WAVE FIELD CODING IN THE SPACETIME FREQUENCY DOMAIN

Francisco Pinto and Martin Vetterli

Ecole Polytechnique Fédérale de Lausanne
francisco.pinto@epfl.ch; martin.vetterli@epfl.ch

ABSTRACT

We present a new method for compressing spatio-temporal audio data for reproduction through Wave Field Synthesis. The data is obtained by sampling the sound field in space at equally-spaced points on a straight line, and transformed into the frequency domain using a spatio-temporal lapped transform. The two-dimensional spectrum is quantized using a psychoacoustic model derived for spatio-temporal frequencies, which estimates the maximum quantization noise power that each frequency can support in order to preserve transparency in the decoded signal. On the decoder side, the inverse lapped transform recovers the spatio-temporal data. In our experimental results, we verified that the bitrate-efficiency can be improved by increasing either the spatial sampling frequency or the spatial resolution of the lapped transform.

Index Terms— Wave field synthesis, spatial audio, perceptual audio coding, frequency masking

1. INTRODUCTION

Reproduction of audio through Wave Field Synthesis (WFS) has gained considerable attention since it was first introduced by Berkhout [1]. One of the main reasons is the potential for reproducing a sound 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.

Most of the research related to WFS is focused essentially on acoustic theory and rendering algorithms, whereas less attention is given to the development of efficient coding techniques. Whenever compression is required, the usual choice is to code each source signal separately using a conventional mono coder, plus side information representing the source positions in space and other spatial cues [2]. The source signals can also be jointly coded in order to reduce the required bitrate [3]. These parametric approaches, however, assume that all source signals in the acoustic scene are separately available, which may not be the case in a complex acoustic scene where there are too many sources to record, or simply when there is a market-related decision of keeping the source signals confidential, like record companies do for commercializing CD-Audio. Furthermore, even if the source signals are available, the decoded signals and parameters must pass through a complex (and possibly confidential) real-time rendering algorithm for reproducing the sound field.

In this paper, we present an alternative coding scheme that is non-parametric, has low complexity (relies on basic signal processing operations), and does not require any source signal or rendering algorithm for reproduction. The proposed scheme exploits a particular WFS configuration, that consists of sampling the sound field in space on equidistant points along a straight line [4] and reconstructing it back with a line-array of loudspeakers. The spatio-temporal samples form a two-dimensional signal that we call *spacetime signal*, and to which we can apply the basic sampling theory (see Section 2 and 3). The encoding method consists of transforming the signal 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. On the decoder side, a spatio-temporal inverse transform recovers the spacetime signal (see Section 3). 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 a spacetime signal generated by one point source in near-field, and estimating the required bitrate for preserving transparency[1]. The results indicate that the spatial sampling frequency and the spatial resolution are key factors for improving the bitrate-efficiency (see Section 4).

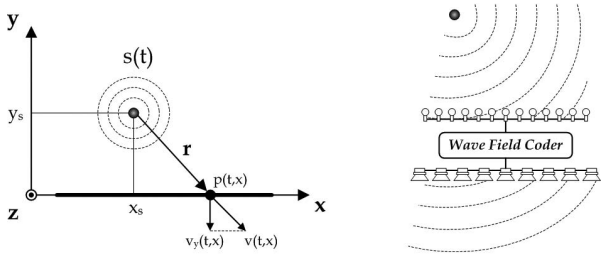
2. SPACETIME SIGNAL ANALYSIS

2.1. Reproduction through WFS

Let $p(t, \mathbf{r})$ be a sound pressure wave on the horizontal xy-plane, generated by a point source located at (x_s, y_s) and

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).

[1] We assume that the encoding/decoding operation is transparent if the decoded spacetime signal has no audible artifacts, whether we listen to the channels jointly (to replicate the sound field) or separately.



(a) (b)

Fig. 1. (a) Point source on the xy-plane; (b) Application example: sound field sampling with microphone array, intermediate coding, and reproduction with loudspeaker array.

driven by the signal $s\,(t)$, where $\mathbf{r}$ is the vector that connects (x_s, y_s) to the point in space where the pressure is measured (see Fig. 1a). From the theory of acoustic wave propagation,

