ANALYSIS/SYNTHESIS AND SPATIALIZATION OF NOISY ENVIRONMENTAL SOUNDS

Charles Verron[1,2]*, Mitsuko Aramaki*[3]*, Richard Kronland-Martinet*[2]*, Grégory Pallone*[1]

[1] Orange Labs, 3D audio technologies, 22307 Lannion, France
[2] CNRS, Laboratoire de Mécanique et d'Acoustique, 13402 Marseille, France
[3] CNRS, Institut de Neurosciences Cognitives de la Méditerranée, 13402 Marseille, France
charles.verron@orange-ftgroup.com

ABSTRACT

The use of stochastic modeling is discussed for analysis/synthesis and transformation of environmental sounds. The method leads to perceptually relevant synthetic sounds based on the analysis of natural sounds. Applications are presented, such as sound effects using parametric signal transformations, or data compression. Moreover, we propose a method which efficiently combines the stochastic modeling with 3D audio techniques. This architecture offers an efficient control of the source width rendering that is often an important attribute of noisy environmental sounds. This control is of great interest for virtual reality applications to create immersive 3D scenes.

1. INTRODUCTION

The class of environmental sounds covers a large variety of sounds since it relates to all events occurring in listener's surroundings. A taxonomy of environmental sounds is proposed in [1] [2]. Three main categories are supported by the physics of the sound-producing events: vibrating solids, aerodynamic and liquid sounds. Real-time generation of such sounds for virtual environments and video games is still a challenge. Most of the studies have focused on synthesizing the sound of vibrating solids with physically-based modal resonance modeling [3] [4] [5]. Physically-based models are proposed in [6] [7] [8] for liquid and aerodynamic sounds, and "had-hoc" synthesis of sea waves and wind sounds is described in [9]. A wavelet approach is presented in [10] for analysis and synthesis of noisy environmental sounds. A comprehensive review of environmental sound synthesis can be found in [11] [12].

In this paper we investigate the modeling of noisy environmental sounds with a purely stochastic model. Our approach is based on techniques usually dedicated to the residual of the sinusoid plus noise model [13] described in the first section. The advantages of stochastic modeling are discussed for noisy environmental sounds, such as sea wave, wind and air swishing ("whoosh") sounds. Then spatialization techniques (in particular sound positioning and source width rendering techniques) are described. Finally we propose an efficient method that simulates the spatial extension of noisy environmental sounds.

2. GENERAL CONTEXT: DETERMINISTIC PLUS STOCHASTIC MODEL

In [13] the authors present an analysis/transformation/synthesis system based on a deterministic plus stochastic modeling of a monophonic sound $x(t)$:

$$x(t) \simeq d(t) + s(t)$$

The deterministic part $d(t)$ is composed by $M(t)$ sinusoids whose instantaneous amplitude $a_m(t)$ and frequency $f_m(t)$ vary slowly in time:

$$d(t) = \sum_{m=1}^{M(t)} a_m(t) cos\left(\int_0^t 2\pi f_m(\tau) \mathrm{d}\tau + \Phi_m \right)$$

while the stochastic part $s(t)$ is a time-varying colored noise, i.e., the output of a "time-varying filter" with a white noise input signal. This deterministic plus stochastic modeling (also called sinusoid plus noise model) has been extensively used for analysis, transformation and synthesis of musical sounds. Environmental sounds are also efficiently modeled with this approach.

The synthesis parameters are determined from the analysis of natural sounds. For the deterministic contribution, the analysis consists in extracting the amplitude, phase and frequency of predominant sinusoids from the short-time Fourier transform (STFT) of the original signal. The partials are then removed to obtain the stochastic residual which is modeled as a time-varying colored noise. It is usually assumed that the stochastic contribution represents only a few components of the original sound that are not included in the deterministic contribution. In the present study, we focus on noisy environmental sounds for which the stochastic contribution is predominant compared to the deterministic contribution. Hence, in the following sections, we investigate analysis/synthesis techniques dedicated to stochastic signals and usually developed for modeling the stochastic residual of the deterministic plus stochastic model.

