CODING OF SPATIO-TEMPORAL AUDIO SPECTRA USING TREE-STRUCTURED DIRECTIONAL FILTERBANKS

Francisco Pinto and Martin Vetterli

Ecole Polytechnique Fédérale de Lausanne
EPFL-IC-LCAV, Station 14, CH-1015 Lausanne, Switzerland
francisco.pinto@epfl.ch; martin.vetterli@epfl.ch

ABSTRACT

We address the problem of integrating directional analysis of sound into the filterbank of a spatial audio coder, with the purpose of processing and coding with some degree of independence the plane waves traveling in different directions. A plane wave represents an elementary waveform in the spatio-temporal analysis of the sound field, the same way a complex exponential is an elementary waveform in the time domain analysis of signals. Since a two-dimensional separable filterbank is not flexible enough for this purpose, we propose a non-separable approach based on the quincunx filterbank with diamond-shaped filters, cascaded with a base transform filterbank. This solution provides an invertible and critically sampled decomposition of the spatio-temporal spectra into subbands representing the different directions of wave propagation.

Index Terms— Spatial audio coding, multidimensional processing, directional filterbanks

1. INTRODUCTION

Spatial audio is the general term used to describe audio data that includes information (implicit or explicit) about the distribution of sound sources in space, in addition to the signals in the time domain, thus providing a better description of the sound field that eventually results in a better immersive experience for all the listeners. A particular class of techniques - most notably wave field synthesis [1] - uses the so-called Huygens principle to replicate an entire sound field within a listening region with no significant sweet-spot restrictions. Concretely, under a few relaxed conditions [1], the theory states that the wavefronts radiated from the outside of an enclosed area $\mathcal{A}$ can alternatively be generated on the inside by a number of secondary point sources located at the boundary contour $\mathcal{L}$, and weighted by the field values at their respective positions.

This project is funded by the *Fundação para a Ciência e Tecnologia* (Portugal) and the *National Center of Competence in Research for Mobile Information and Communication Systems* (Switzerland).

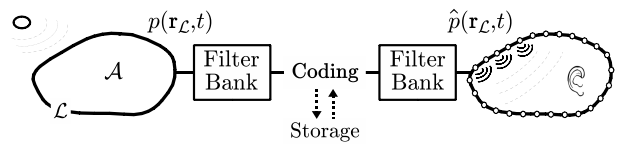

Figure 1: Basic structure of a spatio-temporal wave field coder, where the input signal is a function of time and the spatial coordinates of $\mathcal{L}$, given by $\mathbf{r}_{\mathcal{L}} = (x_{\mathcal{L}}, y_{\mathcal{L}})$, and the output is a processed or decoded version.

The Huygens principle can be conveniently formulated as a digital signal processing problem (Fig. 1), where the input signal is the boundary pressure $p(\mathbf{r}_{\mathcal{L}}, t)$ taken at any given region in space, and the output signal is a modified version $\hat{p}(\mathbf{r}_{\mathcal{L}}, t)$ that can eventually be used to reconstruct the sound field at a different region in space. This opens up the possibility for processing, coding, and storing the wavefront before it can reach the ear of the listeners.

In a previous work [2], we have introduced a new spatial audio coding technique based on plane-wave encoding using perceptual criteria. The coder operates by performing a local spatial analysis around $\mathcal{L}$ and applying a spatio-temporal Modified Discrete Cosine Transform (MDCT) filterbank, which translates the plane-wave content of the sound field into a sparse frequency-based representation. In this paper, we further improve the compressive potential of the MDCT filterbank by combining it with a directional filterbank [3] that is capable of breaking the sound field into multiple directional components, such that each component falls into a different frequency subband representing a small range of directions of wave propagation. The result is a non-parametric representation of the sound field that closely matches a parametric model, with the additional attribute of providing a critically sampled and perfectly reconstructible output, suitable for audio coding.

