Spatialized Synthesis of Noisy Environmental Sounds

Charles Verron[1,4], Mitsuko Aramaki[2,3], Richard Kronland-Martinet[4], and Grégory Pallone[1]

[1] Orange Labs, OPERA/TPS,
Avenue Pierre Marzin, 22307 Lannion, France
{charles.verron;gregory.pallone}@orange-ftgroup.com
[2] CNRS - Institut de Neurosciences Cognitives de la Méditerranée,
31, chemin Joseph Aiguier 13402 Marseille Cedex 20, France
[3] Aix-Marseille - Université,
58, Bd Charles Livon 13284 Marseille Cedex 07, France
aramaki@incm.cnrs-mrs.fr
[4] CNRS - Laboratoire de Mécanique et d'Acoustique,
31, chemin Joseph Aiguier 13402 Marseille Cedex 20, France
kronland@lma.cnrs-mrs.fr

Abstract. In this paper, an overview of the stochastic modeling for analysis/synthesis of noisy sounds is presented. In particular, we focused on the time-frequency domain synthesis based on the inverse fast Fourier transform (IFFT) algorithm from which we proposed the design of a spatialized synthesizer. The originality of this synthesizer remains in its one-stage architecture that efficiently combines the synthesis with 3D audio techniques at the same level of sound generation. This architecture also allowed including a control of the source width rendering to reproduce naturally diffused environments. The proposed approach led to perceptually realistic 3D immersive auditory scenes. Applications of this synthesizer are here presented in the case of noisy environmental sounds such as air swishing, sea wave or wind sound.We finally discuss the limitations but also the possibilities offered by the synthesizer to achieve sound transformations based on the analysis of recorded sounds.

1 Introduction

The creation of auditory environments is still a challenge for Virtual Reality applications such as video games, animation movies or audiovisual infrastructures. The sensation of immersion can be significantly improved by taking into account an auditory counterpart to the visual information. In this context, synthesis models can constitute efficient tools to generate realistic environmental sounds. A comprehensive review of environmental sound synthesis can be found in [1, 2] based on the classification of everyday sounds proposed by W. Gaver [3, 4]. The author defined three main categories from the physics of the sound-producing events: vibrating solids (impacts, deformations...), aerodynamic sounds (wind, fire...) and liquid sounds (sea wave, drop...). Physically-based synthesis models were proposed for these categories. For instance, one can cite [5–7] for contact sounds, [8, 9] for aerodynamic sounds and [10] for liquid sounds.

In this paper, we focus on the synthesis of noisy environmental sounds that represent a huge variety of everyday sounds such as sea wave, wind, fire or air swishing ("whoosh") sounds. Note that they cover most of the categories of everyday sounds defined by W. Gaver. Several authors proposed relevant models describing the physical phenomena to simulate the resulting sound. For example, [8, 9] studied the vortex sounds produced by aerodynamic phenomena such as wind and combustion. First they pre-compute sound textures by making use of computational fluid dynamics. Then they use the sound textures for realtime rendering of aerodynamic sounds. The synthesis is driven by high-level parameters such as the fluid velocity. The model described in [9] handles the rubbling combustion noise and the authors can synthesize a complete fire by adding pre-recorded crackling sounds. They couple their sound modeling with graphics rendering to achieve a complete audiovisual simulation of aerodynamic phenomena.