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

where c is the speed of sound. Suppose that the point source is always located at $y > 0$, and we wish to replicate $p\,(t, \mathbf{r})$ at $y < 0$ without knowing $s\,(t)$, as in the example of Fig. 1b. The theory behind WFS [1] states that $p\,(t, \mathbf{r})$ can, in fact, be reproduced at $y < 0$ by knowing only the y-projection of the particle velocity, $v_y\,(t, \mathbf{r})$, at $y = 0$ (the x-axis). Once $v_y\,(t, x)$ is measured, $p\,(t, \mathbf{r})$ can be replicated using a line source (or an infinite number of point sources) placed on the x-axis and driven by a signal that depends on $v_y\,(t, x)$. Conversely, we can also measure $p\,(t, x)$ and reproduce $p\,(t, \mathbf{r})$ using a directive line source placed on the x-axis. The reason for measuring the particle velocity instead of the pressure is that it allows the use of regular omnidirectional loudspeakers, instead of directive loudspeakers[2], for reproduction. Nevertheless, since the WFC approach is equally valid for both types of signals, the analysis is focused on the sound pressure. We call $p\,(t, x)$ the *continuous-spacetime signal*.

2.2. Frequency representation

A spacetime signal $p\,(t, x)$ can be represented as a linear combination of complex exponentials with temporal frequency Ω and spatial frequency Φ. The spatio-temporal Fourier transform is defined by

$$P\,(\Omega, \Phi) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p\,(t, x)\, e^{-j(\Omega t + \Phi x)} dt dx\,, \qquad (2)$$

where $P\,(\Omega, \Phi)$ is the *continuous-spacetime spectrum.* It can be shown [4] 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. The energy can be either spread over the whole region, which happens when the curvature of the sound field is very stressed (near-field), or reduced to a single line, which happens when the sound field has no curvature (far-field). These details are further developed in Section 3.3.

[2]The particle velocity can be measured with directive microphones, which are more common than directive loudspeakers.

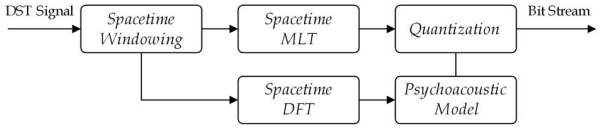


Fig. 2. Block diagram of the wave field encoder.

3. WAVE FIELD CODING

The WFC scheme can be interpreted as a spatio-temporal extension of a traditional perceptual mono coder, where the input spacetime signal is converted into an encoded bit stream. In this section, the encoding steps illustrated in the block diagram of Fig. 2 are described in detail.

3.1. Sampling

In practice, $p\,(t, x)$ can only be measured on discrete points along the x-axis, which, of course, affects the quality of the reconstruction. To obtain a *discrete-spacetime signal* (DST signal), we consider two possible scenarios: (i) in case $s\,(t)$ is available, we compute $p\,(t, x)$ mathematically on equidistant points along the x-axis, using (1); (ii) in case $s\,(t)$ is not available, we physically measure $p\,(t, x)$ on the x-axis using microphones (as shown in Fig. 1b). In either case, the goal is to encode only the spacetime signal $p\,(t, x)$; there is no need to store or even know $s\,(t)$. This is one advantage of WFC.

3.2. Spacetime-frequency mapping

After sampling, the input spacetime signal is transformed into the spatio-temporal frequency domain by applying a spatio-temporal block transform. For simplicity, we assume that the transformation in the spacetime domain 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 us represent the input discrete signal $p_{n,m}$ as a matrix $\mathbf{P}$ of size $N \times M$[3], where n and m are the temporal and spatial sample indexes. Also, let $\tilde{\mathbf{\Psi}}$ and $\tilde{\mathbf{\Upsilon}}$ be two generic transformation matrices for generating the temporal and spatio-temporal spectral matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. The matrix operations that define the spacetime-frequency mapping can be organized as shown in the following table.

[3] N and M are the total number of temporal and spatial samples. If we use a microphone array to sample the sound field, then M is the number of microphones.

Domain:	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 $N \times M$. By combining all transformation steps in the table, one gets $\hat{\mathbf{P}} = \tilde{\mathbf{\Psi}}\tilde{\mathbf{\Psi}}^{\mathbf{T}} \cdot \mathbf{P} \cdot \tilde{\mathbf{\Upsilon}}\tilde{\mathbf{\Upsilon}}^{\mathbf{T}}$. Therefore, it is clear that 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 Modulated Lapped Transform[4] (MLT) [5], which is applied to both temporal and spatial dimensions. The MLT allows 50% of overlap between adjacent windows and still generates a critically sampled spectrum. The transformation matrix $\tilde{\mathbf{\Psi}}$ (or $\tilde{\mathbf{\Upsilon}}$ for space) is defined by

$$\tilde{\mathbf{\Psi}} = \begin{bmatrix} \mathbf{\Psi_1} & & \\ \mathbf{\Psi_0} & \mathbf{\Psi_1} & \\ & \mathbf{\Psi_0} & \ddots \\ & & \ddots \end{bmatrix}, \quad (3)$$