2. SPACETIME DOMAIN REPRESENTATION

2.1. Short-space analysis of the sound field

Let $p(x, t)$ be the sound pressure generated at the x-axis by a point source driven by $s(t)$, and $P(\Phi, \Omega)$ the respective

spatio-temporal Fourier transform such that

$$P(\Phi,\Omega) = \int\int_{-\infty}^{\infty} p(x,t)e^{-j(\Phi x+\Omega t)}dtdx. \quad (1)$$

Depending on whether the source is located in the near-field (NF) or in the far-field (FF), the results (without fixed amplitude factors) are given by [4]

$$P_{\text{nf}}(\Phi,\Omega) = S(\Omega)H_o^{(1)*}\left(y_o\sqrt{\left(\frac{\Omega}{c}\right)^2 - \Phi^2}\right)e^{-jx_o\Phi} \quad (2)$$

$$P_{\text{ff}}(\Phi,\Omega) = S(\Omega)\delta\left(\Phi - \cos\alpha\frac{\Omega}{c}\right), \quad (3)$$

where Φ and Ω are the spatial and temporal frequencies, $S(\Omega)$ is the Fourier transform of $s(t)$, $H_o^{(1)}$ is the zeroth-order Hankel function of the first kind, (x_o, y_o) is the position of the near-field source, α is the angle of incidence of a far-field wavefront in the x-axis, and c is the speed of sound. The result in (2) represents a triangular spectral pattern where most of the energy is distributed across the region $|\Phi| \leq \left|\frac{\Omega}{c}\right|$, whereas in (3) the whole energy is concentrated in a single Dirac line defined by $\Phi = \cos\alpha\frac{\Omega}{c}$.

If the Fourier transform is taken over a short spatial window, $P(\Phi,\Omega)$ varies smoothly between (2) and (3) as the source gets closer and farther away from the windowed region, with the additional ripple effect caused by the window function [5]. From a coding standpoint, it is advantageous to perform a short-space decomposition of $p(\mathbf{r}_\mathcal{L}, t)$ in order to exploit the local curvature of the sound field, considering the fact that (3) is easier and more efficient to code than (2) [2, 5]. This, in addition to a short-time analysis of $p(\mathbf{r}_\mathcal{L}, t)$, generates a cylindrical manifold as depicted in Fig. 2, where the x-axis of each spatio-temporal block represents the tangent to the respective boundary region. As the figure shows, the spectral blocks that are closer to the source have a near-field spectral pattern, whereas the ones farther away have more far-field characteristics.

2.2. Far-field components

A particular theoretical aspect with important implications in spatio-temporal analysis is that $p(\mathbf{r}_\mathcal{L}, t)$ can be decomposed into a linear combination of plane waves and evanescent waves traveling in all directions, where the later have only residual energy [6]. This implies that $p(\mathbf{r}_\mathcal{L}, t)$ is mostly composed of plane waves.

In the frequency domain, a plane wave corresponds to a single point within the triangular region $|\Phi| \leq \left|\frac{\Omega}{c}\right|$, whereas evanescent waves are on the outside of this region. This means that, according to (3), a plane wave taken at

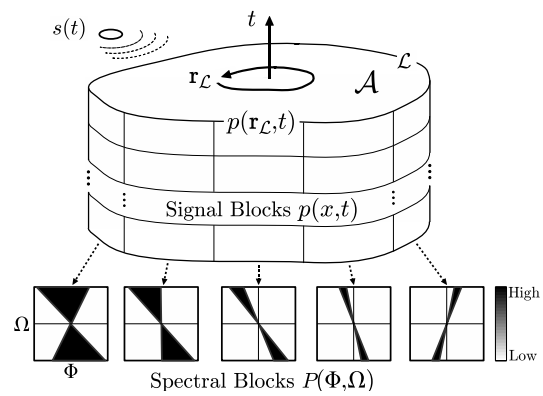

Figure 2: Local spatial analysis of the sound field. The cylindrical manifold is split up into multiple spatio-temporal blocks, each with a different spectral pattern and potential coding gain.

