CONTROLLING A SPATIALIZED ENVIRONMENTAL SOUND SYNTHESIZER

Charles Verron, Grégory Pallone

Orange Labs
2, avenue Pierre Marzin
22307 Lannion, France

Mitsuko Aramaki, Richard Kronland-Martinet

CNRS
Institut de Neurosciences Cognitives de la Méditerranée
Laboratoire de Mécanique et d'Acoustique
13402 Marseille, France

ABSTRACT

This paper presents the design and the control of a spatialized additive synthesizer aiming at simulating environmental sounds. First the synthesis engine, based on a combination of an additive signal model and spatialization processes, is presented. Then, the control of the synthesizer, based on a hierarchical organization of sounds, is discussed. Complex environmental sounds (such as a water flow or a fire) may then be designed thanks to an adequate combination of a limited number of basic sounds consisting in elementary signals (impacts, chirps, noises). The mapping between parameters describing these basic sounds and high-level descriptors describing an environmental auditory scene is finally presented in the case of a rainy sound ambiance.

Index Terms— Sound synthesis, environmental sounds, spatialization, control.

1. INTRODUCTION

The synthesis of environmental sounds is of great importance for numerous audio applications such as virtual reality, video games or movie sound effects. Several authors already addressed such a problem by developing synthesis models based on physical [1] [2] or signal [3] [4] approaches in order to give the sound designer adequate tools. For some applications such as video games, constraints based on interactivity and realism in the process of 3D immersion lead us to propose an environmental sound synthesizer based on signal model and to integrate the spatialization process at its primary level [5]. This architecture, based on an additive synthesis engine designed in the frequency domain, has proven its ability to provide realistic environmental sounds by extracting synthesis parameters from the analysis of natural sounds (sounds examples can be found in [6]).

In this paper, we address the problem of the control of such a synthesizer so that the user can easily generate spatialized environmental sounds and 3D auditory scenes from high-level descriptors (based on sound taxonomy proposed by Gaver [7]) that can be manipulated interactively from either standard MIDI sound interfaces or from data provided by the visual world. For that purpose, we propose a hierarchical sound design process starting from "basic sounds" (also called "atoms") from which most of environmental sounds can be efficiently constructed (such as drops, wind, fire,...) up to auditory scenes. These elements are controlled from high-level descriptors (such as size, intensity, trajectory,...) that are directly linked to the way basic sounds are combined and adjusted. The complete 3D auditory scene is obtained by combining the environmental sounds in an appropriate way, usually defined by intuitive high-level general descriptors. We briefly describe the synthesis engine previously developed and including the spatialization [5]. Then, we focus on an intuitive control of the sound synthesizer by presenting how environmental sounds can be designed with a suitable combination of basic sounds (the example of a water drop is fully described). We finally discuss the design and control of a 3D auditory scene based on combination of various environmental sounds.

[1]This research was partially supported by the French National Research Agency (ANR, JC05-41996, "senSons").

2. THE SYNTHESIS ENGINE

The synthesis engine is based on an efficient combination of additive synthesis and 3D positional audio modules. Sound synthesis and spatialization are implemented at the same level of the sound generation, by contrast with classical approaches that consist in synthesizing a monophonic sound in a first stage and positioning the source in 3D space in a second stage. Here, we briefly describe the spatialized additive synthesis engine based on additive signal model (see [5] for details). The complete synthesis process allows generating a sound $x(t)$ modeled by the summation of two separate entities, i.e., the deterministic part $s_D(t)$ and the stochastic part $s_S(t)$. Practically, the synthetic sound $x(t)$ is constructed in real-time with a frame by frame approach in three stages that are successively described below.

