A 3-D Immersive Synthesizer for Environmental Sounds

Charles Verron, *Student Member, IEEE*, Mitsuko Aramaki, *Member, IEEE*, Richard Kronland-Martinet, *Senior Member, IEEE*, and Grégory Pallone

***Abstract*—Nowadays, interactive 3-D environments tend to include both synthesis and spatialization processes to increase the realism of virtual scenes. In typical systems, audio generation is created in two stages: first, a monophonic sound is synthesized (generation of the intrinsic timbre properties) and then it is spatialized (positioned in its environment). In this paper, we present the design of a 3-D immersive synthesizer dedicated to environmental sounds, and intended to be used in the framework of interactive virtual reality applications. The system is based on a physical categorization of environmental sounds (vibrating solids, liquids, aerodynamics). The synthesis engine has a novel architecture combining an additive synthesis model and 3-D audio modules at the prime level of sound generation. An original approach exploiting the synthesis capabilities for simulating the spatial extension of sound sources is also presented. The subjective results, evaluated with a formal listening test, are discussed. Finally, new control strategies based on a global manipulation of timbre and spatial attributes of sound sources are introduced.**

***Index Terms*—Environmental sounds, frequency-domain additive synthesis, sound analysis/synthesis, sound spatialization, source spatial extension, source width evaluation, spatial sound effects.**

I. Introduction

From the early stages, virtual reality applications used to focus on graphics rendering to create realistic immersive 3-D scenes. However, combining multimodal aspects of our environment increases the sensation of immersion and in particular, the auditory modality brings complementary information to the vision. Sound design by synthesis methods made crucial progress, providing efficient tools to generate high quality sounds either from models (physical-based or signal-based) or from the analysis of natural sounds [1], [2]. Linking visual and auditory modalities implies the construction of coherent visual and sound events. For that purpose, a relationship between visual scenes and their sound counterpart has to be accurately established. In the case of 3-D environmental scenes, these relationships are mainly linked to two main attributes of sound sources: their intrinsic timbre and their spatial features (i.e., the position relative to the listener, the width and the directivity). These attributes are intricately linked since spatial distribution of sound energy depends on the physical sources and on the way they are excited. For this reason, they would gain in being associated at the same level of the sound generation.

In this paper, we present the design of a 3-D immersive synthesizer dedicated to environmental sounds to be used in the framework of virtual reality applications such as video games, animation, audiovisual immersion, etc. The synthesizer aims at generating realistic sounds evoking a wide variety of phenomena from our natural environment (wind, liquids, impacts, etc.) and unusual sounds for special audio effects purposes. It also aims at providing an accurate tool to create complex scenes implying several sound sources that can be spread in the 3-D space and controlled in their spatial width and in motion. On top of these considerations, the synthesizer has to run in real-time to fulfill the constraints of interactivity.

Several authors have proposed methods for audio generation in virtual environments and video games (see [3]–[7] for a review of recent progress). Some works are attached to real-time synthesis with physically based models [5]. Particularly, most studies focused on vibrating solid sounds and modal resonance modeling [7]–[9]. Physically based models for other types of environmental sounds (like wind, fire, explosion and water sounds) have also been investigated [10]–[12]. Likewise, signal-based approaches have been used successfully for synthesis of environmental sounds [3], [4]. For example, the "harmonic plus noise" and "source filter" models as well as granular and pitch-synchronous overlap-add (PSOLA) techniques allow synthesis (and transformations) of a wide class of musical, speech, and environmental sounds [13], [14], [1], [15].

Spatialization techniques aim at creating spatial sound attributes (see [16] for a review of the different methods). Most studies have focused on simulating the position and motion of sources around the listener [17]–[19], [20] as well as the sound propagation in enclosed spaces (room reverberation). Computational issues related to the rendering of complex scenes with hundreds of sound sources have also been investigated [21]. Complementary to recent works attached to spatialization (like [22]–[24]) the system proposed in this paper includes both synthesis and spatialization of the virtual sources.

In virtual auditory environments, synthesis and spatialization are generally processed in two separated stages of sound generation, i.e., first monophonic sounds are synthesized with specific

Manuscript received December 30, 2008; revised October 11, 2009. First published November 24, 2009; current version published July 14, 2010. This work was supported in part by the French National Research Agency (ANR, JC05–41996, "senSons"; http://www.sensons.cnrs-mrs.fr/). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick Naylor.

C. Verron is with the Department of Audiovisual and Speech for Quality (OPERA), Orange Labs, 22307 Lannion, France, and also with the CNRS-Laboratoire de Mécanique et d'Acoustique (LMA), 13402 Marseille, France. (e-mail: verron@lma.cnrs-mrs.fr).

M. Aramaki is with the CNRS-Institut de Neurosciences Cognitives de la Méditerranée, 13402 Marseille, France.

R. Kronland-Martinet is with the CNRS-LMA, 13402 Marseille, France.

G. Pallone is with the Department of Audiovisual and Speech for Quality (OPERA), Orange Labs, 22307 Lannion, France.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2037402

synthesis methods, second they are spatialized with appropriate spatialization techniques. A few studies have merged additive synthesis and spatialization aspects for simulating directional sound sources [25]–[27], or creating musical effects [28]. In [29] the authors present a sound rendering algorithm that combines sinusoidal synthesis and binaural filtering at the synthesis stage. They use this approach to synthesize and spatialize efficiently impact sounds in a virtual environment. Recently, we proposed new architectures that handle the synthesis of sinusoids and filtered noise, compatible with several 3-D audio formats (multichannel, Ambisonics, binaural) [30], [31].

This paper presents the design of a unified 3-D immersive synthesizer for environmental sounds. It is organized as follows: first we specify the notion of environmental sounds leading to a sound classification to be used for an efficient representation of the sonic real world. Then we review the existing methods for synthesis, spatialization and source extension. Based on these considerations, we propose an original "spatialized synthesis engine" for synthesizing, spatializing and spatially extending virtual sound sources on arbitrary loudspeaker setups. The proposed architecture is based on a single-stage combination of inverse fast Fourier transform synthesis and amplitude-based 3-D positioning. The additional cost per source is significantly reduced compared to classical two-stage implementations and the synthesizer satisfies real-time requirements. The spatial extension is achieved by a novel method that bypasses the classical decorrelation filtering stage and that is directly included in the synthesis process. A subjective evaluation of the synthesizer is presented in the case of various categories of environmental sounds (wind, bubble noise, drops, sea waves, etc.) with a formal validation protocol of the spatial extension control parameter. Finally, the possibilities offered by the synthesizer are presented. Sound examples are available online [32].

II. Environmental Sounds

The class of environmental sounds covers a large variety of sounds since it relates to all events occurring in listener's surroundings. Consequently, a generic definition of such sounds cannot be easily found. According to the field of environmental sound research, we assume that the generic environmental sounds class relates to sounds naturally occurring in our everyday life other than speech, music, animal communication, and electronic abstract sounds [33]–[35]. Several studies investigated the identification and classification of such sounds [33], [36]–[39]. Identification experiments revealed that listeners asked to listen to sounds and to tell what they heard, spontaneously described the event that caused the sound rather than the sound itself. In particular, they described the action (door opened, hammering…), the object that is actioned (doorbell, telephone…) or the context/location in which the object stands (traffic, office, etc.). These experiments support the ecological theory proposed by Gibson [40] assuming that a listener tends to identify its environment, i.e., that he perceives the properties of a sound event (distal stimulus) rather than the properties of the acoustical signal (proximal stimulus). It is worth noticing that in some cases, this ability to recover the sound event properties can be misled. For instance, in the case of impact sounds, the material identification is quite good for wood and metal gross-categories but the confusion between wood and plastic or between metal and glass is high [41]. The object shape identification (e.g., circle, square, or triangle shape) is also largely poor [42].

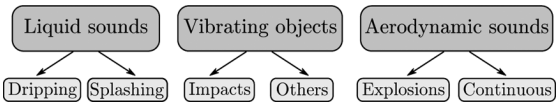

Fig. 1. Main categories of environmental sounds (from Gaver [38]).

Based on these ecological considerations, Gaver defined this source-oriented listening as "everyday listening" in contrast with "musical listening" which is related to the perception of the quality of the sound itself [38], [39]. In addition, he proposed a taxonomy of environmental sounds supported by the physics of sound events. This classification is hierarchically organized and the first level of sound-producing events is divided in three categories: vibrating solids, aerodynamic and liquid sounds. Interestingly, the author pointed out that according to previous studies, nobody confused the sounds between these categories from a perceptual point of view. At the low level of this hierarchy, basic levels of sound-producing events are defined (e.g., impacts for vibrating solids, explosions for aerodynamic or dripping for liquid sounds). Based on these categories, Gaver further proposed a general "map" of everyday sounds which includes more complex events (e.g., rain on surface, fluttering cloth, etc.) that are located at overlapping regions [38].