$(\Phi,\Omega) = (\cos\alpha\frac{\Omega_o}{c}, \Omega_o)$ corresponds to a far-field wavefront generated by $s(t) = e^{j\Omega_o t}$ and hitting the x-axis with angle α. Since $s(t)$ itself can be expressed as a linear combination of complex exponentials, $P(\Phi,\Omega)$ is equivalently a linear combination of far-field components:

$$P(\Phi,\Omega) \approx \int_0^\pi S(\alpha,\Omega)\delta\left(\Phi - \cos\alpha\frac{\Omega}{c}\right)d\alpha\,. \quad (4)$$

where $S(\alpha,\Omega)$ represents all the spectral content arriving from direction α. Conversely, $S(\alpha,\Omega)$ can be approximately recovered from $P(\Phi,\Omega)$ by integrating both sides of (4), such that, for any given α,

$$\lim_{\Delta\to 0^+}\int_{\cos\left(\alpha+\frac{\Delta}{2}\right)\frac{\Omega}{c}}^{\cos\left(\alpha-\frac{\Delta}{2}\right)\frac{\Omega}{c}} P(\Phi,\Omega)d\Phi \approx S(\alpha,\Omega). \quad (5)$$

From this perspective, the spectral pattern generated by a near-field source can be intuitively understood as a high concentration of far-field components around an average angle of incidence α. This result is the basis for directional analysis of the sound field, introduced in the next section.

3. DIRECTIONAL SUBBAND DECOMPOSITION

Let $p[\mathbf{n}] = p[n_x, n_t]$ represent the uniformly sampled version of $p(x,t)$ such that $p[\mathbf{n}] = p\left(n_x\frac{2\pi}{\Phi_S}, n_t\frac{2\pi}{\Omega_S}\right)$, where $\Phi_S = 2\frac{\Omega_{\max}}{c}$ and $\Omega_S = 2\Omega_{\max}$ are the spatial and temporal sampling frequencies that strictly satisfy the Nyquist conditions. Let also $P[\mathbf{b}] = P[b_x, b_t]$ denote the respective transform-domain coefficients. The design of a transform filterbank that operates in the spacetime domain is characterized by a four-dimensional tiling that spans all the variables n_x, n_t, b_x, and b_t. In the case of Fourier-based transforms, such as the Modified Discrete Cosine Transform

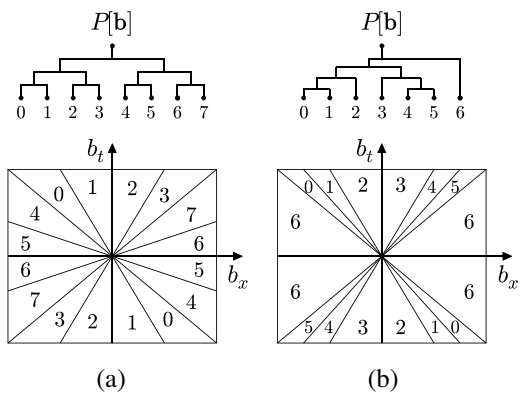

Figure 3: Decomposition of the spatio-temporal spectra into directional subbands using a tree-structured filterbank. The subband partitioning can be either (a) uniform or (b) non-uniform. A partitioning similar to (b) is more suitable for sound field spectral patterns, since it gathers up all the evanescent waves into a single subband and has variable resolution in the plane-wave region.

(MDCT), the tiling is separable and uniform. For directional analysis, however, the filterbank must be designed according to a non-separable decomposition of the spectrum.

The spatio-temporal MDCT was developed in a previous work [2, 5] as the filterbank of a perceptual wave field coder, which operates by encoding plane waves in the MDCT domain. One of the limitations is that, even though plane waves are sparsely represented in the MDCT domain, the filterbank lacks an important property described in the previous section: the directional analysis of sound. In other words, the filterbank must be capable of representing elementary waveforms (far-field components) as efficiently as possible. For this reason, we propose a hybrid extension of the MDCT filterbank that further decomposes the spatio-temporal spectra into directional subbands representing the spectral content arriving from each different direction.

The goal here is to translate the result in (5) into a discrete-domain architecture that can be coupled to the output of a spatio-temporal MDCT filterbank. Conceptually, one can think of a filterbank that decomposes the spectrum into directional subbands defined in the range $\cos\left(\alpha_o \pm \frac{\Delta}{2}\right)\frac{\Omega}{c} \leq \Phi \leq \cos\left(\alpha_o \mp \frac{\Delta}{2}\right)\frac{\Omega}{c}$, for $\Omega = \pm|\Omega|$, where α_o is the central direction and Δ is the directional bandwidth, as illustrated in Fig. 3-a. Additionally, if we assume that each pair of subbands is obtained by slicing a larger band in half, the filterbank can be designed as an iterated 2-channel structure, providing more flexibility to obtain non-uniform directional decompositions such as the one in Fig. 3-b. This way, the biggest effort goes into designing a 2-channel filterbank that can be used in all nodes of the tree.