Usually, especially for noisy sounds, the physics beyond environmental phenomena becomes rapidly complex (they refer to stochastic processes) and the physical approach may not answer to the real time constraints imposed by Virtual Reality applications. In particular, these applications require constraints of interactivity so that the surrounding sounds should be continuously updated according to the users' locations and actions in the virtual world. Thus, we considered signal-based synthesis models that aim at generating signals from their time-frequency representations, independently from their physical correlates. These models provided a wide palette of timbres and were extensively used for analysis, transformation and synthesis of musical sounds. Environmental sounds are also efficiently modeled with this approach. For instance, a wavelet approach was presented in [11] for analysis and synthesis of noisy environmental sounds. The authors proposed a method with four stages: analysis, parameterization, synthesis and validation. During the analysis, a wavelet decomposition of the sound was computed. The parameterization consisted in finding relevant manipulations of the wavelet coefficients so as to produce new sounds. The synthesis reconstructed the sounds from the manipulated wavelet coefficients. Finally the validation consisted in perceptual evaluations of the model quality. Several models were presented for sounds such as rain, car engine or footsteps... In addition, an "had-hoc" synthesis approach, that concentrates on the perception of environmental sounds by the listener rather than on analysis/synthesis has been presented in [12]. The author used time-varying filtered noise (subtractive synthesis) for creating and controlling sea waves and wind sounds. The goal was not to obtain audiorealistic sounds but to reproduce the main acoustic invariants, so that the sounds were easily recognizable and conveyed informations about the source. The synthetic sounds were used as auditory icons in an auditory display to monitor the progress of various operations. Finally, several analysis/synthesis techniques dedicated to stochastic signals were developed in the context of the additive signal model in which the sound is defined as a sum of deterministic and stochastic contributions. The deterministic part is composed of sinusoids whose instantaneous amplitude and frequency vary slowly in time while the stochastic part is modeled by a time-varying colored noise. We will describe the main synthesis techniques in the following Section.

The notion of source position and spatial extension is of great importance for the creation of immersive auditory environments. For example, the generation of sea or windy environments involves the synthesis of extended sea wave or wind sounds around

a fixed 3D position. It may also be of interest to control the spatial extension of an initial point-like sound source according to the distance source-listener: for example, the perceived width of a fire in a chimney increases when we move closer to the source and decreases when we move away (note that the timbre of the fire sound also varies). Thus, to take into account this aspect of sound rendering, we proposed the design of a spatialized synthesizer that efficiently combines synthesis and spatialization techniques at the same level of sound generation.

The paper is organized as follows: we first present the modeling of stochastic signals developed for additive signal models. Then, we describe the implementation of the spatialized synthesizer with a specific interest in noisy sound synthesis. Finally, we discuss the limitations and the possibilities offered by the synthesizer to generate realistic 3D environments.

2 Analysis/synthesis of stochastic signals

Sound synthesis can be implemented either in the time (using oscillator banks) or in the frequency domain [13]. We here focus on synthesis processes based on the short-time Fourier transform (STFT) since the proposed synthesizer (section 3) is based on this technique. The pioneer works were conducted by McAulay and Quatieri for speech synthesis [14]. The speech signal was approximated by a sum of short-time sinusoids which amplitudes, frequencies and phases were determined from the STFT. In [15], Serra and Smith brought improvements to the McAulay and Quatieri's model by developing the Spectral Modeling Synthesis (SMS). This method provided a complete analysis/transformation/synthesis scheme by taking into account both deterministic and stochastic parts of the signal. In this model, the stochastic residual $s(t)$ was defined by:

$$s(t) = \int_0^t h(t,\tau)x(\tau)d\tau \tag{1}$$

where $x(t)$ is a white input noise and $h(t,\tau)$ the impulse response of a "time-varying" filter. This stochastic part is usually assumed to represent a minor part of the original signal. However, for noisy environmental sounds, it would be predominant compared to the deterministic contribution.

In [16], Hanna and Desainte-Catherine presented an analysis/synthesis scheme to model noisy signals as a sum of sinusoids (CNSS). The frequency and phase of the sinusoids were randomly chosen at each frame (following a uniform distribution). A spectral density (number of sinusoids per frequency band) was additionally estimated in the analysis stage.

In [17, 18], the authors proposed the Bandwidth-Enhanced Additive Model that allowed synthesizing sinusoidal signals with noisy components by representing the noisy components as part of each partial. Their model follows ridges in a time-frequency analysis to extract partials having both sinusoidal and noise characteristics. They also proposed to use time-frequency reassignment methods to enhance the analysis accuracy [19].