Alternative strategies to categorize environmental sounds were also investigated. For instance, several studies examined the sound categorization according to their acoustic properties independently from the nature of the event. In particular, dissimilarity or classification protocols were conducted by using tasks oriented to perceptual properties of sounds [35], [43]. Data analysis revealed that the harmonicity (related to pitch) or periodicity (related to rhythmic patterns) were relevant parameters to "explain" the acoustic categories. In a context of musical composition, Schaeffer proposed a general sound typology based on what he called the "acousmatic" listening [44]. Here, sounds are considered as "auditory objects" for human perception, whatever their physical causes, and are classified according to their morphological features: form (iterative, continuous, impulsive, etc.), matter (inharmonic, rough, etc.) and variation (melodic profile). This typology may constitute an interesting tool to better investigate the acoustic features that convey relevant signification for human perception [45].

For our concern, we consider that the sound taxonomy proposed by Gaver is well adapted to accurately represent the different categories of environmental sounds since it corresponds to a very intuitive way of describing the sounds. Fig. 1 shows these main categories that we used in the user interface to control the environmental sound synthesizer.

III. Synthesizer Design

In this section, we discuss the different processes involved in the design of a generic 3-D immersive synthesizer. In particular, techniques adapted to our purposes are discussed for sound synthesis, 3-D positioning and source spatial extension.

A. Sound Synthesis

Since the pioneer's works of Mathews [46], the field of sound synthesis has been widely investigated, leading to numerous techniques to generate sounds. These techniques can be decomposed into two main families: the physical-based methods, aiming at simulating the physics of sound sources and the signal-based methods, aiming at reproducing perceptual effects independently of the sound event [1].

On one hand, physical approaches are of great interest for environmental sounds since, as discussed in the previous section, ecological categorization of sounds based on physical events has been adopted. Some authors have successfully used physical models for simulating for example wind and fire noises [10], [11] or rolling sounds [47]. Nevertheless, the physics beyond these phenomena is often complicated, generally involving the knowledge of mechanical characteristics (size, shape material, etc.) and their possible interaction with surrounding fluids, liquids, or solids. Moreover, simulating some common environmental sounds lead to complex dynamical models. For instance, modeling accurately the acoustical phenomena corresponding to a sea wave sound is still an open issue. Thus, the difficulty of designing a general physical model for the whole environmental sound family, in addition to the complexity of the calibration processes and the often heavy calculations makes unrealistic the choice of such an approach for our purpose.

On the other hand, signal-based models aim at generating nonstationary signals from their morphological description (usually related to their time-frequency representation behavior). Both nonlinear and linear models have been proposed in the literature and provide efficient algorithms to generate a wide palette of timbres. In particular, nonlinear models such as frequency modulation [48] or waveshaping [49] allow sound creation by acting on a few control parameters. By contrast, linear models such as additive, subtractive and granular synthesis generally necessitate the control of a huge amount of parameters [1]. Nevertheless, recent works on the control of linear synthesis models allow to easily control sound timbres thanks to adequate mapping strategies based on perceptual considerations, allowing us to better understand the relationship between the sound signal and the way our auditory system processes the information [50].

These fundamental synthesis aspects led us to choose linear signal-based models to design the environmental sound synthesizer. Besides, we considered the so-called "physically informed" linear signal-based models. These models allow generating the synthesis parameters not only from signal properties but also from physical considerations, allowing the connection between signal and physics. Note that linear methods present the great advantage of permitting inversion processes since synthesis parameters can be determined from the analysis of natural sounds [2] (see Section V-A). Efficient signal transformations can be processed, based on the parametric representation resulting from the analysis stage. This approach has been used for generating a complete environmental scene starting from a small set of recorded sounds [68]. High-quality transformations can be achieved when the analyzed signal matches the chosen model. For example, the source-filter model used in PSOLA methods is well adapted to simulate speech signals, and allows precise and realistic time–frequency manipulations such as pitch-shifting with formant preservation, time-stretching, etc.[51].

In practice, we used the additive linear model to generate signals constituted of time-varying spectral lines and noisy contributions [13], [14]. The complete synthesis process allows generating a sound $x(t)$ modeled by summing two separate entities, i.e., the deterministic part $d(t)$ and the stochastic part $s(t)$

$$x(t) = d(t) + s(t).$$

The deterministic part $d(t)$ is given by summing M sinusoids:

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

where $f_m(t)$ and $a_m(t)$ are the instantaneous frequency and amplitude for component m and Φ_m is the phase at $t = 0$. Instantaneous amplitudes and frequencies are supposed to vary slowly in time (i.e., they are constant within a 20-ms window). The stochastic part $s(t)$ is modeled by a time-varying filtered noise.

B. 3-D Positional Audio

3-D positional audio techniques have been developed to simulate the position of a sound relative to the listener: monophonic sounds are converted into point-like virtual sources for creating an immersive audio space. Several techniques are now well established and widely used. Ambisonics and its extension high-order ambisonics (HOA) [17], pair-wise or triplet-wise amplitude panning such as VBAP [18] and wave field synthesis (WFS) [17] have been developed for spatialization over a multichannel loudspeaker system. The binaural technique uses head-related transfer functions (HRTF) for generating the soundfield at the entrance to the listener's ear canals for reproduction over headphones. A comprehensive comparison of these methods can be found in [16] and [19]. In the context of the study, we do not focus on one particular technique: we want the synthesizer to be compatible with several 3-D positioning methods.

C. Spatial Extension Methods

Some environmental sound sources, such as sea waves or wind in a tree, are naturally diffused and spatially extended. It is then of great importance that the synthesizer takes into account the control of the spatial extent of the sources to reproduce more realistic and immersive sound scenes. Several experiments already showed that the perceived width of auditory events is related to signal frequency, loudness, and duration, and to the interaural cross-correlation (IACC) of binaural signals [52]–[55]. Thus, simulating a wide source from an original monophonic sound requires manipulating one or several of these parameters. Spatial extension techniques usually manipulate only the interaural cross-correlation to conserve the timbre of the original sound as much as possible. Several authors proposed to create secondary sources (decorrelated versions of the original sound) positioned at various locations to produce an extended sound source [56], [57]–[60]. The decorrelated secondary sources can be produced by passing the original sound through orthogonal all-pass filters [56] or complementary linear-phase filters [57].

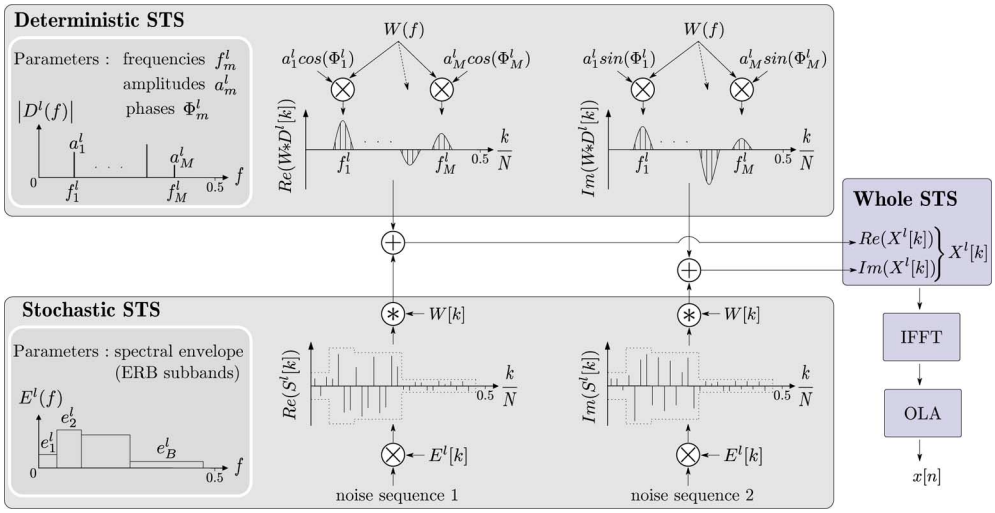

Fig. 2. Inverse FFT synthesis of a stochastic+deterministic signal $x[n]$. The continuous normalized frequency is noted f, k is the discrete frequency index and N the synthesis window size. At each frame l, real and imaginary parts of the deterministic short-time spectrum (STS) are computed from the synthesis parameters (i.e., the amplitudes a_m^l, frequencies f_m^l, and phases Φ_m^l of the M sinusoidal components, composing a ray spectrum noted $D^l(f)$) by accumulation of the spectral motif $W(f)$ for each component. Real and imaginary parts of the stochastic STS are computed from the amplitude spectral envelope $E^l(f)$ (defined by B subband coefficients $(e_1^l, \ldots, e_B^l)$ on the equivalent rectangular bandwidth scale) multiplied by two noise sequences and circularly convolved with the discrete spectral motif $W[k]$. Real and imaginary parts of the whole STS $X^l[k]$ are the sum of the two respective contributions. The synthetic time-domain signal $x[n]$ is reconstructed after IFFT and OLA processes.