3. ANALYSIS/SYNTHESIS OF STOCHASTIC SIGNALS

In [13] a time-varying line-segment spectral envelope is proposed to model the stochastic residual of the sinusoid plus noise model. The spectral envelope is obtained by measuring the average energy of the residual on a set of contiguous frequency bands covering the whole frequency range. This representation has the advantage to be very flexible and it is generally accurate enough for sound synthesis applications. Analysis and synthesis of the residual in this manner can be seen as a particular use of the channel vocoder introduced by Dudley in 1939 [14] [15] [16]. Formerly used for speech coding, the channel vocoder reconstructs an original signal based only on its short time power spectrum (note that this contrasts with the so-called "phase vocoder" that keeps the phase information). The short time power spectrum of the speech is measured with a bank of bandpass filters. A pulse train (that simulates the glottal excitation) or noise (for transient parts of the speech) feed the same filterbank at the synthesis stage. Note that for efficiency, the analysis/synthesis filterbank is typically implemented with the short-time Fourier transform [16]. The whole process is illustrated on Figure 1 and described in the following two sections. For our concern the excitation is assumed to be only noise.

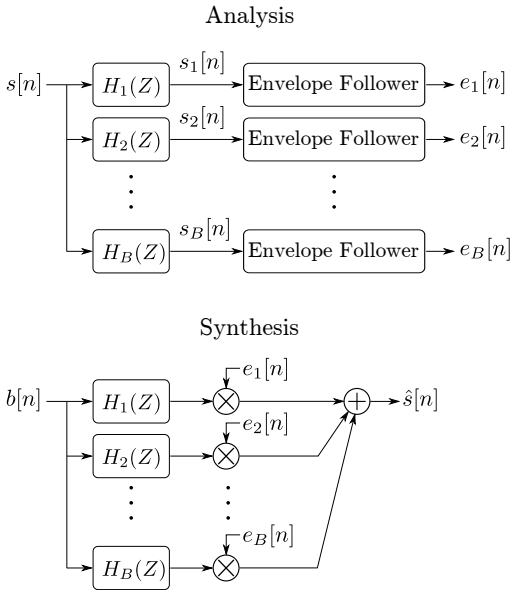

Figure 1: *Channel vocoder analysis/synthesis. Analysis: the original sound $s[n]$ is passed through a bank of B bandpass filters and the envelope is estimated in each subband, resulting in time-varying spectral envelope $E[n] = (e_1[n], \ldots, e_B[n])$. Synthesis: a white noise is passed through the filterbank and weighted by the spectral envelope. The output $\hat{s}[n]$ has approximately the same time-frequency characteristics as $s[n]$.*

3.1. Analysis

The analysis stage aims at characterizing the short-time power spectrum of the original sound $s[n]$. This is usually done by passing the signal through a bank of B contiguous bandpass filters and estimating the subband envelopes $(e_1[n], \ldots, e_B[n])$ with an envelope follower. A filterbank satisfying the perfect reconstruction constraints with subbands evenly spaced on the Equivalent Rectangular Bandwidth (ERB) scale is presented in [17]. Subband envelopes can be estimated by:

$$e_b[n] = \sqrt{\frac{1}{I}\sum_{i=0}^{I-1}(v[i]s_b[n+i])^2}$$

where $s_b[n]$ is the b^{th} subband signal and $v[n]$ an analysis window of size I (a simple rectangular window for instance). To reduce the amount of data, a discrete version of the envelopes is usually stored:

$$E^r = (e_1^r, \ldots, e_B^r) = (e_1[rR], \ldots, e_B[rR])$$

where r is the index of the frame and R the analysis hop size.