2.1 Characterization of the time-varying spectral envelope

The stochastic signal is fully defined by its power spectral density (PSD), i.e., the expected signal power with respect to the frequency. This means that the instantaneous phases can be ignored for the resynthesis. In the SMS model, the analysis of the stochastic part consisted in measuring at each time step the average energy in a set of contiguous frequency bands covering the whole frequency range. The obtained amplitude curve corresponded to a piecewise linear function and is commonly called the "time-varying spectral envelope". To take into account human hearing system, Goodwin [20] proposed to define this spectral envelope on the ERB (Equivalent Rectangular Bandwidth) scale defined by:

$$ERB(f) = 21.4 \log_{10} \left(4.37 \frac{f}{1000} + 1\right) \tag{2}$$

where f is expressed in Hz. The amplitude was assumed to be constant in each ERB subband. The spectral modeling based on ERB scale allowed efficiently coding the residual part (excluding transients) without loss of sound quality.

This approach is close to the analysis/synthesis scheme implemented in the channel vocoder developed by Dudley in 1939 [21–23]. Formerly used for speech coding, the channel vocoder reconstructs an original signal based only on its short time power spectrum (by contrast with the so-called "phase vocoder" that keeps the phase information). The whole process is illustrated on Figure 1. At the analysis stage, a short time power spectrum of the input signal $s[n]$ is measured with a bank of M contiguous bandpass filters ($H_1[z]$,...,$H_M[z]$). Then, a time-varying envelope $e_m[n]$ is estimated on each subband signal $s_m[n]$ by using an envelope follower, and can be defined by:

$$e_m[n] = \sqrt{\frac{1}{I} \sum_{i=0}^{I-1} (v[i] s_m[n+i])^2} \tag{3}$$

where $v[n]$ is an analysis window of size I (a rectangular window for instance). At the synthesis stage, a pulse train (that simulates the glottal excitation) or a noise (that simulates the transient parts of the speech) noted $b[n]$ feeds the same filterbank ($H_1[z]$,..., $H_M[z]$) and is weighted by the subband spectral envelope $E[n] = (e_1[n], \ldots, e_M[n])$ so that the output $\hat{s}[n]$ is close to the original signal $s[n]$.

Note that for efficiency, the short time Fourier transform is typically used to implement the analysis/synthesis filterbank [23]. Furthermore, only a discrete version of the envelopes is usually computed:

$$E^r = (e_1^r, \ldots, e_M^r) = (e_1[rR], \ldots, e_M[rR]) \tag{4}$$

where r is the index of the frame and R the analysis hop size. This representation allows saving a significant amount of data compared to the original signal. When using an analysis hop size of 512 samples (e.g., a 1024-tap analysis window with an overlap factor of 50%) and 32 subbands, the compression ratio is $512/32 = 16$. Note that this ratio can be increased for relatively stationary sounds since longer analysis windows can be used without degrading the quality.