Practically, the filtering process may produce artifacts such as alterations to the transients and/or to the timbre of the original sound. Based on synthesis processes, the proposed synthesizer will benefit of a new spatial extension method that overcomes these problems by using specific properties of the additive modeling, avoiding the tricky issue linked to the filter design.

IV. Novel Architecture for the Synthesizer: A 3-D Immersive Synthesis Engine

In this section, we present a one-stage fast implementation for synthesizing/positioning point-like sound sources (composed of both sinusoidal and noisy components) allowing a significant computational cost reduction per source compared to the traditional two-stage implementation. Then, we present a new method based on the additive modeling for simulating spatially extended sound sources.

A. One-Stage Synthesis/Spatialization

Sound synthesis and spatialization are usually implemented at separated stages of the sound generation. First, monophonic sounds are synthesized for each source. Second, 3-D positional audio algorithms are applied to create point-like virtual sources. Here, we present successively the implementation of additive synthesis by inverse fast Fourier transform (IFFT) and 3-D positional audio methods. IFFT synthesis is to be approximately 10 to 30 times faster than using time-domain oscillator banks for sinusoidal sounds [61] and fulfills the real-time constraints of the synthesizer. Moreover, we take advantage of the fact that IFFT synthesis is computed in the frequency domain to propose a new one-stage architecture that integrates IFFT synthesis and 3-D positional audio modules to reduce the computational cost per source.

1) IFFT Synthesis: IFFT synthesis is a method for synthesizing time-varying filtered noise and sinusoids [13], [62]–[65]. From a theoretical point of view, IFFT synthesis is an implementation of additive synthesis in the time–frequency domain: first an approximation to the short-time Fourier transform (STFT) of the desired signal is computed from the synthesis parameters (frequencies, amplitudes and phases for the deterministic contribution, and spectral envelope for the stochastic contribution) then the inverse STFT is processed to get the synthetic signal. Practically, the deterministic $d[n]$ and stochastic $s[n]$ contributions of a monophonic sound $x[n]$ are constructed in real-time with a frame by frame pattern. Short-time spectra (STS) are built at each frame from the synthesis parameters. STS are inverse fast Fourier transformed, weighted by a synthesis window $w[n]$ and overlap-added (OLA) to get the reconstructed synthetic signal $x[n]$. Practical implementation of the whole process is illustrated in Fig. 2.

IFFT synthesis of the deterministic contribution $d[n]$ has been proposed in [62]. The method cleverly exploits the inherent sparsity of sinusoids in the frequency domain to reduce the computational cost compared to time-domain computation of sinusoids. At each frame l, the M sinusoidal components to be synthesized form a ray spectrum noted $D^l(f)$, where f is the continuous normalized frequency. Each component m is defined by its amplitude a_m^l, frequency f_m^l and phase Φ_m^l parameters. Since the synthetic signal is real-valued in the time domain, its spectrum is conjugate-symmetric in the frequency domain. Thus, we only consider positive frequencies in this document, and ignore their negative counterparts. The sinusoidal signal corresponding to the inverse Fourier transform of $D^l(f)$ is noted $d^l[n]$:

$$d^l[n] = \sum_{m=1}^{M} a_m^l e^{j\left(2\pi f_m^l n + \Phi_m^l\right)}.$$

Synthesizing arbitrary frequencies implies that f_m^l is defined on a continuous scale between 0 and 0.5. Consequently, the computed deterministic STS at frame l cannot be a simple version of $D^l(f)$ discretized on N evenly spaced frequency bins: all components whose frequency is not a multiple of $1/N$ would be lost. In [62], the authors propose to solve this issue by computing the discrete deterministic STS given by

$$(W * D^l)[k] = \sum_{m=1}^{M} a_m^l e^{j\Phi_m^l} W\left(\frac{k}{N} - f_m^l\right) \quad (1)$$

where $W(f)$ (called "spectral motif") is the Fourier transform of the synthesis window $w[n]$, N is the number of frequency bins (i.e., the synthesis window size), k the discrete frequency index (i.e., $W[k] = W(k/N)$). Note that the synthesis window $w[n]$ is assumed to be symmetric in the time domain so that $W(f)$ is real. The constructed STS is the discrete Fourier transform of $w[n]d^l[n]$. As $D^l(f)$ is a ray spectrum, the periodic convolution $\left(W * D^l\right)(f)$, which results in the multiplication $w[n]d^l[n]$ in the time domain, is performed at no cost when constructing the STS, simply by circularly shifting $W(f)$ in the frequency domain.

The crucial factor to reduce the computational complexity is the choice of the synthesis window $w[n]$. If the window's energy is sufficiently concentrated in a narrow frequency band, then the spectral motif can be truncated to K frequency bins (typically $K < 10$) without losing much information. Then, synthesizing N points of a sinusoidal component in the time domain requires only modifying K points of the STS in the frequency domain. It drastically reduces the number of multiplications per component and makes IFFT synthesis significantly more efficient than time-domain oscillator banks [62].

IFFT synthesis of the stochastic contribution $s[n]$ has been proposed in [13], [63], and [64]. Short-time stochastic signals $s^l[n]$ are constructed in the frequency domain. The corresponding STS $S^l[k]$ have a piecewise magnitude spectral envelope $E^l(f)$ and random phases [13]. To take into account perceptual properties of human hearing, $E^l(f)$ is defined as a sequence of B subband coefficients $(e_1^l, \ldots, e_B^l)$ evenly spaced on the equivalent rectangular bandwidth (ERB) scale [63] defined by

$$\mathrm{ERB}(f) = 21.4 \log_{10}\left(4.37\frac{f}{1000} + 1\right)$$

where f is the frequency in Hz [66]. Then the maximum number B of ERB subbands that can be synthesized with a N-point synthesis window is given by

$$B = \left\lfloor \frac{\mathrm{ERB}\left(\frac{f_s}{2}\right)}{\mathrm{ERB}\left(\frac{f_s}{N}\right)} \right\rfloor \quad (2)$$

where f_s is the sampling frequency and $\lfloor . \rfloor$ the floor function. A large window allows a high frequency resolution (i.e., a high number of subbands) but leads to a poor time resolution, inadequate for synthesizing transients. By contrast, a short synthesis window is well adapted to reproduce transients, but only a few subbands can be accurately synthesized. We propose to use different window sizes for synthesizing both sharp temporal envelopes and narrow spectral envelopes. In practice, two synthesis windows of 128 taps and 1024 taps were chosen. Five ERB subbands can be synthesized with the 128-tap window, 26 with the 1024-tap window [see (2)]. ERB subbands may also be approximated by linear subbands in the low frequencies to increase the frequency resolution. Following this principle, eight frequency subbands are synthesized with the 128-tap window and 32 subbands with the 1024-tap window. We found that such time/frequency resolutions are adequate for reproducing a wide variety of environmental sounds (see Section V-A).

The combination of both deterministic and stochastic contributions is realized in the frequency domain. Since the multiplication by the synthesis window $w[n]$ is included in the deterministic STS but not in the stochastic STS, $S^l[k]$ has to be circularly convolved by $W[k]$. As $W[k]$ is nonzero for only K frequency bins, the convolution does not increase the complexity dramatically. Summing deterministic and stochastic contributions results in the STS

$$X^l[k] = (W * \left(D^l + S^l\right))[k].$$

The IFFT is processed on $X^l[k]$ for each frame l. The resulting short-time signals are overlap-added to reconstruct the whole time-domain signal:

$$x[n] = \sum_{l=-\infty}^{\infty} w[n - lL]\left(d^l[n - lL] + s^l[n - lL]\right)$$

where L is the synthesis hop size.

2) 3-D Positioning: Even if each 3-D positional audio method has its own characteristics, a general implementation strategy has been described in [19]. In practice, positioning a point-like sound source is decomposed into three stages. First, the monophonic sound is spatially encoded, i.e., processed through a C-channel filterbank H that depends on the intended virtual direction defined by the azimuth and elevation (θ, Ψ). This stage results in one multichannel signal per sound source. Second, source mixing is realized to obtain a single C-channel signal y. Third, the spatial decoding is performed by matrixing and filtering the C channels of y, to finally determine the contribution of each loudspeaker in the reproduction setup.