3.2. Synthesis

The synthesis can be implemented either in the time or in the frequency domain [17]. Here we focus on efficient frequency domain implementations described in [13] [18]. The synthesis is performed with a frame by frame approach. When the spectral envelope is estimated from the analysis of a natural sound $s[n]$ (see Section 3.1), the analysis and the synthesis windows and hope sizes can be different. In that case, the spectral envelope is interpolated at the synthesis frame rate. Figure 2 illustrates the synthesis process. For each frame, a short-time spectrum (STS) is created

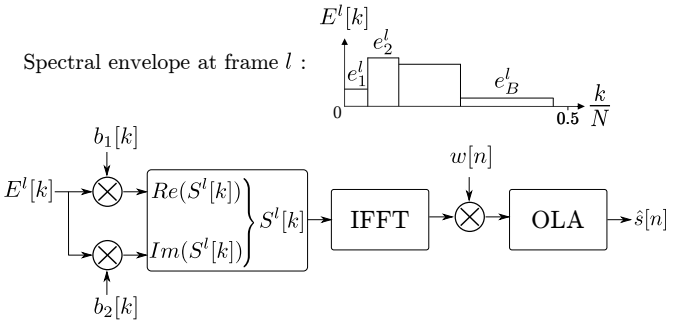

Figure 2: *Synthesis of a noisy signal $\hat{s}[n]$ from the spectral envelope $E^l[k]$ at frame l. First, $E^l[k]$ is multiplied by two random sequences $b_1[k]$ and $b_2[k]$ to get a short-time spectrum, then inverse fast Fourier transform (IFFT) is processed. The resulting frames are weighted by the synthesis window $w[n]$ and overlap-added (OLA).*

with the spectral envelope magnitude, and phases randomly distributed between 0 and 2π. The STS are inverse fast Fourier transformed (IFFT) weighted by the synthesis window and overlap-added (OLA) to obtain the time-domain stochastic signal $\hat{s}[n]$. Let N be the synthesis bloc size, k the discrete frequency index and $S^l[k]$ the N-point stochastic STS at frame l. Since the synthetic signal is real-valued in the time domain, its spectrum is conjugate-symmetric in the frequency domain. Thus, only positive frequencies (i.e., $k = 0, \ldots, \frac{N}{2}$) need to be considered. The spectral envelope E^l is resampled for $k = 0, \ldots, \frac{N}{2}$ and multiplied by two random sequences $b_1[k]$ and $b_2[k]$ to get the real and imaginary parts of $S^l[k]$:

$$Re(S^l[k]) = b_1[k]E^l[k]$$
$$Im(S^l[k]) = b_2[k]E^l[k]$$

The STS $S^l[k]$ is inverse fast Fourier transformed to get a short-time stochastic signal $s^l[n]$:

$$\text{with} \quad s^l[n] = \frac{1}{N}\sum_{k=0}^{N-1} S^l[k]e^{j2\pi\frac{k}{N}n}$$

Then all short-time signals are weighted by the synthesis window $w[n]$ of size N and overlap-added to get the whole reconstructed signal $\hat{s}[n]$:

$$\hat{s}[n] = \sum_{l=-\infty}^{\infty} w[n - lL]s^l[n - lL]$$

where L is the synthesis hop size. A normalization of $b_1[k]$, $b_2[k]$ and $w[n]$ is necessary so that $\hat{s}[n]$ and $s[n]$ have the same average power [17]. Additionally, in [13] [18] $b_1[k]$ and $b_2[k]$ satisfy:

$$b_1[k]^2 + b_2[k]^2 = 1 \text{ for } k = 0, \ldots, \frac{N}{2} \qquad (1)$$

so that the magnitude of the STS is exactly equal to the spectral envelope $E^l[k]$. However, if we consider the discrete Fourier transform $B[k]$ of a zero-mean N-point sequence of Gaussian white noise with variance σ^2, it is shown in [19] that the magnitude of $B[k]$ is a Rayleigh distribution (and the phase a uniform distribution). The real and imaginary parts of $B[k]$ are independent Gaussian sequences with variance $\frac{N}{2}\sigma^2$. Informal listening tests have confirmed that letting $b_1[k]$ and $b_2[k]$ be two independent Gaussian sequences, i.e., not satisfying Eq.1, leads also to good perceptive results.

4. APPLICATIONS FOR ENVIRONMENTAL SOUNDS

4.1. Tuning the model

The stochastic modeling can be perceptively relevant for many environmental sources. The main issue is to find a set of analysis/synthesis parameters satisfying time and frequency resolution constraints to be suitable for a wide range of sounds. On the one hand, signals whose temporal envelope varies rapidly (transient signals) require short analysis/synthesis windows (for example, less than 128 taps); on the other hand, signals whose spectral envelope varies rapidly in frequency (narrow-band noises) require long analysis/synthesis windows (for example, more than 2048 taps).