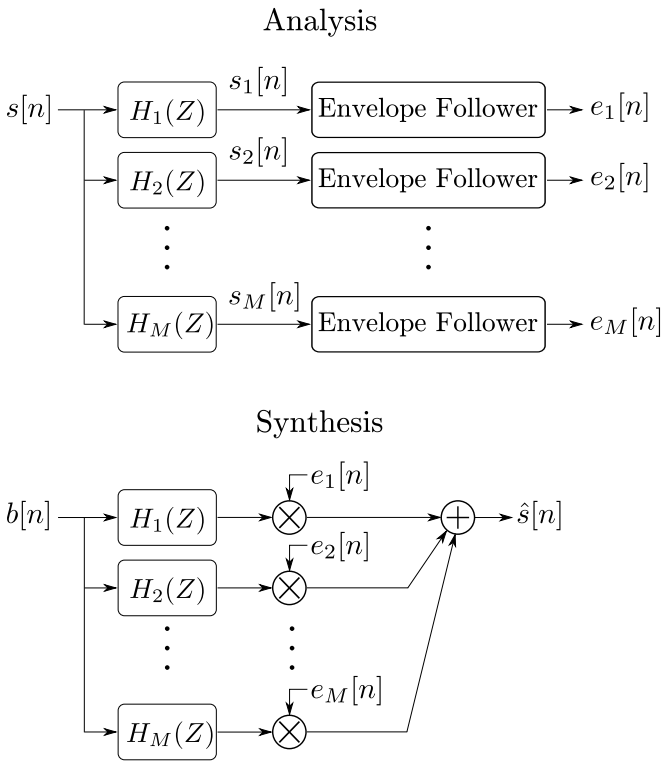

Fig. 1. Channel vocoder. *Analysis stage:* the original sound $s[n]$ is passed through a bank of M bandpass filters $H_m[z]$ and an envelope $e_m[n]$ is estimated in each subband, resulting in time-varying spectral envelope $E[n] = (e_1[n], \ldots, e_M[n])$. *Synthesis stage:* the input signal $b[n]$ is passed through the same filterbank $H_m[z]$ and weighted by the estimated set of spectral envelopes $E[n]$.

2.2 Time-frequency domain synthesis

For a given time-varying spectral envelope, the stochastic signal can be synthesized in the time-frequency domain using the inverse fast Fourier transform (IFFT) algorithm developed by Rodet and Depalle [24] and commonly called "IFFT synthesis". From a theoretical point of view, an approximation to the STFT is computed from the synthesis parameters (i.e., the expected spectral envelope for noise), then the inverse STFT is processed. IFFT synthesis is an implementation with a frame by frame pattern. Short-time spectra are created and the IFFT is performed. The resulting short-time signals are weighted by the synthesis window and overlap-added (OLA) to reconstruct the temporal signal[(1)].

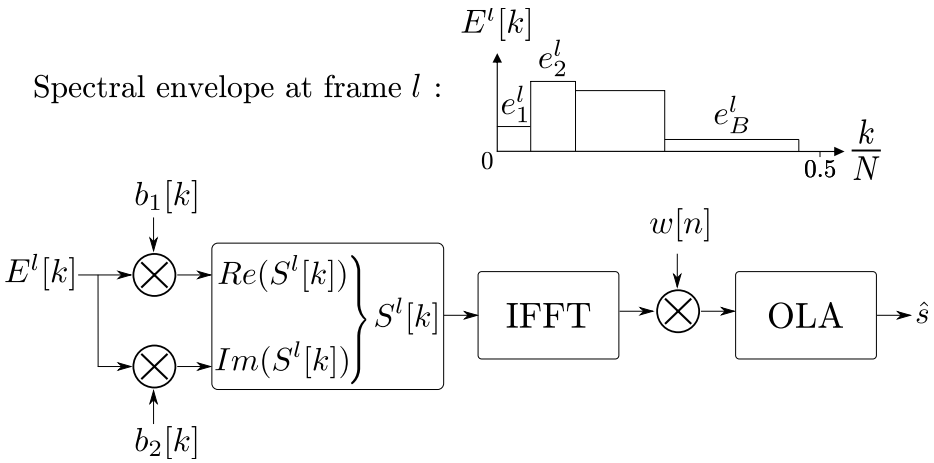

Fig. 2. IFFT synthesis of a noisy signal $\hat{s}[n]$ from a given time-varying spectral envelope $E^l[k]$. At each frame l, the real and imaginary part of the short-time spectrum $S^l[k]$ is reconstructed by multiplying $E^l[k]$ by two random sequences $b_1[k]$ and $b_2[k]$. The resulting frames are inverse fast Fourier transformed (IFFT), weighted by the synthesis window $w[n]$ and finally overlap-added (OLA) to obtain the signal $\hat{s}[n]$.