For the encoding stage, the spatial filterbank H reduces to a vector of position-dependent gains (called "spatial gains" in this document) for Ambisonics, HOA and amplitude panning techniques. H also reduces to a vector of spatial gains for several multichannel implementations of binaural synthesis [59], [67]. Concerning WFS, H includes position-dependent delays.

For the decoding stage, y is multiplied by a matrix adapted to the loudspeaker setup when using Ambisonics and HOA. For multichannel implementations of binaural synthesis, y is decoded with spatial filters and downmixed to two channels [59], [67]. For amplitude panning and WFS, the decoding is unnecessary: y directly feeds the loudspeakers.

3) Combining IFFT Synthesis and 3-D Positioning: To efficiently reduce the computational cost, we propose the frame-based frequency-domain architecture composed of three stages that combine IFFT synthesis and 3-D audio modules: Synthesis part 1, 3-D positioning and Synthesis part 2. The whole architecture is depicted in Fig. 3 and the stages are described as follows.

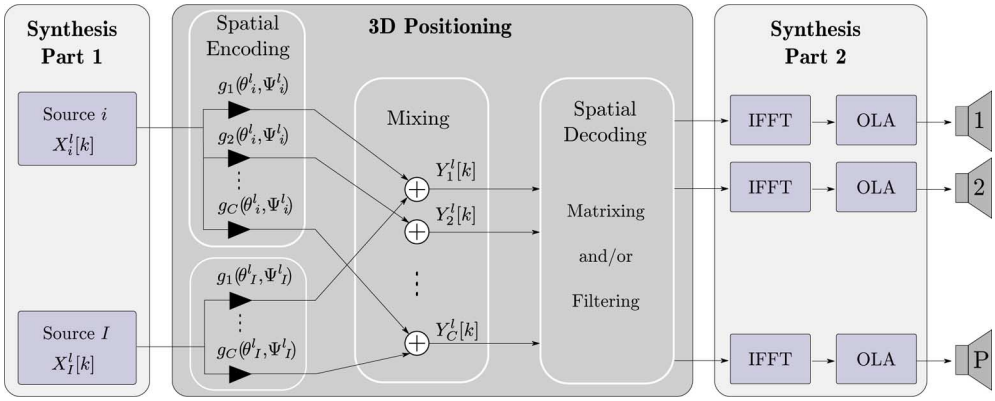

Fig. 3. Fast implementation for generating I point-like sound sources. Synthesis and positioning are combined at the same level of sound generation. To simulate a source i at a given direction defined by the azimuth and elevation (θ_i^l, Ψ_i^l) at frame l, the short-time spectrum $X_i^l[k]$ is build (Synthesis part 1) and spatially encoded by applying spatial gains $(g_1(\theta_i^l, \Psi_i^l), \ldots, g_C(\theta_i^l, \Psi_i^l))$ (Spatial encoding). The same process is realized for other sound sources and all C-channel encoded signals are mixed together (Mixing). The resulting short-time spectrum $(Y_1^l[k], \ldots, Y_C^l[k])$ is matrixed and filtered (Spatial decoding) and finally the IFFT/OLA process is performed (Synthesis part 2) to obtain the signals to feed the P loudspeakers.

- Synthesis part 1: at a given frame, the whole STS of each source (obtained by summing deterministic and stochastic STS) is constructed as described in Section IV-A1.
- 3-D positioning: the spatial encoding is applied to each STS. As presented in Section IV-A2, the encoding depends on the 3-D positioning method. When using Ambisonics, HOA, multichannel binaural and amplitude panning, the encoding consists in multiplying the monophonic sounds by real-valued spatial gains $(g_1, \ldots, g_C)$. Such gains can be applied to the STS in the frequency domain. Regarding WFS, the spatial encoding requires delaying the monophonic sounds. It is more difficult and less computationally efficient to implement such delays per block in the frequency domain: zero-padding is required to avoid circular time aliasing and the whole STS must be multiplied by a linear-phase spectrum. In the present paper, we do not consider WFS and base our architecture on "amplitude-based" 3-D audio methods, which avoid delays in the 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 cth channel is given by

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

  where $X_i^l[k]$ is the STS of the ith source at frame l, g_c is the cth position-dependent spatial gain, (θ_i^l, Ψ_i^l) is the position of the ith source at frame l, and I is the total number of sound sources. Then, the spatial decoding is performed by matrixing and/or filtering the C channels of Y^l, depending on the 3-D audio method (see Section IV-A2).
- Synthesis part 2: IFFT and OLA are processed after the decoding stage for each loudspeaker channel. Note that with Ambisonics and HOA the number of loudspeakers P can be higher than the number of internal channels C. In that case only, the IFFT/OLA process is performed before the spatial decoding on the internal channels.

B. Spatial Extension Based on the Additive Modeling

Extended sound sources can be simulated by positioning several secondary sources at different locations (see Section III-C). Compared to existing methods [56]–[60] our approach bypasses the filtering process so that the transients and the timbre of the original sound are well preserved. Here we propose new possibilities for computing decorrelated secondary sources from the synthesis parameters of the original sound. For the stochastic contribution, decorrelation is achieved by synthesizing the STS of each secondary source with the same original spectral envelope but with different noise sequences. For the deterministic contribution, the STS are synthesized with the same original amplitude and frequency parameters, but random phases are generated at $t = 0$ for each sinusoidal component. This way, deterministic secondary sources are effectively decorrelated if they contain a sufficient number of components. The proposed method allows computing an unlimited number of decorrelated signals and it preserves the transients and the timbre of the original sound.

Using this decorrelation method, we propose a spatial extension effect in the synthesizer by using a maximum of eight virtual secondary sources evenly spaced on a circle surrounding the listener (see Fig. 4). The spatial extension parameter Ω controls the relative contributions of the eight sources via a set of "extension gains" $(u_1, \ldots, u_8)$. For instance, for reproducing a point source $(\Omega = 0\%)$ u_1 is set to 1 while other gains are set to 0. For simulating a completely diffused sound surrounding the listener $(\Omega = 100\%)$ all extension gains are set to 1. The first secondary source (with gain u_1) is positioned at the intended virtual position. Other secondary sources are controlled by pairs (i.e., $u_2 = u_3$, $u_4 = u_5$ and $u_6 = u_7$). When Ω increases from 0 to 100%, the number of secondary sources with a nonzero extension gain increases gradually from one to eight: first u_2 increases to 1, then u_4 increases to 1, followed by u_6 and finally u_8. To ensure that the energy of the extended source is constant $\forall\, \Omega$, extension gains are scaled by the factor

$$\gamma = \sqrt{\sum_{i=1}^{8} u_i^2(\Omega)}.$$

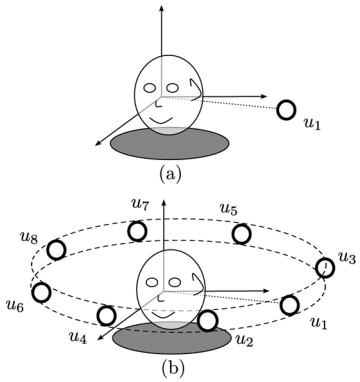

Fig. 4. Simulation of an extended sound source. (a) A single virtual source produces a narrow auditory event. (b) Positioning several decorrelated copies of the original sound at different locations allows producing an extended auditory event. The spatial extension parameter Ω weights the contribution of eight secondary sources by gains $(u_1, \ldots, u_8)$ to control the apparent source width.

A perceptual evaluation of the parameter Ω was conducted to validate the accuracy of this method for simulating spatially extended sound sources (see Section V-B).

C. Efficiency

In contrast with a classical two-stage implementation (synthesis of monophonic sources then spatialization) that requires one IFFT per source, the proposed novel architecture requires one IFFT per loudspeaker, independently of the number of sources, and becomes a serious advantage in the case of auditory environments with hundreds of sound sources.

The algorithm is particularly attractive when using a multichannel implementation of binaural synthesis for 3-D audio rendering on headphones (e.g., for mobile-phone applications). First, because the spatial decoding requires linear filtering that can be efficiently applied in the frequency domain: multiplying the channels of Y^l with the discrete Fourier transform of the spatial filters results in a fast implementation of the time-domain circular convolution. Note that the synthesis window $w[n]$ and the impulse responses of the spatial filters are properly zero-padded before the calculation of their Fourier transforms to avoid time aliasing. Second, since the C channels of Y^l are downmixed to two binaural signals after the spatial filtering, only two IFFT are computed whatever the number of sound sources.