Regarding the frequency resolution, we found that 32 subbands evenly spaced on the ERB scale is usually sufficient to reproduce the salient properties of noisy environmental sounds, such as wind, sea waves and "whoosh" sounds. Regarding the time resolution, different analysis window sizes may be used according to the stationarity of the original sound. However, when using the IFFT synthesis algorithm described in Section 3.2, it is desirable to have a single synthesis window. This way sounds can be added directly in the frequency domain, by summing the STS, so that only one IFFT is processed per frame, whatever the number of sound sources. We experimented with several synthesis window sizes and found that 1024 taps lead to accurate resynthesis of a wide range of environmental sounds. Wind, sea waves and air swishing sounds usually do not have very sharp transients so that 1024 taps (i.e., 21 milliseconds at 48kHz) is sufficiently short. Furthermore, 1024 taps is sufficiently long for synthesizing 32 subbands evenly spaced on the ERB scale. This compromise between time and frequency resolutions appears to be relevant for many noisy environmental sounds. Sound examples can be found in [20]. Analysis/synthesis of a "whoosh" sound with this time/frequency resolution is illustrated on Figure 3.

4.2. Advantages of the model

Compared to classical wavetable synthesis, the stochastic modeling of environmental sounds has two main advantages: reducing the amount of data and allowing parametric transformations of the signal.

4.2.1. Data compression

The stochastic modeling allows saving a significant amount of data compared to the original signal. When using an analysis hop size of 512 samples (e.g., a 1024-tap analysis window with an overlap factor of 50%) and 32 ERB subbands, the compression ratio is $512/32 = 16$. When the original sound is relatively stationary, longer hop sizes can be used to increase the compression ratio without degrading the quality.

4.2.2. Signal transformations

In the context of analysis/synthesis, the analysis stage leads to the determination of the synthesis parameters for reconstructing the original sound. Modifying these synthesis parameters allows the creation of new sounds (see Figure 4). In [21] the authors use this framework for generating complex sound scenes from a small set of recorded environmental sounds.

The spectral envelope is a parametric representation of the stochastic signal that allows realizing signal transformations efficiently, such as pitch-shifting, time-stretching and morphing. Subband equalization of the reconstructed signal is possible by weighting the spectral envelope with a set of coefficient in the synthesis process. It provides an efficient and intuitive control of the spectral shape of the reconstructed sound.

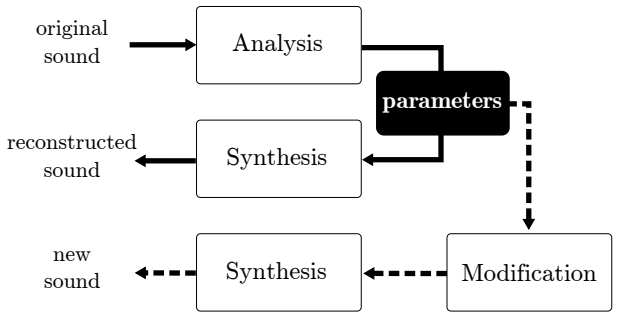

Figure 4: *Analysis/transformation/synthesis framework: the synthesis parameters extracted from the analysis of an original sound are used to resynthesize the original one or to create new sounds by signal transformation (e.g., by pitch-shifting, time-stretching, filtering or morphing).*

Additionally, the stochastic modeling allows controlling the correlation between several resynthesized versions of the same original sound. This property is very attractive for spatial sound transformations, such as simulating spatially extended sound sources (see Section 5.2).

5. SPATIALIZED SYNTHESIS OF NOISY ENVIRONMENTAL SOUNDS

From a general point of view, spatialization of sounds relates to 3D sound positioning, source width rendering, source directivity and reverberation. For our concern, we focus on the first and second aspects. We will see that the stochastic modeling can be combined with 3D audio modules for creating efficiently point-like and extended sound sources.