Stage 1: Time-frequency domain synthesis At each frame l, an approximation of the short-time Fourier transform of the sound is built by summing the deterministic S_D^l and stochastic S_S^l contributions. Real and imaginary parts of the deterministic short-time spectrum (STS) are computed by convolving the theoretical ray spectrum (formed by the M sinusoidal component of amplitude a_m^l, frequency f_m^l and phase Φ_m^l) with the "spectral motif" W (Fourier transform of the synthesis window $w[n]$) [8]:

$$S_D^l[k] = \sum_{m=1}^{M} a_m^l e^{j\Phi_m^l} W(\frac{k}{N} - f_m^l) \qquad (1)$$

where N is the number of frequency bins (i.e., the synthesis window size) and k the discrete frequency index (i.e., $W[k] = W(\frac{k}{N})$). Real and imaginary parts of the stochastic STS are computed from the amplitude spectral envelope (defined by B subband coefficients $(e_1^l, \ldots, e_B^l)$ on the equivalent rectangular bandwidth (ERB) scale) multiplied by two noise sequences and circularly convolved with the spectral motif W. Real and imaginary parts of the whole STS $X^l[k]$ are obtained by summing the two respective contributions.

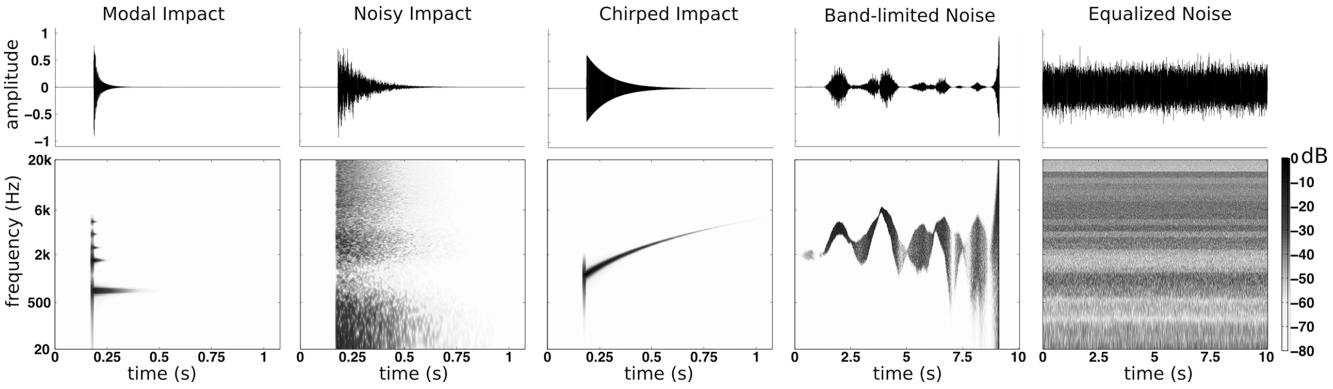

Figure 1: Temporal (first row) and time-frequency representations (second row) of the five atomic elements. Environmental sounds are generated by composition of theses primitives. The first three atoms (modal, noisy and chirped impacts) are impulsive and the last two atoms (band-limited and equalized noises) are continuous sounds.

Stage 2: Spatialization The spatial encoding depends on the 3D positioning method. When using "amplitude-based" 3D audio methods, which avoid delays in the spatial encoding stage (e.g., Ambisonics, HOA, multichannel binaural and amplitude panning methods) the C-channel encoding consists in applying real-valued spatial gains $(g_1, \ldots, g_C)$ on each monophonic sound source. Practically, such gains are directly applied to the whole STS $X_i^l[k]$ (from Stage 1) of each source i. After spatial encoding, the mixing stage consists in summing encoded STS together, channel by channel. It results in a single C-channel frame signal Y^l whose c^{th} channel is:

$$Y_c^l[k] = \sum_{i=1}^{I} g_c(\theta_i^l, \Psi_i^l) X_i^l[k]$$

where g_c is the c^{th} position-dependent spatial gain, (θ_i^l, Ψ_i^l) the position of the i^{th} source and I the number of sound sources.