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[5] of the transpose of the basis matrix $\mathbf{\Psi}$, which is given by

$$\psi_{b,2B-1-n} = w_n \sqrt{\frac{2}{B}} \cos\left[\frac{\pi}{B}\left(n + \frac{B+1}{2}\right)\left(b + \frac{1}{2}\right)\right], \quad (4)$$

$$b = 0, 1, \ldots, B-1 \; ; \;\; n = 0, 1, \ldots, 2B-1 \,,$$

where n (or m) is the signal sample index, b is the frequency band index, B is the number of spectral samples in each block, and w_n is the window sequence. For perfect reconstruction, the window sequence must satisfy the Princen-Bradley conditions [5], $w_n = w_{2B-1-n}$ and $w_n^2 + w_{n+B}^2 = 1$.

3.3. Short-spacetime analysis

By dividing the input spacetime signal into smaller spatio-temporal blocks, the MLT is performing both *short-time* and *short-space* analysis. The goal of short-space analysis is to exploit variations in the curvature of the sound field along the spatial sampling axis. The size of the spatial window w_m determines the spatial resolution. As explained in Section 2.2, the spectral energy of a spacetime signal with near-field characteristics is more spread over the triangular region (more information to code), whereas with far-field characteristics it is more concentrated on a single line (less information to code).

[4] Also known as the Modified Discrete Cosine Transform (MDCT).

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

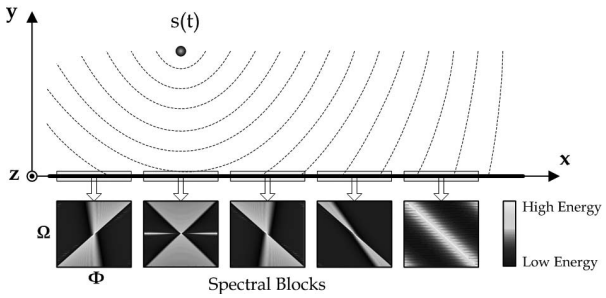


Fig. 3. Short-space analysis of the sound field.

Thus, by performing a localized analysis on the x-axis, as illustrated in Fig. 3, the far-field partitions of the sound field can be isolated from the near-field ones, and their spectrum coded independently, resulting in an increased coding gain.

3.4. Psychoacoustic model

The theory of wave propagation states that any sound field can be decomposed into a linear combination of plane waves and evanescent waves travelling 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 [4]. Since the energy outside the triangle is residual, we can discard evanescent waves and represent the sound field solely by a linear combination of plane waves, which have the elegant property described next.

The spacetime signal generated by a plane wave is given by $p_\alpha(t,x) = s_\alpha\left(t + \frac{\cos\alpha}{c}x\right)$, where $s_\alpha(t)$ is the far-field source signal that produces the plane wave with angle of arrival α. By applying (2), the spacetime spectrum can be shown to be $P_\alpha(\Omega,\Phi) = S_\alpha(\Omega)\,\delta\left(\Phi - \frac{\cos\alpha}{c}\Omega\right)$, where $S_\alpha(\Omega)$ is the one-dimensional spectrum of the source signal $s_\alpha(t)$. If we sum $p_\alpha(t,x)$ and $P_\alpha(\Omega,\Phi)$, respectively, over all possible angles of arrival α, the results are $p(t,x)$ and $P(\Omega,\Phi)$.

Using the plane wave decomposition, we derive the psychoacoustic model for spatio-temporal frequencies by estimating and adding the masking curves produced by all far-field components $s_\alpha(t)$ in the sound field. The masking surface is then $M(\Omega,\Phi) = \sum_\alpha M_\alpha(\Omega)\,\delta\left(\Phi - \frac{\cos\alpha}{c}\Omega\right)$, where $M_\alpha(\Omega)$ is the masking curve generated by $S_\alpha(\Omega)$. In this model, we discard spatial masking effects by assuming total separation of the plane waves by the auditory system.