5.1. Source positioning

Several approaches exist for positioning sound sources in virtual environments. High Order Ambisonics and Wave Field Synthesis aims at reconstructing the sound field in a relatively extended area with a multichannel loudspeaker setup. Discrete panning (time or amplitude panning) reconstructs some aspects of the sound field at the "sweet spot". Binaural Synthesis reconstructs the sound field at the entrance to the ear canals by filtering the monophonic sound with Head Related Impulse Responses. It is mainly for headphone reproduction but can also be extended to loudspeaker setup (commonly referred to as "Transaural"). Even if all the techniques cited above have their own characteristics, a general implementation strategy is described in [22]. The sound positioning is decomposed into three modules:

1. Directional encoding thanks to a filterbank depending on the source virtual direction.
2. Mixing all sources in the multichannel encoded domain.
3. Directional decoding by matrixing and/or filtering the multichannel signal. The decoding stage is common for all sources: it does not depend on individual source position.

In [23] an architecture combining efficiently the spatialization modules and additive synthesis is proposed. This "spatialized synthesis" architecture is compatible with the stochastic modeling and the frequency-domain implementation described in Section 3.2. It handles all positioning methods that use only gains in the spatial

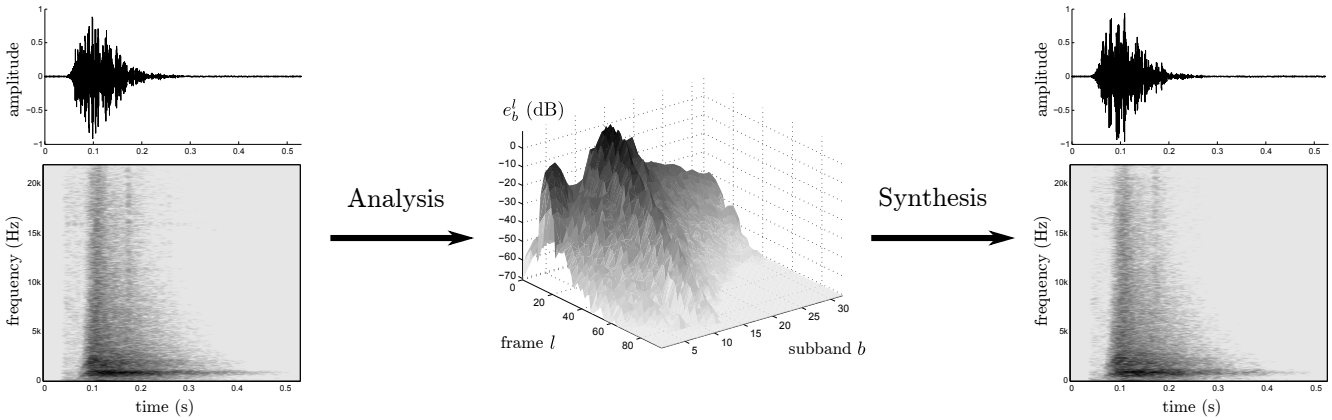

Figure 3: *Analysis-Synthesis of an air swishing ("whoosh") sound. (Left) Original signal and its STFT. (Middle) Time-varying 32-subband spectral envelope. (Right) Reconstructed signal and its STFT. The reconstructed signal is perceptively similar to the original one.*

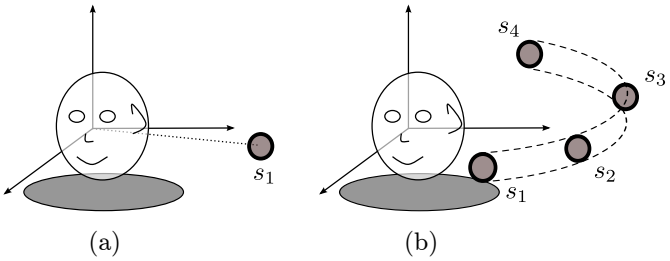

Figure 5: *(a) A single point-like source produces a narrow auditory event. (b) Several decorrelated copies of the source located at several positions around the listener produce a wide auditory event. The perceived source width can be adjusted by changing the relative contributions (i.e., gains) of the decorrelated sources.*

encoding module, i.e., Ambisonics, HOA, amplitude panning and some multichannel implementations of binaural synthesis. WFS is excluded since it requires delays in the spatial encoding module.