Stage 3: Reconstruction in the time-domain The spatial decoding is performed by matrixing and/or filtering the C channels of Y^l, depending on the 3D audio method. Then, they are inverse fast Fourier transformed and overlap-added to reconstruct the synthetic signal $x_c[n]$ for the c^{th} loudspeaker:

$$x_c[n] = \sum_{l=-\infty}^{\infty} g_c(\theta_i^l, \Psi_i^l) w[n - lL] \Big(s_D^l[n - lL] + s_S^l[n - lL] \Big)$$

where s_D^l and s_S^l are the deterministic and stochastic short-time signals at frame l and L is the synthesis hop size.

Simulating wide sound sources The architecture of the synthesizer benefits of easily extending the perceived width of the sound source. To simulate an extended sound source, decorrelated secondary sources are computed from the same original set of synthesis parameters and positioned in 3D space around the listener. For the deterministic contribution, the STS are synthesized with the same original amplitude and frequency parameters, but phases at the origin are randomized for each sinusoidal component. For the stochastic contribution, decorrelation is achieved by synthesizing each STS with the same original amplitude spectral envelope but noise sequences differ for each decorrelated version. Practically, eight virtual secondary sources are synthesized and evenly positioned on a horizontal circle surrounding the listener. We proposed a spatial extension parameter that modifies the perceived width of the sound source by acting on the relative contributions of the secondary sources [5].

Efficiency By contrast with a classical two-stage implementation (synthesizing monophonic sources before spatialization) that requires one IFFT per source, this architecture requires one IFFT per loudspeaker, independently of the number of sources. As complex auditory environments can contain hundreds of sound sources, it becomes a serious advantage to provide this architecture. It is particularly attractive when using a multichannel implementation of binaural synthesis for mobile-phone applications with 3D audio on headphones. First because the spatial decoding requires linear filtering that can be efficiently applied in the frequency domain. Second, since the C channels of Y^l are downmixed to two binaural signals after the spatial filtering, only two IFFT are computed per frame whatever the number of sound sources.

3. DESIGNING SPATIALIZED ENVIRONMENTAL SOUNDS

The synthesis engine allows reproducing realistic spatialized environmental sounds by estimating synthesis parameters from the analysis of natural sounds. At low level, the sound generation necessitates acting on hundreds of synthesis parameters the manipulation of which is not adapted to interactive constraints of sound design. To address this issue, we propose here an efficient way to design environmental sounds by a combination of few basic elements. Indeed, environmental sounds refer to a wide range of various sounds but interestingly, their acoustic morphology calls for common signal characteristics allowing for a granular-like synthesis process. Based on our investigations, we identified five fundamental "atoms" which relevancy and sufficiency were tested by analyzing and resynthesizing various environmental sounds in each category: Solids, Liquids, Aerodynamics. These atoms are shown in Figure 1 and defined below. Note that this atom dictionary may be completed in the future without compromising the proposed methodology.

- the modal impact atom is a sum of M exponentially decaying sinusoids with amplitudes a_m, frequencies f_m, phases Φ_m

and damping coefficients α_m:

$$x_1(t) = \sum_{m=1}^{M} a_m \cos\big(2\pi f_m t + \Phi_m\big) e^{-\alpha_m t}$$

- the noisy impact atom is a sum of B exponentially decaying subbands of noise $s_b(t)$ with amplitudes a_b and damping coefficients α_b:

$$x_2(t) = \sum_{b=1}^{B} a_b s_b(t) e^{-\alpha_b t}$$

- the chirped impact is an exponentially decaying swept sinusoid with amplitude a, damping coefficient α, time-varying instantaneous frequency $f(t)$ and phase at origin ϕ:

$$x_3(t) = a \cos\Big(2\pi \int_0^t f(\nu) d\nu + \phi\Big) e^{-\alpha t}$$

- the band-limited noise is a time-varying filtered noise whose spectral envelope is defined as:

$$X_4(f) = \begin{cases} a(t) & \text{if } |f - f_0(t)| < \frac{B(t)}{2} \\ a(t) e^{-\alpha(t)\left(|f - f_0(t)| - \frac{B(t)}{2}\right)} & \text{if } |f - f_0(t)| > \frac{B(t)}{2} \end{cases}$$

where $f_0(t)$ is the center frequency, $a(t)$ the filter gain, $B(t)$ the bandwidth and $\alpha(t)$ the filter spectral slope.