These successive steps illustrated in Figure 2 are now detailed in the case of stochastic signal synthesis. Let N be the block size of the synthesis window $w[n]$ and $S^l[k]$ the N-point short-time spectrum (STS) to be reconstructed from a given spectral envelope E^l at frame l. This envelope is first resampled for $k = 0, \ldots, N/2$ since only positive frequencies (i.e., $k = 0, \ldots, N/2$) need to be considered for a real-valued signal (the corresponding spectrum is conjugate-symmetric). Then, the obtained envelope $E^l[k]$ is multiplied by two Gaussian random sequences $b_1[k]$ and $b_2[k]$ to get the real and

[1] If the initial spectral envelope is estimated from the analysis of a natural sound, the analysis window, the synthesis window and hop size can be different. The spectral envelope is interpolated at the synthesis frame rate.

imaginary parts of the STS $S^l[k]$:

$$\begin{cases} \Re\{S^l[k]\} = b_1[k]E^l[k] \\ \Im\{S^l[k]\} = b_2[k]E^l[k] \end{cases} \tag{5}$$

Additionally, $b_1[k]$ and $b_2[k]$ should satisfy [15, 20]:

$$b_1[k]^2 + b_2[k]^2 = 1 \quad \text{for} \quad k = 0, \ldots, N/2 \tag{6}$$

so that the magnitude of $S^l[k]$ fits the desired spectral envelope $E^l[k]$. However, if we consider the discrete Fourier transform $G[k]$ of a zero-mean N-point sequence of Gaussian white noise with variance σ^2 , it is shown in [25] that the magnitude of $G[k]$ follows a Rayleigh distribution and the phase a uniform distribution. The real and imaginary parts of $G[k]$ are independent Gaussian sequences with variance $\sigma^2 N/2$. Informal listening tests confirmed that letting $b_1[k]$ and $b_2[k]$ be two independent Gaussian sequences, i.e., not satisfying Equation (6), leads to good perceptive results.

Then, the short-time spectrum $S^l[k]$ is inverse fast Fourier transformed to obtain the short-time signal $s^l[n]$:

$$s^l[n] = \frac{1}{N} \sum_{k=0}^{N-1} S^l[k] e^{j2\pi n \frac{k}{N}} \tag{7}$$

and after weighted by the synthesis window $w[n]$. Finally, the overlap-add is processed on the weighted short-time signals to get the whole reconstructed signal $s[n]$:

$$s[n] = \sum_{l=-\infty}^{\infty} w[n - lL] s^l[n - lL] \tag{8}$$

where L is the synthesis hop size.

For the purpose of combining the stochastic synthesis with sinusoidal synthesis, the multiplication by the synthesis window is performed as a convolution in the frequency domain [26]. In that case, to reduce the computational cost of the convolution, the synthesis window is usually defined as a 4-term Blackman-Harris window (side-lobe level at -92 dB) whose spectrum is truncated to its main-lobe and sampled on 9 frequency bins. For our concern, we use a digital prolate spheroidal window (DPSW) since this window obtains the optimal energy concentration in its main-lobe [27]. The DPSW family constitutes a particular case of the "discrete prolate spheroidal sequences" developed by Slepian [28]. We compute the DPSW with bandwidth $\frac{3.5}{N}$ (N being the window size), truncate its spectrum to its main-lobe (the side-lobe is at -82 dB) and sample it on 7 frequency bins. Comparison between the Blackman-Harris window and this DPSW is illustrated on Figure 3. For an accurate synthesis, [29] showed that the following condition:

$$\sum_{l=-\infty}^{\infty} w[n - lL]^2 = 1 \qquad \forall n \in Z \tag{9}$$

should be satisfied to avoid modulations in the resulting signals and to make sure to obtain a flat power spectrum in the case of white noise. The latter condition (Equation (9))

was satisfied by the DPSW window provided a small synthesis hop size, typically corresponding to an overlap of 75% between two successive frames. To increase the efficiency of the method, Rodet and Depalle proposed to reduce the overlap to 50% and to smooth discontinuities between frames with an additional window, that is different from the synthesis window [24]. In practice, Bartlett, Hann or Bartlett-Hann windows are well adapted to this use.