The proposed architecture is also efficient to synthesize/spatialize sound sources with a small number M of sinusoidal components, and/or to position components individually in space. Indeed, it is not necessary to apply the spatial gains to the entire STS (of length $N/2$) to spatialize one sinusoid; the multiplication of the spectral motif (of length K) is sufficient. Thus, the computational cost per channel is reduced from N to $2K$ real multiplications per component. Consequently, if $2KM < N$, it is more efficient to apply the spatial gains directly on the motif for each sinusoid rather than multiplying the entire STS of the source once. In practice, based on (1), the spatially encoded C-channel STS for a given source at frame l is constructed at once and is expressed by

$$\left(W * D_c^l\right)[k] = \sum_{m=1}^{M} g_c\left(\theta_m^l, \Psi_m^l\right) a_m^l e^{j\Phi_m^l} W\left(\frac{k}{N} - f_m^l\right)$$

where $g_c(\theta_m^l, \Psi_m^l)$ is the cth position-dependent spatial gain for component m. In the same way, the stochastic contribution can be spatially encoded by applying spatial gains directly to the B subband coefficients $(e_1^l, \ldots, e_B^l)$ of the spectral envelope. Note that for a point-like sound source, the spatial gains are identical for all components and need to be computed only once.

V. Subjective Evaluation of the Synthesizer

The possibilities offered by the synthesizer have been perceptually evaluated. We verified by informal listening tests that it allows synthesizing/positioning accurately point-like environmental sources. We conducted a formal listening test to validate the method described previously to control the spatial extension of sound sources.

A. Synthesis of Positioned Sound Sources

In this section, the 3-D immersive synthesizer is used to simulate typical examples of environmental sounds and the analysis/synthesis process is detailed for each sound category (i.e., vibrating solids, liquids, and aerodynamics). A complete analysis/transformation/synthesis system is presented in [13] for the additive signal model. In the analysis stage, the time-varying parameters associated with the deterministic contribution (i.e., amplitude, phase and frequency of predominant sinusoidal components) are first estimated from a time–frequency decomposition of the original sound. Then, the deterministic contribution is removed from the original sound to quantify the stochastic residual that is modeled by its average energy in subbands [13], [63]. For our concern, we adapted these analysis techniques and combined them with physically based approaches for generating several types of environmental sounds (impacts, wind, fire, whoosh, sea waves, etc.). We refer the reader to [32] where sound examples can be found.

1) Vibrating Solids: Impact sounds are efficiently simulated by a sum of exponentially decaying sinusoids [3], [7]

$$x(t) = \sum_{m=1}^{M} e^{-\alpha_m t} a_m cos(2\pi f_m t + \Phi_m). \quad (3)$$

From a physical point of view, the frequencies f_m correspond to the modal frequencies that characterize the shape of the impacted object; the amplitudes a_m depend on the excitation point; the decay factors α_m are mainly characteristic of the object's material. In our case, these modal parameters were estimated by signal analysis of recorded impact sounds. Modal frequencies can also be replaced by bands of noise to synthesize impact sounds with a high modal density (with overlapping modes) without altering the perceptual rendering [69].

The attack of real impact sounds can be very short (less than a few milliseconds). Synthesizing such short transients is a challenge for the synthesizer because the IFFT technique assumes that the sound characteristics (frequency and amplitude)

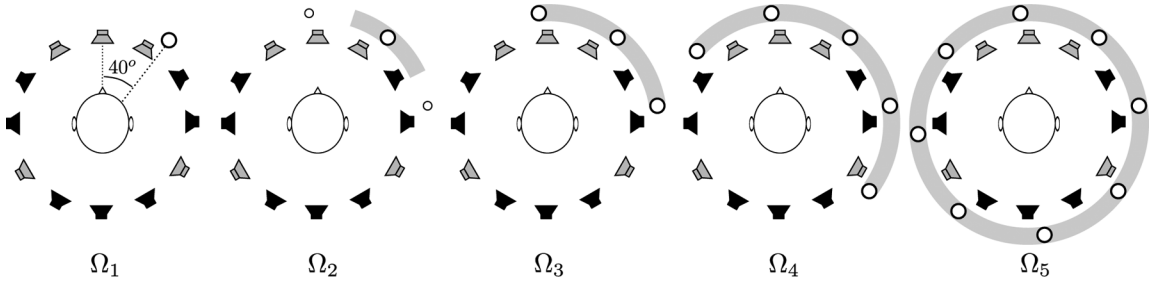

Fig. 5. Design of the five spatial extensions ($\Omega_1, \ldots, \Omega_5$) evaluated in the listening test. The original point-like sound source positioned at 40° on the right of the listener is given by Ω_1. The four spatially extended versions of the original source were created by using three secondary sources (white circles) with different gains (Ω_2), three sources with equal gain (Ω_3), five sources (Ω_4), and eight sources (Ω_5). The gray area represents the intended perceived source width in each case. The figure also shows the two reproduction systems used in the test: a standard 5.0 loudspeaker setup (gray) and a binaural downmix of 12 virtual loudspeakers (gray + black) for headphones.

vary slowly in time compared to the block size. In [70], the authors propose a scalable approach using different synthesis window sizes to address this issue. For our concern, we use a short synthesis window for producing sounds that contain transients. Informal listening tests have shown that a 128-tap window (equivalent to 3 ms at 44 100 Hz) is sufficiently short for convincing impact sounds. Furthermore, it introduces a minor delay which is desirable for low-latency interactive applications such as video games.

We are currently investigating the generation of sounds made by other types of solid interactions. Physically based simulation of solids usually decomposes the vibrating system into an excitation function and modal resonators. The impact sound model given by (3) corresponds to an impulse excitation. Other excitation functions can produce rolling and rubbing sounds [7]. Since the synthesizer is based on a frame by frame synthesis approach, the excitation function must be discretized and it is assumed to be constant within a frame. Informal listening tests have shown that convincing rolling sounds can be synthesized with a hop size sufficiently short (e.g., 32 samples).

2) Aerodynamic Sounds: In [10] and [11], Dobashi *et al.* use computational fluid dynamics to precompute sound textures that are then played in real-time at different speeds related to the fluid velocity. Their method can produce realistic sounds like swinging swords, wind blowing, and fire sounds. For our concern, since such sounds are usually very noisy, the analysis/synthesis approach based on the stochastic subband modeling proposed in [63] is very promising. To synthesize relatively stationary sounds such as wind, we use 32 subbands quasi-evenly spaced on the ERB scale with a 1024-tap analysis/synthesis window. To synthesize fast transients, such as fire crackling, we use only eight subbands with a 128-tap analysis/synthesis window.

3) Liquid Sounds: A physically based model for liquid sound synthesis was proposed in [12]. Bubble sounds are simulated as swept sinusoids whose amplitude exponentially decays in time. A stochastic model is used to excite a population of bubbles of different sizes to create complex liquid stream. We use this technique for synthesizing drops and bubble-like sounds with the synthesizer. Some liquid sounds, for instance sea waves, are perceptually closer to broadband noise than bubbles. For such sounds we use the same analysis/synthesis approach than for aerodynamic sounds with the stochastic subband modeling.

B. Synthesis of Extended Sound Sources

This section presents a formal subjective evaluation of the spatial extension control provided in the synthesizer for simulating spatially extended sound sources (see Section IV-B).

1) Experimental Setup and Subjects: The test was performed on two reproduction systems: Loudspeakers and Headphones. The first system was a standard 5.0 loudspeaker setup. The second was a binaural downmix for headphones with 12 virtual loudspeakers evenly spaced around the subject (implemented with generic HRTFs). Both systems were restricted to 2-D sound spatialization (see Fig. 5). We used VBAP [18] to calculate the spatial gains in the 3-D positioning stage of the synthesizer (cf. Section IV-A3).

Two groups of 20 participants took part in the experiment: the first group (15 men, 5 women, 30 years old on average) was tested with Loudspeakers and the second group (17 men, 3 women, 35 years old on average) with Headphones. Eight people belonged to both groups. They were volunteered students and engineers involved in audio or acoustics at Orange Labs or CNRS-LMA. They reported no hearing problem.

2) Stimuli: We designed four sound items representative of the main categories of environmental sounds (cf. Section II): bells for vibrating solids, sea waves and drops of water for liquid sounds, wind for aerodynamic sounds. First, the sounds were synthesized/spatialized at 40° on the right of the listener to create a point-like source (see Fig. 5). Then we used the decorrelation method described in Section IV-B to create spatially extended versions of the source. Since the decorrelation method differs for deterministic and stochastic parts, we equilibrated the items as follows: two items contained only time-varying filtered noise (sea waves and wind) and two items contained only sinusoidal components (bell and drops of water).