- the equalized noise is a sum of 32 ERB subbands of noise $s_b(t)$ with time-varying amplitudes $a_b(t)$:

$$x_5(t) = \sum_{b=1}^{32} a_b(t) s_b(t)$$

Figure 2 illustrates the decomposition process of a water drop from the analysis of a natural sound. In practice, a drop can be obtained by combining two types of atoms: noisy and chirped impacts. Hence, most of environmental sounds can be designed by combining some of these atoms and then spatialized by positioning them in the 3D space.

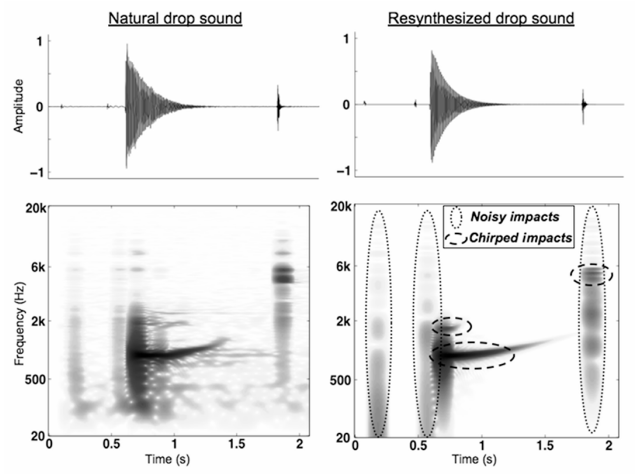

Figure 2: Temporal (first row) and time-frequency representations (second row) of a natural (left) and resynthesized (right) drop sound. The drop is reconstructed by combining three noisy impacts and three chirped impacts.

4. INTUITIVE CONTROL OF THE SYNTHESIZER

The sound taxonomy of the synthesizer accessible to the user is based on the classification proposed by Gaver: environmental sounds are indexed in Solids, Liquids, Air or Explosion category. The user can act on several high-level control parameters related to both timbre and spatial attributes. Some examples are shown in Table 1.

Similarly to the design of environmental sounds by combining atoms, complex 3D auditory scenes can be intuitively designed by combining spatialized environmental sounds. As an example, we describe below the design of a rainy weather sound ambiance. For that, we combine a streaming, a rain shower and drops, each of these environmental sounds being independently spatialized. The table in Figure 3 illustrates construction of the auditory scene from environmental sounds that are themselves constructed from atoms. Figure 3 also shows the evolutive control of the Environmental sound parameters during the control of Intensity parameter of the rainy weather (Mapping 2). In particular, both the Flow parameter of Streaming and Intensity parameter of Rain shower increase. The Occurrence of Drops also increases in accordance with the Intensity curve of the Rain shower.

5. CONCLUSION AND PERSPECTIVES

We have proposed a general environmental sound synthesizer allowing for an interactive synthesis of spatialized auditory scenes. The originality of the system relates first in the combination of an additive signal synthesis and spatialization processes at the primarily level of the architecture, allowing for a real-time implementation. Second, the design of environmental 3D scenes is based on a hierarchical organization enabling easy and intuitive controls of sounds thanks to high-level descriptors directly mapped to the signal parameters of "atoms". In practice, these controls can be achieved from either MIDI interfaces in an interactive way or automatically from data provided by a video game engine. Sound examples illustrating the paper are available [9].

6. REFERENCES

[1] 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.

[2] 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.

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

[4] A. Farnell, *Designing sound, procedural audio for games and film*. Applied Scientific Press, 2008.

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

[6] *www.lma.cnrs-mrs.fr/∼kronland/spatsynthIEEE/index.html*.

[7] 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.

[8] X. Rodet and P. Depalle, "Spectral envelopes and inverse fft synthesis," in *Proc. of the 93rd AES Conv.*, 1992.

[9] *www.lma.cnrs-mrs.fr/∼kronland/spatsynthWaspaa/index.html*.