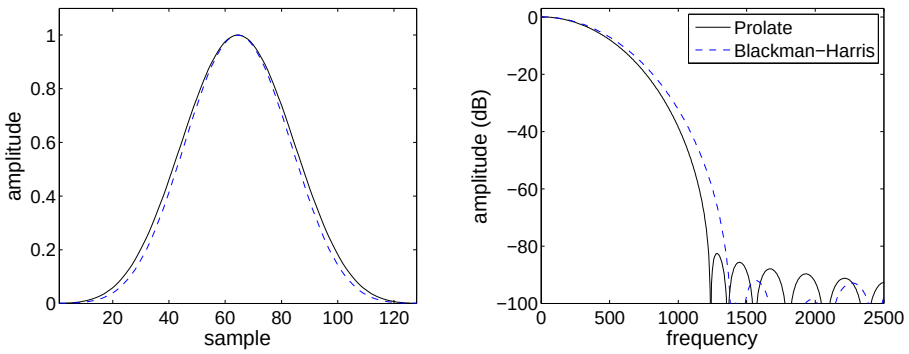

Fig. 3. Comparison of the 4-term Blackman-Harris window and the DPSW with bandwidth $\frac{3.5}{N}$ (N being the window size). Here $N = 128$ and the sampling frequency is 44.1 kHz.

3 The spatialized synthesizer

Based on this IFFT synthesis, we proposed the design of a spatialized synthesizer to simulate sound sources in a 3D space. By contrast with traditional implementations that consist in synthesizing a monophonic source before spatialization, the synthesizer has the advantage to efficiently combine synthesis and spatialization modules at the same level of sound generation in a unified architecture. In the following sections, we summarize the implementation of the synthesizer by briefly describing the source positioning module and how it was included in a one-stage architecture. Then, we present a method for the spatial extension rendering that was included in the synthesizer. This latter effect will be of great importance to simulate naturally diffused noisy sources such as sea wave or wind sounds. We refer the reader to [30, 31] for more details on the implementation of the synthesizer.

3.1 Source positioning

Several approaches exist for positioning sound sources in virtual environments. For instance, High Order Ambisonics (HOA) and Wave Field Synthesis aim at reconstructing the sound field in a relatively extended area with a multichannel loudspeaker setup. Discrete panning (time or amplitude panning) reconstructs main aspects of the sound field

at the “sweet spot”. In the case of binaural synthesis, the sound field at the entrance to the ear canals is reconstructed by filtering the monophonic sound with Head Related Impulse Responses. The binaural synthesis is mainly for headphone reproduction but can also be extended to loudspeaker setup commonly referred to as “Transaural” setup. A general implementation strategy applicable to all the techniques cited above can be found in [32].

For the synthesizer, we proposed an architecture relying on “amplitude-based” positioning, for which the spatial filterbank is reduced to a vector of position-dependent gains. Consequently the synthesizer was compatible with several 3D positioning methods such as Ambisonics, HOA, amplitude panning and some multichannel implementations of binaural synthesis.

3.2 One-stage architecture

The one-stage implementation was made possible while the synthesis and the spatial encoding can be performed in the frequency domain. Thus, it handled all positioning methods that use only gains in the spatial encoding module. WFS was excluded since it required delays in the spatial encoding module, expensive to compute in the frequency domain. The integration was effectuated following three main stages:

1. *time-frequency domain synthesis:* based on the IFFT synthesis described in Section 2.2, the real and imaginary parts of the STS are computed at each frame from a given spectral envelope associated to each source.
2. *3D positioning:* the spatial encoding is processed by directly applying spatial gains to the STS of each source. The encoded STS are summed channel by channel in the mixing stage. The spatial decoding is performed by matrixing and/or filtering the multichannel signal. Note that the decoding stage is common for all sources and do not depend on individual source position.
3. *reconstruction of the time-domain signal:* the decoded STS are inverse fast Fourier transformed and the successive short-time signals are overlap-added to get the synthetic signal for each channel of the loudspeaker setup.

The proposed architecture reduced the computational cost since it requires only one IFFT per frame for each loudspeaker channel, independently of the number of sound sources. This is particularly attractive for applications over headphones using binaural synthesis because only two IFFT are computed per frame, while the scene can contain hundreds of sources.