We use this spatialized synthesis architecture for simulating point-like noisy sources in the 3D space. Compared to the traditional implementation that consists in synthesizing a monophonic source before spatialization, it reduces the complexity. All the processing is performed in the frequency domain, so that the filtering of the spatial decoding module can be performed efficiently. Furthermore, one IFFT is required per frame for each loudspeaker channel, independently of the number of sound sources in the scene [23]. This is particularly attractive for applications over headphones because only two IFFT are computed per frame, while the scene can contain hundreds of sources.

5.2. Source width rendering

When a monophonic signal is simply duplicated to feed a multichannel loudspeaker setup, the listener perceives a relatively sharp phantom image. By contrast, a wide spatial image is perceived when decorrelated versions of the signal feed the loudspeakers. Several authors have proposed to create wide sound sources by positioning decorrelated copies of a signal at several locations around the listener [24] [25] [26] [27] (see Figure 5). Each copy is called a secondary source. Note that multiple-direction amplitude panning [28] is a similar process except that the secondary sources are not decorrelated so as to keep a sharp spatial image.

Various filtering techniques have been proposed for decorrelating the original signal to compute the secondary sources [29] [30] [31] [24]. The problem with filtering is that it may alter the transients and the timbre of the original sound. We propose an alternative approach to produce the decorrelated sources at the synthesis stage, thanks to the specific properties of the stochastic modeling. Let $s_1[n]$ and $s_2[n]$ be two resynthesized versions of the same original sound with the technique described in Section 3.2. At each frame l, $S_1^l[k]$ and $S_2^l[k]$ can be created with different noise sequences multiplied with the spectral envelope $E^l[k]$. The resulting signals $s_1[n]$ and $s_2[n]$ are two versions of the same original sound, but they are statistically uncorrelated. This way, an unlimited number of decorrelated secondary sources can be created by using different random sequences at the synthesis stage. Then, positioning the secondary sources at different locations produces wide sound sources.

Informal listening tests have shown that the proposed method can lead to realistic wide spatial images of noisy sources. Since the decorrelation is realized at the synthesis stage (i.e., without filtering) the timbre and the transients of the secondary sources do not suffer unnatural alteration. Sound examples can be found in [20]. A formal perceptual validation of the technique will be carried out in the future.

Note that in some cases it is also of interest to control the inter-channel correlation. For that purpose, a correlation C can be introduced between $s_1[n]$ and $s_2[n]$ and can be accurately controlled by creating $S_2^l[k]$ with:

$$Re(S_2^l[k]) = C\ Re(S_1^l[k]) + \sqrt{(1-C^2)}\ b_1[k]E^l[k]$$
$$Im(S_2^l[k]) = C\ Im(S_1^l[k]) + \sqrt{(1-C^2)}\ b_2[k]E^l[k]$$

where $b_1[k]$ and $b_2[k]$ are two independent Gaussian noise sequences. For instance, for stereo (2-channel) applications, this control allows going progressively from a sharp spatial image of the sound source towards a completely diffused source between the two loudspeakers.

6. CONCLUSION

An original approach has been presented for analysis/synthesis and spatialization of noisy environmental sounds. The use of stochastic modeling allowed us to propose an architecture where spatialization techniques operate directly at the synthesis stage for simulating wide sound sources. Informal listening tests have shown that our technique produces wide and realistic environmental sound

sources such as sea waves and wind. Furthermore, using the stochastic modeling for environmental sounds has several advantages. Compared to wavetable synthesis, the amount of data to be stored is significantly reduced and new parametric effects based on the analysis/transformation/synthesis framework are possible.

7. REFERENCES

[1] W. W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception," *Ecological Psychology*, vol. 5(1), pp. 1–29, 1993.

[2] W. W. Gaver, "How do we hear in the world? explorations in ecological acoustics," *Ecological Psychology*, vol. 5(4), pp. 285–313, 1993.