3.5. Bitrate estimation

The minimum bit density required to encode each spectral block can be estimated by the perceptual entropy [6]. For a spectral block of size $B_N \times B_M$, the entropy is given by

$$H_{g,l} = \frac{1}{B_N B_M} \sum_{u=0}^{B_N-1} \sum_{v=0}^{B_M-1} \log_2\left(1 + \sqrt{SMR_{u,v}}\right), \quad (5)$$

$$g = 0, 1, \ldots, K_N - 1 \; ; \;\; l = 0, 1, \ldots, K_M - 1 \, ,$$

where g and l are the block indexes, K_N and K_M are the total number of temporal and spatial blocks, and $\mathbf{SMR}$ is the *signal-to-mask ratio* matrix. The bitrate is then given by

$$\text{Bitrate} = \frac{\Omega_S}{2\pi N} \sum_{g=0}^{K_N-1} \sum_{l=0}^{K_M-1} B_N B_M H_{g,l} \, , \qquad (6)$$

where Ω_S is the temporal sampling frequency, $B_N B_M H_{g,l}$ is the minimum number of bits required to code all spectral samples in block (g, l), and $\frac{2\pi N}{\Omega_S}$ is the total signal length in seconds. Additionally, we compute the average bitrate per channel, $\frac{\text{Bitrate}}{M}$, which provides a point of comparison between wave field coding and separate channel coding[6].

4. EXPERIMENTAL RESULTS

To test the WFC scheme in Matlab, we generated a spacetime signal $\mathbf{P}$ (see Fig. 4a) by placing one point source near the array on the left side, similarly to the example of Fig. 3, and using a 2s-long music sequence as the source signal $s\,(t)$. We applied the spatio-temporal MLT with $B_N = 512$ and $B_M = 8$ to the signal $\mathbf{P}$, and added random quantization noise to the spacetime spectrum $\mathbf{Y}$ based on the perceptual model described in Section 3.4. We considered three different cases, in which the width of the array was maintained, but the number of channels modified. We also informally confirmed that the decoded spacetime signal $\hat{\mathbf{P}}$ had no audible artifacts. The results were the following.

Number of Channels M	**24**	**48**	**96**
Temporal Sampl. Freq. (s^{-1})	44100	44100	44100
Spatial Sampl. Freq. (m^{-1})	6.7	13.3	26.7
Bitrate (Kbit/s)	1994	3224	5894
Bitrate / Channel (Kbit/s)	83	67	61

These results clearly show that, for a fixed array width, the bitrate-efficiency of WFC increases with the number of channels in the array. This happens mainly because of the higher density of spatial samples on the x-axis, which allows us to better exploit the local curvature of the sound field by increasing the spatial resolution of the MLT, as explained in Section 3.3. As Fig. 4b shows, the locations on the array where the curvature is more stressed have a higher perceptual entropy, whereas locations with less curvature have lower perceptual entropy. Another reason for these results is that the increased spatial sampling frequency results in a spectrum with less dispersed energy [4], and thus less information to code.

[6] In separate channel coding, we compress each spatial channel, $p_{n,0}$, $p_{n,1}$,..., $p_{n,M}$, independently.

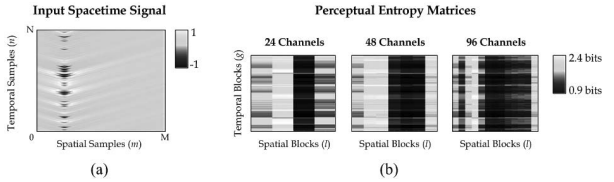


Fig. 4. (a) Spacetime signal; (b) Perceptual entropy matrices.

We compared these results with the ones produced by separate channel coding (SCC), where each channel is coded separately as a mono signal, and using the exact same psychoacoustic model applied to $S_\alpha\,(\Omega)$ in the WFC scheme (see Section 3.4). For SCC, we obtained an average bitrate per channel of 80Kbit/s, independently of M. Looking at the table, we conclude that WFC is at least as efficient as SCC for a small number of channels, and increasingly efficient as the number of channels goes up.

It is important to mention that the masking surface obtained with the method of Section 3.4 is highly underestimated, since only temporal-frequency masking is considered. The bitrate can be reduced if the spatial masking effect over α is also exploited (which is not possible in SCC). The most optimal approach, however, would be to analyze the masking effect produced by each individual spatio-temporal frequency (Ω, Φ), and combine all contributions into a non-separable spatio-temporal masking surface. Studying these approaches is part of our future work.

5. REFERENCES

[1] A. Berkhout, "Wave-front 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. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," in *IEEE Transactions on Signal Processing*, 2006, vol. 54, pp. 3790–3804.

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

[6] J. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *International Conference on Acoustics, Speech, and Signal Processing*, 1988, vol. 5, pp. 2524–2527.