Such a filterbank has been extensively studied [7, 8, 3], and is known as Quincunx Filterbank (QFB). The QFB is a non-separable perfect reconstruction filterbank defined by

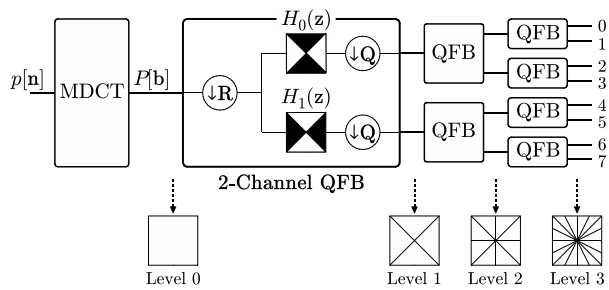

Figure 4: Cascade of MDCT filterbank and iterated Quincunx filterbank. The matrices $\mathbf{R}$ and $\mathbf{Q}$ represent a parallelogram resampler followed by a quincunx resampler.

two diamond-shaped half-band filters, $H_0(\mathbf{z})$ and $H_1(\mathbf{z})$, preceded by a parallelogram resampler $\mathbf{R}$ and followed by a quincunx resampler $\mathbf{Q}$ [9], as illustrated in Fig. 4. The seven variants of $\mathbf{R}$ and $\mathbf{Q}$ are given by

$$\mathbf{Q}_0 = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \mathbf{Q}_1 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \mathbf{R}_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$
$$\mathbf{R}_1 = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \mathbf{R}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \mathbf{R}_3 = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \tag{6}$$

plus the identity matrix $\mathbf{R} = \mathbf{I}$, and have resampling densities given by $|\det(\mathbf{R})| = 1$ and $|\det(\mathbf{Q})| = 2$. The selection among these matrices depends on the tree level and branch considered [3]. The output of both channels is maximally decimated due to a global resampling density of $|\det(\mathbf{RQ})| = 2$.

The intuition behind the iterated QFB structure is that, instead of using different filters to obtain different subbands, we use the resampling matrices to rotate and skew the subbands before slicing them with a fixed filter.

The design of the half-band filters $H_0(\mathbf{z})$ and $H_1(\mathbf{z})$, and the respective synthesis filters $F_0(\mathbf{z})$ and $F_1(\mathbf{z})$, can be done using two-dimensional filter design techniques. Notably, an efficient structure by Phoong *et al* [8] reduces the procedure to the design of an FIR all-pass filter $A(z) = z^{-\frac{1}{2}(2M-1)}$ of order M, such that

$$H_0(\mathbf{z}) = \frac{1}{2}z_x^{-2M} + \frac{1}{2}z_x^{-1}A(z_x z_t^{-1})A(z_x z_t) \tag{7a}$$
$$H_1(\mathbf{z}) = -A(z_x z_t^{-1})A(z_x z_t)H_0(\mathbf{z}) + z_x^{-4M+1} \tag{7b}$$
$$F_0(\mathbf{z}) = -H_1(-\mathbf{z}) \tag{7c}$$
$$F_1(\mathbf{z}) = H_0(-\mathbf{z}), \tag{7d}$$

where $\mathbf{z} = (z_x, z_t)$ represents the z-transform variables. The design of the all-pass filter $A(z)$ controls the frequency selectivity of the half-band filters, which has an influence in the final amount of cross-band artifacts.

