (kbit/s/ch)	7	7

Table 5: Circle setup with 24 channels. Plane-wave method.

for scale factors, Huffman codebook selection, and horizontal/vertical pair selection, according to Section 4.7.

The results in Tables 1 and 2 show that increasing the spatial resolution slightly decreases the entropy, although after Huffman encoding the bitrate is nearly the same. The importance of processing $\mathbf{P}$ in short-space, specially in the circle configuration, becomes evident when the quantization noise is increased, where we verify that for $B_m = 96$ the artifacts become audible faster than with $B_m = 12$. This is because for $B_m = 96$ the circle can not be approximated by a polygon (see Section 3.3), and therefore the plane-wave decomposition is not valid anymore. The results also suggest that when $\mathbf{P}$ is downsampled in space (less loudspeakers) the bitrate per channel is not significantly affected, although there is less flexibility in selecting B_m.

The results in Tables 3 and 4 show that the spatial-frequency based estimation of the masking surface (equivalent, as mentioned, to DCT-based decorrelation of the channels) is not as optimal as the plane-wave method.

B_m	24	12
Entropy (bit/sample)	3.0	3.0
Data Bitrate (kbit/s/ch)	147	154
Side Info. Bitrate (kbit/s/ch)	7	7

Table 6: Line setup with 24 channels. Plane-wave method.

These results were compared to coding each channel independently, using the exact same masking curve estimation method (the operator $\mathbb{M}[\cdot]$ in Section 4.4). This coding approach required around 175 kbit/s/ch for spectral data and 13 kbit/s/ch for side information. The bitrate reduction achieved by WFC is therefore around 15%. Note, however, that the purpose of this publication is not to exhaustively optimize the coder's design and performance, but only to demonstrate the potential of this new technique. In fact, the bitrate can be further reduced if the spatial masking effect is also exploited, which is not possible when coding the channels independently. Improving the spatio-temporal psychoacoustic model is part of our future work.

6. REFERENCES

[1] A. Berkhout, D. de Vries, and P. Vogel, Wavefront synthesis: A new direction in electroacoustics, in Audio Engineering Society 93th Convention, 1992.

[2] U. Horbach, E. Corteel, R. Pellegrini, and E. Hulsebos, Real-time rendering of dynamic scenes using wave field synthesis, in IEEE International Conference on Multimedia and Expo, 2002, vol. 1, pp. 517 520.

[3] C. Faller, Parametric joint-coding of audio sources, in Audio Engineering Society 120th Convention, 2006.

[4] T. Sporer, J. Plogsties, S. Brix, CARROUSO - An European Approach to 3D-Audio, 110th AES Convention, 2001, pp 5314.

[5] P. Morse, K. Ingard, Theoretical acoustics, Princeton University Press, 1987.

[6] T. Ajdler, L. Sbaiz, and M. Vetterli, The plenacoustic function and its sampling, in IEEE Transactions on Signal Processing, 2006, vol. 54, pp. 3790 3804.

[7] H. Malvar, Signal processing with lapped transforms, Artech House Publishers, 1992.

[8] M. Bosi, R. Goldberg, Introduction to digital audio coding and standards, Springer, 2002.

[9] ISO/IEC, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Part 3: Audio".