For each item, five spatial extensions were computed by acting on the relative contributions of the eight secondary sources as follows.

- $\Omega_1 : u_i = 1$ for $i = 1$ and 0 otherwise.
- $\Omega_2 : u_1 = \sqrt{10}/2\sqrt{3}$, $u_2 = u_3 = 1/2\sqrt{3}$, $u_i = 0$ for $i = 4, \ldots, 8$.
- $\Omega_3 : u_i = 1/\sqrt{3}$ for $i = 1, \ldots, 3$ and 0 otherwise.
- $\Omega_4 : u_i = 1/\sqrt{5}$ for $i = 1, \ldots, 5$ and 0 otherwise.
- $\Omega_5 : u_i = 1/\sqrt{8}$ for $i = 1, \ldots, 8$.

3) Procedure: For both rendering systems, the subjective evaluation of the spatial extension was conducted by a paired

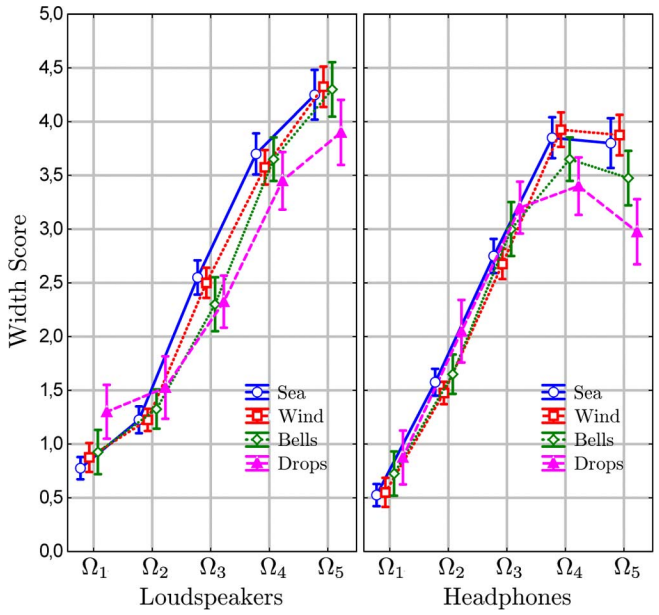

Fig. 6. Width scores averaged across the 20 subjects participating to the listening test. The scores are depicted for five spatial extensions ($\Omega_i, i = 1 \ldots 5$) for two reproduction systems (Loudspeakers and Headphones) and four sound items (Sea, Wind, Bells, Drops). Vertical bars represent the 95% confidence intervals.

comparison experiment. We used the "Subjective Training and Evaluation Program" by Audio Research Labs for A-B comparison. A total of 40 pairs (10 pairs corresponding to all the combinations between the five extended versions for each of the four items) was presented in a random order. Each pair A-B was composed of two different extended versions of a same item (i.e., two different items were not directly compared) so that Ω is the only changing parameter between A and B.

Participants were asked to listen to each pair A-B (as often as they wanted) and to evaluate the difference of perceived widths by choosing one of the 3 following possibilities: "A is wider than B," "A and B have the same width," "B is wider than A." A training session was performed. The whole test lasted about 20 min on average.

4) Results: Data were collected into 5×5 matrices Δ_p^q for each participant p and for each item q. Each matrix was completed as follows: if the participant judged that sounds A (Ω_i) and B (Ω_j) had the same width, then the cell $\Delta_p^q(i,j)$ was set to 0.5; if sound A was judged wider than sound B, then $\Delta_p^q(i,j)$ was set to 1; if sound B was judged wider than sound A, then $\Delta_p^q(j,i)$ was set to 1. Since identical pairs A-A were not tested, the diagonal of Δ_p^q was set to 0.5 by default. As indicated in [71], each matrix Δ_p^q was converted to construct a one-dimensional subjective scale, i.e., a 5×1 vector of scores, by computing the sum value of each column of Δ_p^q. The scores (obtained for each participant p for each item q) reveal the relative perceived source widths for the five spatial extensions (Ω_i, $i = 1 \ldots 5$). Width score values averaged across participants are illustrated on Fig. 6.

The width scores were submitted to repeated measures analysis of variance (ANOVA) including "Extension" $(\Omega_1, \ldots, \Omega_5)$, "Item" (sea, wind, bells, and drops) as within-subject factors and "System" (Loudspeakers versus Headphones) as between-subject factor. When interactions between two or more factors were significant, post-hoc comparisons (Tukey tests) were computed. Results revealed a main effect of the Extension factor ($F(4,152) = 745.96$, $p < .001$). Moreover, the Extension by System interaction was significant ($F(4,152) = 24.19$, $p < .001$): all extensions Ω_i differed from each other for both systems ($p < .01$), except between Ω_4 and Ω_5 for the Headphones system ($p = 0.733$). The Item by Extension interaction was also significant ($F(12,456) = 8.59$, $p < .001$): the Extension effect was smaller for Drops than for the three other items, in particular for extreme values (between Ω_1 and Ω_2 and between Ω_4 and Ω_5).

5) Discussion: The analysis revealed that the spatial extension parameter Ω has a relevant action on the perceived width: independently of the sound items, the perceived width increases with respect to Ω. In particular, the difference of source width between the different Ω_i was well perceived on both reproduction systems, except between Ω_4 and Ω_5 on Headphones. In addition, the source extension was well perceived for all type of Items with the smallest score range for Drops. Thus, these results show that the control of spatial extension in the synthesizer performs well globally on classical reproduction systems as well as for all environmental sound categories.

These results also revealed some limitations of the control parameter. The first one concerns the simulation of extremely wide sources on Headphones that is revealed by a higher score for Ω_4 than for Ω_5 (but they are not statistically different). In other words, our control fails to reproduce sources larger than 180° on Headphones by contrast with Loudspeakers. Several reasons could explain this restriction. First, a main difference between Loudspeakers and Headphones is that subjects could freely move their head to explore the soundfield on Loudspeakers while subjects' head was virtually immobilized on Headphones. It is known that head movements are useful to qualitatively perceive the borders of an extended source [72]. Since head movements were not allowed on Headphones, the evaluation of subjects was slightly limited compared with Loudspeakers condition. Second, a widely extended source (like Ω_5 condition) is not properly externalized on Headphones: in practice, the most extended source in Ω_5 condition may be perceived inside the head and consequently, its spatial width is underestimated. Note that the source in Ω_4 condition may be better externalized because it was lateralized. Finally, Headphones introduce a front–back confusion (a typical artifact of binaural synthesis, especially when head movements are not allowed [73]) that may have also blurred width assessment. We think that using a head-tracking system on Headphones could compensate these artifacts, i.e., take into account head movements, better externalize the source and reduce the front–back confusion. In this way, results on Headphones may be improved so that to be similar to the ones on Loudspeakers.

A second limitation may be due to the decorrelation method that is unsuitable to efficiently extend sources with a small number of sinusoidal components (see Section IV-C). This is precisely the case of the Drops item that contains only about ten simultaneous sinusoids. Its score range is significantly reduced (i.e., the score curve is compressed compared to the curves

for the other items; see Fig. 6). For comparison, the Bells item contains 65 sinusoidal components and the corresponding score range is much larger. An informal test confirmed that the spatial extension is more accurate when the number of components increases. However, this limitation is not a crucial aspect since from a conceptual point of view, a Drop sound (and more generally a sound containing few components) is not devoted to be spatially extended as widely as a Sea or Fire sound. Nevertheless, to simulate large spatial extension with a few scattered drops, we suggest positioning them individually at different locations distributed on a wide space (each drop being considered as an independent sound source in this case).

VI. Sound Effect Possibilities

Combining both synthesis and spatialization processes at the same level of sound generation opens new ways of addressing the design of spatial sound effects. Since each sound component is defined by a set of parameters characterizing its timbre (synthesis parameters) and its spatial attributes (position parameters), sound manipulations may act on these parameters in a global way rather than in a separate way. One may for instance propose sound effect processes linking intimately the timbre and the spatial distribution of the sound.