Category	Environmental sounds	High-level control parameters
SOLIDS	Impact	Size, Material, Force, 3D position
	Rolling	Speed, Regularity, Trajectory
	Crumpling	Size, Speed, 3D position
LIQUIDS	Drop	Size, 3D position
	Streaming	Flow, Trajectory, Width
	Rain shower	Intensity, 3D position, Width
	Wave	Size, Intensity, 3D position, Width
AIR	Wind	Force, Perturbations, 3D position, Width
	Whoosh	Speed, Length, Trajectory
	Periodic whoosh	Periodicity, 3D position, Width
EXPLOSION	Spark	Intensity, 3D position
	Fire	Intensity, crackling density, 3D position, Width
	Shot	Size, Intensity, 3D position, Width
	Thunder	Intensity, Duration, 3D position, Width

Table 1: Control of the synthesizer: the set of environmental sounds is indexed in Solids, Liquids, Air or Explosion category. For each category, some examples of environmental sounds are given and their corresponding high-level control parameters (related to both timbre and spatial attributes) are defined.

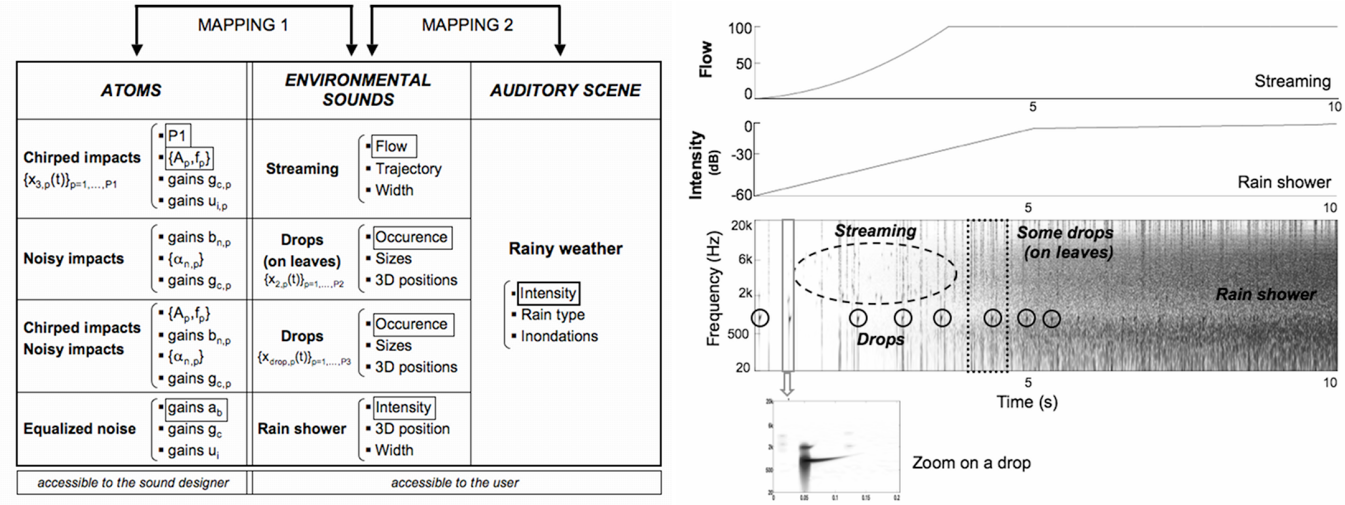

Figure 3: **Left:** Design of a 3D auditory scene evoking a rainy weather ambiance by combining a Streaming, a Rain shower and Drops. Each of these environmental sounds is designed by combining atoms. **Right:** Example of Mapping 2: when the Intensity parameter of the rainy weather increases, both the Flow of the Streaming (expressed in number of chirped impacts per second) and the Intensity of the Rain shower (in dB) increase. The occurrence of Drops increases in accordance with the Intensity curve of the Rain shower. Environmental sounds that are involved in the scene are highlighted in the time-frequency representation.