3.3 Source width extension

The extension effect cannot be reproduced by simply feeding a multichannel loudspeaker setup with duplicated copies of a monophonic signal: in that case, a relatively sharp phantom image is created. The creation of a wide spatial image necessitates feeding the loudspeaker setup with decorrelated versions of the monophonic signal (see Figure 4). Each decorrelated copy is called “secondary source”. In this context, various

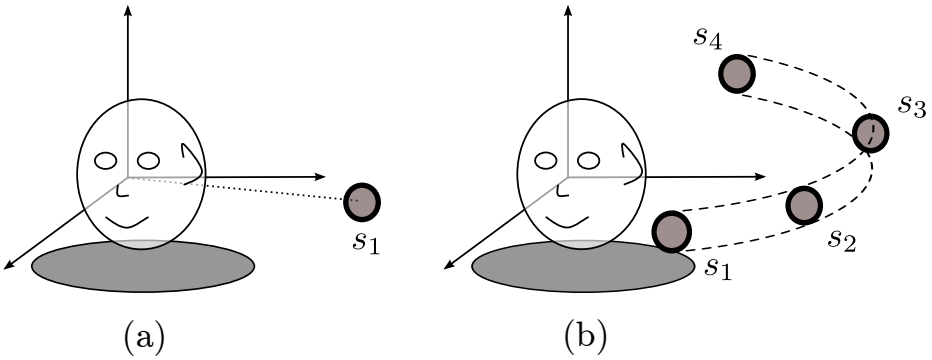

Fig. 4. (a) A single point-like source produces a narrow auditory event. (b) Several decorrelated copies of the source located at evenly spaced positions around the listener produce a wide auditory event. The perceived source width can be adjusted by changing the relative contributions (i.e., gains) of the decorrelated sources.

filtering techniques were proposed to create decorrelated versions of an original signal [33–36] that were further used to create wide source images [36–39]. Nevertheless, filtering techniques may alter the transients and the timbre of the original sound.

The proposed architecture presents the advantage to overcome this drawback by effectuating the decorrelation at the synthesis stage, i.e., without filtering process. In particular, different versions of the STS $S^l[k]$ (noted $S_i^l[k]$) can be created from a same original spectral envelope $E^l[k]$ with different noise sequences $b_i[k]$ (see Figure 2). The resulting signals $\hat{s}_i[n]$ correspond to different versions of the same original sound, and they are statistically uncorrelated. This way an unlimited number of decorrelated secondary sources can be created from different random sequences. Based on this technique, we included a spatial extension effect in the synthesizer by using a maximum of eight virtual secondary sources evenly spaced on a circle surrounding the listener. The control of source width acted on the relative contributions of the eight sources via a set of extension gains.

Furthermore, it was of interest to control the interchannel correlation. For that purpose, a correlation C was introduced between the signals $s_1[n]$ and $s_2[n]$ that can be accurately controlled by creating $S_2^l[k]$ with:

$$\begin{cases} \Re\{S_2^l[k]\} = C \times \Re\{S_1^l[k]\} + \sqrt{(1-C^2)} \times b_1[k]E^l[k] \\ \Im\{S_2^l[k]\} = C \times \Im\{S_1^l[k]\} + \sqrt{(1-C^2)} \times b_2[k]E^l[k] \end{cases} \tag{10}$$

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

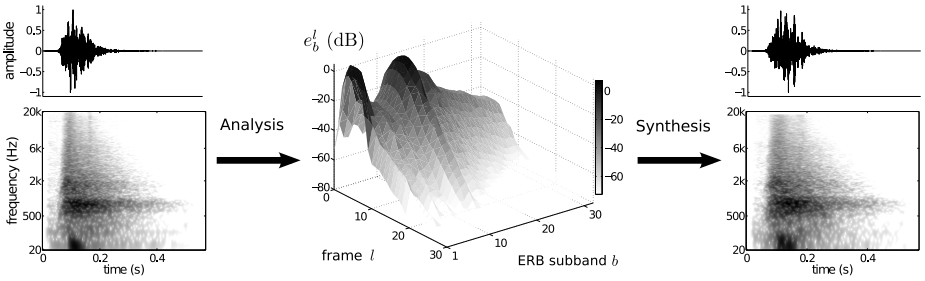

Fig. 5. Analysis/Synthesis of an air swishing ("whoosh") sound, analysis with 32 subbands using and a 1024-tap window with 75% overlap. The amount of data is reduced by a factor 8 compared to the original sound. The synthesis uses a 1024-tap window with 75% overlap.