Several approaches are currently under investigation. Based on advantages of the 3-D immersive synthesizer architecture, a first approach concerns the independent positioning of sinusoidal and/or noisy components at specific locations with possibly multiple locations for a same component (see Section IV-C). This approach is very attractive to simulate a "spatial distribution effect" and perceptual effects of spreading the sinusoidal components in space have been studied in a musical context [28]. Informal listening tests have shown that spatializing sound components at different locations tends to produce broad spatial images. A second approach under investigation is the definition of a specific spatial trajectory for each component, allowing perceptual time-varying spreading effects. A third approach consists in modifying the timbre of the source as a function of its spatial position. This can lead to the simulation of physical effects (due for instance to the air absorption) but also to unnatural sound effects such as transforming the perceptual properties of a sound source (for instance, the nature of the perceived struck material) according to specific localization rules. In particular, it is worth noting that the synthesizer can be easily extended to simulate a "directivity effect," i.e., sound sources with directional radiation. When building the STS, spectral components can be weighted by "directivity gains" depending on the relative orientation between the source and the listener. The source-listener distance can also be simulated by an additional "distance gain" per source, producing for example a 6-dB attenuation of the sound when doubling the distance.

Finally, a last investigation concerns the spatialized sound morphing, consisting in a process that acts both on timbre and on spatial properties. For example, this sound effect would allow creating an initial point-like sound source in space, from which some components could "escape" to constitute individual sources (with closely related or different timbre attributes) having their own trajectory in space.

VII. Conclusion

In this paper, we addressed the design of a real-time 3-D immersive synthesizer for environmental sounds in the framework of virtual reality applications. The implementation is based on a novel architecture that merges together synthesis, spatialization, and spatial extension methods and is compatible with several 3-D audio formats (multichannel, Ambisonics, binaural). In particular, we proposed a one-stage synthesis/spatialization approach that combines frequency-domain additive synthesis and amplitude-based positioning at the prime level of sound generation. This approach reduces the computational cost per source since it requires only one IFFT per channel as opposed to one IFFT per source. Thus, the synthesizer provides an efficient tool to create complex scenes implying several sound sources in the 3-D space. It is particularly attractive for interactive mobile applications using binaural rendering over headphones (only two IFFT are required per frame, whatever the number of sound sources). We also presented a new method to spatially extend sound sources, that uses specific properties of the additive modeling to avoid filtering artifacts.

Formal and informal listening tests assessed the capabilities of the synthesizer to reproduce realistic environmental sounds from the main categories (wind, liquids, impacts, etc.) and accurately control their perceived width. Sound examples are available online [32]. Based on advantages of the unified architecture, the synthesizer offers numerous sound effect possibilities, such as spatialized sound morphing (i.e., modifying timbre, position and spatial width at once in an interactive way) that could hardly be reproduced by a classical two-stage implementation.

We are currently investigating the other possibilities of the synthesizer by designing a generic control interface. Especially, we aim at providing high-level controls for each category of environmental sounds. We further expect to validate these controls by perceptual tests. We are also investigating the integration of sound reverberation at the prime level of sound generation, and experimenting for the creation of original spatial effects. For that purpose, we will base our exploration both on musicians and sound designers advises to further clarifying their practical needs.

Acknowledgment

The authors would like to thank Dr. P. Depalle for advised suggestions on inverse FFT synthesis, Drs. J. Daniel and M. Emerit for helpful discussions on sound spatialization, Dr. J. Faure for fruitful comments on statistical analysis, T. Voinier, and all people who gracefully accepted to participate to the perceptual evaluation of the synthesizer.

References

[1] C. Roads, *The Computer Music Tutorial*. Cambridge, U.K.: MIT Press, 2000.

[2] R. Kronland-Martinet, P. Guillemain, and S. Ystad, "Modelling of natural sounds by time-frequency and wavelet representations," *Organised Sound*, vol. 2, no. 3, pp. 179–191, 1997.

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

[4] A. Farnell, *Designing Sound, Procedural Audio for Games and Film*. London, U.K.: Applied Scientific, 2008.

[5] D. Rocchesso and F. Fontana, "The Sounding Object," 2003 [Online]. Available: http://www.soundobject.org/

[6] "CLOSED Project," [Online]. Available: http://closed.ircam.fr/

[7] K. van den Doel, P. G. Kry, and D. K. Pai, "Foleyautomatic: physically-based sound effects for interactive simulation and animation," in *Proc. 28th Annu. Conf. Comput. Graphics Interactive Techniques*, 2001, pp. 537–544.

[8] J. F. O'Brien, C. Shen, and C. M. Gatchalian, "Synthesizing sounds from rigid-body simulations," in *Proc. 2002 ACM SIGGRAPH/Eurographics Symp. Comput. Animation*, 2002, pp. 175–181.

[9] N. Raghuvanshi and M. C. Lin, "Interactive sound synthesis for large scale environments," in *Proc. 2006 Symp. Interactive 3-D Graphics Games*, 2006, pp. 101–108.

[10] Y. Dobashi, T. Yamamoto, and T. Nishita, "Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics," *ACM Trans. Graphics (Proc. SIGGRAPH 2003)*, vol. 22, no. 3, pp. 732–740, 2003.

[11] Y. Dobashi, T. Yamamoto, and T. Nishita, "Synthesizing sound from turbulent field using sound textures for interactive fluid simulation," in *Proc. EUROGRAPHICS*, 2004, vol. 23, no. 3, pp. 539–546.

[12] K. van den Doel, "Physically-based models for liquid sounds," in *Proc. ICAD 04-10th Meeting Int. Conf. Auditory Display*, 2004.

[13] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, pp. 12–24, 1990.

[14] X. Rodet, "Musical sound signal analysis/synthesis: Sinusoidal + residual and elementary waveform models," in *Proc. IEEE Time-Frequency and Time-Scale Workshop (TFTS'97)*, 1997.

[15] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[16] T. Funkhouser, J.-M. Jot, and N. Tsingos, "Sounds good to me! computational sound for graphics, virtual reality, and interactive systems," in *SIGGRAPH 2002 Course Notes*, 2002.

[17] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging," in *Proc. 114th AES Conv.*, 2003.

[18] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *JAES*, vol. 45, no. 6, pp. 456–466, 1997.

[19] 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.

[20] P. G. Georgiou and C. Kyriakakis, "A multiple input single output model for rendering virtual sound sources in real time," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, 2000, pp. 253–256.

[21] T. Moeck, N. Bonneel, N. Tsingos, G. Drettakis, I. Viaud-Delmon, and D. Aloza, "Progressive perceptual audio rendering of complex scenes," in *Proc. ACM SIGGRAPH Symp. Interactive 3-D Graphics and Games*, 2007.

[22] N. Peters, T. Matthews, J. Braasch, and S. McAdams, "Spatial sound rendering in max/msp with vimic," in *Proc. Int. Comput. Music Conf. (ICMC)*, 2008.

[23] C. Ramakrishnan, J. Gossmann, and L. Brümmer, "The zkm klangdom," in *Proc. 2006 Int. Conf. New Interfaces Musical Express.(NIME06)*, 2006.

[24] J. Nixdorf and D. Gerhard, "Real-time sound source spatialization as used in challenging bodies: Implementation and performance," in *Proc. 2006 Int. Conf. New Interfaces Musical Express. (NIME06)*, 2006.

[25] D. L. James, J. Barbič, and D. K. Pai, "Precomputed acoustic transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources," *ACM Trans. Graphics (Proc. SIGGRAPH 2006)*, vol. 25, no. 3, pp. 987–995, 2006.

[26] R. Corbett, K. van den Doel, J. E. Lloyd, and W. Heidrich, "Timbrefields—3D interactive sound models for real-time audio," *Presence: Teleoperators and Virtual Environments*, vol. 16, no. 6, 2007.

[27] O. Warusfel and N. Misdariis, "Sound source radiation synthesis: From stage performance to domestic rendering," in *Proc. 116th AES Conv.*, 2004.

[28] D. Topper, M. Burtner, and S. Serafin, "Spatio-operational spectral (S.O.S.) synthesis," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFx'02)*, 2002.

[29] Q. Zhang and J. Shi, "Progressive sound rendering in multimedia applications," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2004.

[30] C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone, "A spatialized additive synthesizer," in *Proc. Inaugural Int. Conf. Music Commun. Science (ICoMCS)*, 2007.

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

[32] [Online]. Available: www.lma.cnrs-mrs.fr/~kronland/spatsynthIEEE/index.html

[33] N. J. Vanderveer, "Ecological acoustics: Human perception of environmental sounds," Ph.D.dissertation, Georgia Inst. Technol., Atlanta, GA, 1979.

[34] B. Gygi and V. Shafiro, "General functions and specific applications of environmental sound research," *Frontiers in Biosci.*, vol. 12, pp. 3152–3166, 2007.

[35] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Percept. Psychophys.*, vol. 69, no. 6, pp. 839–855, 2007.