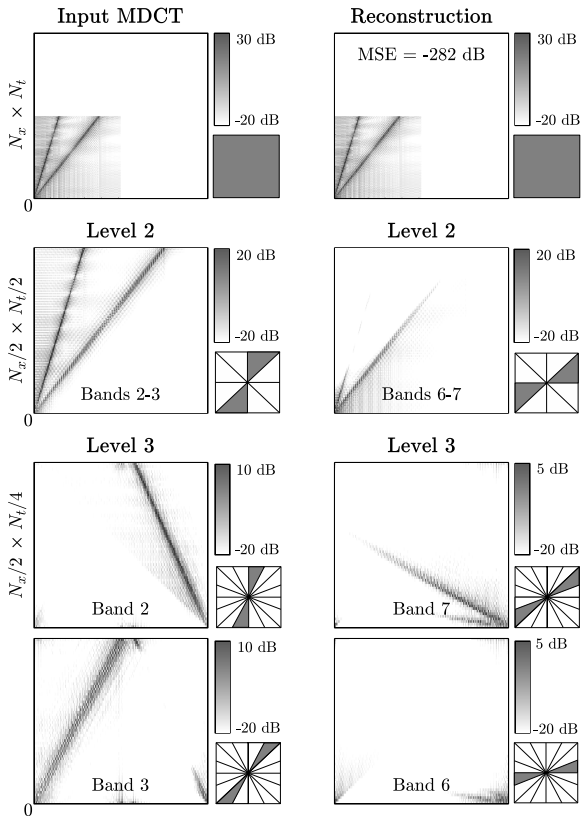

Figure 5: Three-level directional subband decomposition of an MDCT block of size $N_x \times N_t$ (with negative frequencies set to zero), using the uniform partitioning of Fig. 3-a. Up to Level 2, the two far-field components are separated from the remaining spectrum, whereas at Level 3 they are separated from each other. The spectral rotation is caused by the non-diagonal resampling matrices.

4. SIMULATION RESULTS

Fig. 5 shows an example of directional subband decomposition of an input MDCT block $P[\mathbf{b}]$ generated by two wide-band sources in the far-field. As expected, the far-field components are well separated by the directional filterbank, each falling into a different subband, and then recombined into the original spectrum with negligible error. However, since the filters are non-ideal, some of the energy is leaked across subbands - specially at lower frequencies. These artifacts are expected to increase with the number of directional subbands, but can be controlled to a certain degree by varying the resolution of the filters in the design phase.

Most importantly, the result shows a concentration of the directional energy into a limited number of subbands - two, in this case - as was initially intended. These bands carry the most relevant information for the subsequent steps in the wave field coding process, particularly in the quantization of $P[\mathbf{b}]$ based on directional psychoacoustic effects associated to the various sources in the scene [5].

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a hybrid filterbank architecture that is able to decompose a non-parametric representation of the sound field into elementary and sparse directional components, while preserving perfect reconstruction and critical sampling of the input data - an important requirement for audio coding. Under ideal filtering conditions, the filterbank output approximates a parametric model as the number of directional subbands increases, where the parameters are the source signal $s(t)$ and the direction α of every far-field component in the acoustic scene.

Further work will include the integration of the hybrid filterbank into a non-parametric wave field coder already in development. Optimization issues will also be taken into account; namely: (i) how to adapt or redesign the half-band diamond filters for optimal operation in the MDCT domain, thus avoiding redundant computations for negative frequencies, and (ii) how to compensate for cross-band artifacts generated by spatial windowing and the finite resolution of the half-band filters.

6. REFERENCES

[1] A. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journal of the Acoustical Society of America*, vol. 93, pp. 2764–2778, 1993.

[2] F. Pinto and M. Vetterli, "Wave field coding in the spacetime frequency domain," in *IEEE Inter. Conf. Acoustics, Speech and Signal Processing*, 2008.

[3] M. Do, "Directional multiresolution image representations," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, 2001.

[4] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Trans. Signal Processing*, vol. 54, pp. 3790–3804, 2006.

[5] F. Pinto and M. Vetterli, "Bitstream format for spatio-temporal wave field coder," in *Audio Engineering Society 124th Convention*, 2008.

[6] E. Williams, *Fourier Acoustics*. Academic Press, 1999.

[7] R. Bamberger and M. Smith, "A filter bank for the directional decomposition of images: Theory and design," in *IEEE Trans. Signal Processing*, 1992.

[8] S. Phoong, C. Kim, P. Vaidyanathan, and R. Ansari, "A new class of two-channel biorthogonal filter banks and wavelet bases," in *IEEE Trans. Signal Processing*, 1995.

[9] P. Vaidyanathan, *Multirate Systems And Filter Banks*. Prentice Hall, 1992.