[3] K. van den Doel, P. G. Kry, and D. K. Pai, "Foleyautomatic: physically-based sound effects for interactive simulation and animation," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 537–544.

[4] J. F. O'Brien, C. Shen, and C. M. Gatchalian, "Synthesizing sounds from rigid-body simulations," in *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2002, pp. 175–181.

[5] N. Raghuvanshi and M. C. Lin, "Interactive sound synthesis for large scale environments," in *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, 2006, pp. 101–108.

[6] Y. Dobashi, T. Yamamoto, and T. Nishita, "Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics," *ACM Transactions on Graphics (Proc. SIGGRAPH 2003)*, vol. 22(3), pp. 732–740, 2003.

[7] Y. Dobashi, T. Yamamoto, and T. Nishita, "Synthesizing sound from turbulent field using sound textures for interactive fluid simulation," *EUROGRAPHICS*, vol. 23(3), pp. 539–546, 2004.

[8] K. van den Doel, "Physically-based models for liquid sounds," in *Proceedings of ICAD 04-Tenth Meeting of the International Conference on Auditory Display*, 2004.

[9] S. Conversy, "Ad-hoc Synthesis of auditory icons," in *Proceedings of of ICAD 98-The fifth International Conference on Auditory Display*, 1998.

[10] N. E. Miner and T. P. Caudell, "Using wavelets to synthesize stochastic-based sounds for immersive virtual environments," in *Proceedings of of ICAD 97-The fourth International Conference on Auditory Display*, 1997.

[11] P. R. Cook, *Real Sound Synthesis for Interactive Applications*, A. K Peters Ltd., 2002.

[12] D. Rocchesso and F. Fontana, *The Sounding Object*, http://www.soundobject.org/, 2003.

[13] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comp. Music. J.*, vol. 14(4), pp. 12–24, 1990.

[14] H. Dudley, "The vocoder," *Bell Labs Record*, vol. 17, pp. 122–126, 1939.

[15] B. Gold and C. M. Rader, "The channel vocoder," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 4, pp. 148–161, 1967.

[16] J. O. Smith, *Spectral Audio Signal Processing, October 2008 Draft*, http://ccrma.stanford.edu/∼jos/sasp/, 2008, online book.

[17] M. M. Goodwin, *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*, Ph.D. thesis, University of California, Berkeley, 1997.

[18] M. M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996.

[19] J.-J. Sung, G.-S. Kang, and S. Kim, "A transient noise model for frequency-dependent noise sources," *IEEE Trans. on CAD of Integrated Circuits and Systems*, vol. 22, no. 8, 2003.

[20] *www.lma.cnrs-mrs.fr/∼kronland/spatsynthIcad09/index.html*.

[21] A. Misra, P. R. Cook, and G. Wang, "A new paradigm for sound design," in *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx'06)*, 2006.

[22] J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A comparative study of 3-d audio encoding and rendering techniques," in *Proc. 16th Int. Conf. AES*, 1999.

[23] C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone, "Spatialized additive synthesis of environmental sounds," in *Proceedings of the 125th AES Convention*, 2008.

[24] G. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, vol. 19(4), pp. 71–87, 1995.

[25] A. Sibbald, "Method of synthesizing an audio signal," United State Patent No. US 6498857 B1, december 2002.

[26] G. Potard and I. Burnett, "Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays," in *Proc. Int. Conf. on Digital Audio Effects (DAFx'04)*, 2004.

[27] J.-M. Jot, M. Walsh, and A. Philp, "Binaural simulation of complex acoustic scenes for interactive audio," in *Proceedings of the 121th AES Convention*, 2006.

[28] V. Pulkki, "Uniform spreading of amplitude panned virtual sources," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

[29] M. R. Schroeder, "An artificial stereophonic effect obtained from a single audio signal," *JAES*, vol. 6(2), 1958.

[30] R. Orban, "A rational technique for synthesizing pseudo-stereo from monophonic sources," *JAES*, vol. 18(2), 1970.

[31] M. A. Gerzon, "Signal processing for simulating realistic stereo images," in *AES Convention 93*, 1992.