[36] M. M. Marcell, D. Borella, M. Greene, E. Kerr, and S. Rogers, "Confrontation naming of environmental sounds," *J. Clinical Experimental Neuropsychol.*, vol. 22, no. 6, pp. 830–864, 2000.

[37] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *J. Experimental Psychol.: Human Percept. Perform.*, vol. 19, no. 2, pp. 250–267, 1993.

[38] W. W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception," *Ecol. Psychol.*, vol. 5, no. 1, pp. 1–29, 1993.

[39] W. W. Gaver, "How do we hear in the world? explorations in ecological acoustics," *Ecol. Psychol.*, vol. 5, no. 4, pp. 285–313, 1993.

[40] J. J. Gibson, *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin, 1979.

[41] B. L. Giordano and S. McAdams, "Material identification of real impact sounds: Effects of size variation in steel, wood, and Plexiglas plates," *J. Acoust. Soc. Amer.*, vol. 119, no. 2, pp. 1171–1181, 2006.

[42] S. Tucker and G. J. Brown, "Investigating the perception of the size, shape and material of damped and free vibrating plates," Dept. Comput. Sci.. Univ. de Sheffield, 2002, Tech. Rep. CS-02-10.

[43] Y. Gérard, "Mémoire Sémantique et Sons de L'environnement (Semantic Memory and Environmental Sounds)," Ph.D. dissertation, Bourgogne Univ., Dijon, France, 2004.

[44] P. Schaeffer, *Traité Des Objets Musicaux*. Paris, France: Seuil, 1966.

[45] A. Merer, S. Ystad, R. Kronland-Martinet, and M. Aramaki, "Semiotics of Sounds Evoking Motions: Categorization and Acoustic Features," in *Computer Music Modeling and Retrieval. Sense of Sounds*. Berlin/Heidelberg, Germany: Springer, 2008, pp. 139–158.

[46] M. V. Mathews, J. E. Mille, F. R. Moore, J. R. Pierce, and J. C. Risset, *The Technology of Computer Music*. Cambridge, MA: MIT Press, 1969.

[47] C. Stoelinga and A. Chaigne, "Time-domain modeling and simulation of rolling objects," *Acustica united with Acta Acustica*, vol. 93, no. 2, pp. 290–304, 2007.

[48] J. M. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *JAES*, vol. 21, no. 7, pp. 526–534, 1973.

[49] M. L. Brun, "Digital waveshaping synthesis," *JAES*, vol. 27, no. 4, pp. 250–266, 1979.

[50] M. Aramaki, R. Kronland-Martinet, T. Voinier, and S. Ystad, "A percussive sound synthetizer based on physical and perceptual attributes," *Comput. Music J.*, vol. 30, no. 2, pp. 32–41, 2006.

[51] G. Peeters, "Modèles et modélisation du signal sonore adaptés à ses caractéristiques locales," Ph.D. dissertation, Univ. Paris VI, Paris, France, 2001.

[52] D. R. Perrott and T. N. Buell, "Judgments of sound volume: Effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise," *JASA*, vol. 72, no. 5, pp. 1413–1417, 1982.

[53] R. Mason, T. Brookes, and F. Rumsey, "Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli," *JASA*, vol. 117, no. 3, pp. 1337–1350, 2005.

[54] T. Hirvonen and V. Pulkki, "Perceived spatial distribution and width of horizontal ensemble of independent noise signals as function of waveform and sample length," in *Proc. 124th AES Conv.*, 2008.

[55] J. Blauert, *Spatial Hearing*. Cambridge, MA: MIT Press, 1997.

[56] G. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Comput. Music J.*, vol. 19, no. 4, pp. 71–87, 1995.

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

[58] A. Sibbald, "Method of Synthesizing an Audio Signal," U.S. patent US 6,498,857 B1, Dec. 2002.

[59] J.-M. Jot, M. Walsh, and A. Philp, "Binaural simulation of complex acoustic scene for interactive audio," in *Proc. 121th AES Conv.*, 2006.

[60] V. Pulkki, "Spatial sound reproduction with directional audio coding," *JAES*, vol. 55, no. 6, pp. 503–516, 2007.

[61] X. Rodet and D. Schwarz, "Spectral Envelopes and Additive + Residual Analysis/Synthesis," in *Analysis, Synthesis, and Perception of Musical Sounds: Sound of Music*. New York: Springer, 2007, pp. 175–227.
[62] X. Rodet and P. Depalle, "Spectral envelopes and inverse FFT synthesis," in *Proc. 93rd AES Conv.*, 1992.
[63] M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996.
[64] A. Freed, "Real-time inverse transform additive synthesis for additive and pitch synchronous noise and sound spatialization," in *Proc. 104th AES Conv.*, 1998.
[65] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, *DAFX: Digital Audio Effects*. New York: Wiley, 2002, pp. 373–438.
[66] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, Nov. 1999.
[67] J.-M. Jot, S. Wardle, and V. Larcher, "Approaches to binaural synthesis," in *Proc. 105th AES Conv.*, 1998.
[68] A. Misra, P. R. Cook, and G. Wang, "A new paradigm for sound design," in *Proc. Int. Conf. Digital Audio Effects (DAFx06)*, 2006.
[69] M. Aramaki and R. Kronland-Martinet, "Analysis-synthesis of impact sounds by real-time dynamic filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 695–705, Mar. 2006.
[70] N. Bonneel, G. Drettakis, N. Tsingos, I. Viaud-Delmon, and D. James, "Fast modal sounds with scalable frequency-domain synthesis," *ACM Trans. Graphics (Proc. SIGGRAPH 2008)*, vol. 27, no. 3, 2008.
[71] H.-A David, *The Method of Paired Comparisons*. New York: Oxford Univ. Press, 1988.
[72] C. Kim, R. Mason, and T. Brookes, "An investigation into head movements made when evaluating various attributes of sound," in *Proc. 122th AES Conv.*, 2007.
[73] J.-M. Pernaux, M. Emerit, J. Daniel, and R. Nicol, "Perceptual evaluation of static binaural sound synthesis," in *Proc. AES 22nd Int. Conf. Virtual, Synth., Entertain. Audio*, 2002.

Charles Verron (S'09) received the M.S. degree in acoustics, signal processing, and computer sciences applied to music from University Paris 6, Paris France, in 2004. He is currently pursuing the Ph.D. degree at Orange Labs, Lannion, France, and at the Laboratoire de Mécanique et d'Acoustique, Marseille, France.

He was a Research Assistant in 2006 in the Computer and Audio Research Laboratory, Sydney University, Sydney, Australia, where he contributed to 3-D audio projects. His main research interests include sound synthesis and sound spatialization for music and virtual reality applications.

Mitsuko Aramaki (M'09) received the M.S. degree in mechanics (speciality in acoustic and dynamics of vibrations) from the University of Aix-Marseille II, Marseille, France, in 1999 and the Ph.D. degree for her work at the Laboratoire de Mécanique et d'Acoustique, Marseille, France, in 2003, on analysis and synthesis of impact sounds using physical and perceptual approaches.

She is currently a Researcher at the Mediterranean Institute for Cognitive Neuroscience, Marseille, where she works on a pluridisciplinary project combining sound modeling, perceptual and cognitive aspects of timbre, and neuroscience methods, in the context of virtual reality.

Richard Kronland-Martinet (SM'10) received the M.S. degree in theoretical physics and the Ph.D. degree in acoustics from the University of Aix-Marseille II, Marseille, France, in 1980 and 1983, respectively, and the "Doctorat d'Etat es Sciences" degree in 1989 from the University of Aix-Marseille II for his work on analysis and synthesis of sounds using time–frequency and time-scale (wavelets) representations.

He is currently a Director of Research at the National Center for Scientific Research (CNRS), Laboratoire de Mécanique et d'Acoustique, Marseille, where he is the Head of the group "Modeling, Synthesis, and Control of Sound and Musical Signals." His primary research interests are in analysis and synthesis of sounds with a particular emphasis on high-level control of synthesis processes. He recently addressed applications linked to musical interpretation and semantic description of sounds using a pluridisciplinary approach associating signal processing, physics, perception, and cognition.

Grégory Pallone received the Ph.D. degree in acoustics, signal processing, and computer music prepared at the Laboratoire de Mécanique et d'Acoustique, Marseille, France, in 2003.

During the Ph.D. degree, he was with Genesis, Inc., a French company specialized in acoustics. He is currently an Audio Research Engineer at Orange Labs, Lannion, France. His main interests are in 3-D sound, sound synthesis, audio signal processing, spatialized communications, and audio coding.