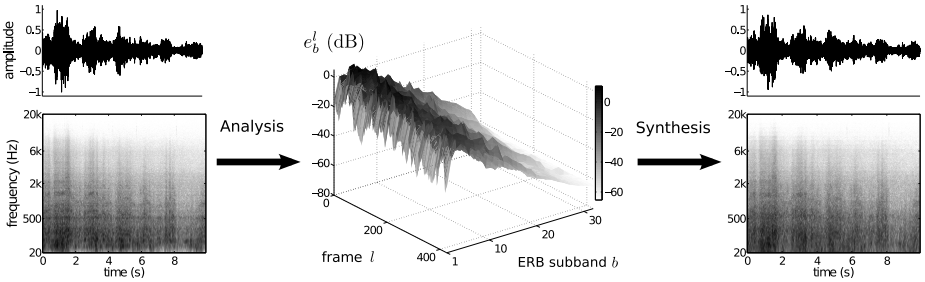

Fig. 6. Analysis/Synthesis of a wind sound, analysis with 32 subbands using and a 4096-tap window with 75% overlap. The amount of data is reduced by a factor 32 compared to the original sound. The synthesis uses a 1024-tap window with 75% overlap.

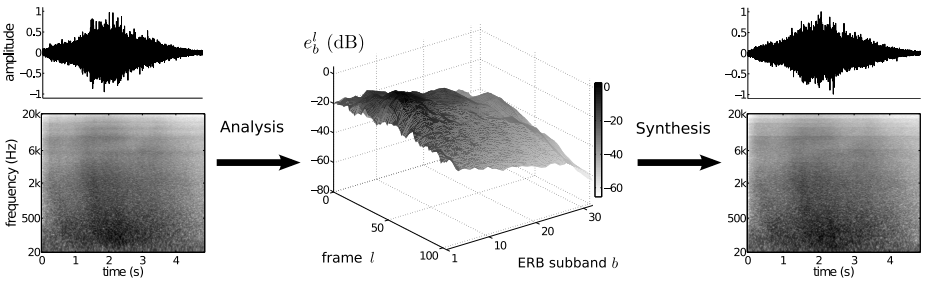

Fig. 7. Analysis/Synthesis of a wave sound, analysis with 32 subbands using and a 8192-tap window with 75% overlap. The amount of data is reduced by a factor 64 compared to the original sound. The synthesis uses a 1024-tap window with 75% overlap. **(Left)** Original signal and its time-frequency representation. **(Middle)** Time-varying spectral envelope defined in 32 ERB subbands estimated from the analysis of the original sound. **(Right)** Reconstructed signal and its time-frequency representation. The reconstructed sound is perceptually similar to the original one. Original and reconstructed sounds are available at [40].

4 Applications to wide noisy environmental sounds

4.1 Choice of the synthesis parameters

The complete version of the spatialized synthesizer allowed generating both deterministic and stochastic sounds, thus covering the main categories of environmental sound categories defined by Gaver (see Introduction). Thus, the synthesizer can be used to simulate various 3D auditory environments. As mentioned previously, we here focus on the generation of specific class of noisy environmental sounds. For the synthesis, we propose to find a set of parameters satisfying time and frequency resolution constraints to generate most of these types of sounds.

We examined several prototypes of noisy sounds among the main environmental sound categories, i.e., wind, sea waves and air swishing sounds. Their time-frequency representations were investigated. We observed that all sounds usually do not have very sharp transients and that 32 subbands evenly spaced on the ERB scale were sufficient to reproduce their salient spectral properties. Based on these observations, we experimented different synthesis window sizes and concluded that a 1024-tap digital prolate window led to a good compromise for an accurate reproduction of the sound source. Regarding the time resolution, this window was short enough to reproduce signals with relatively fast variations (i.e., 21 milliseconds at 48kHz sampling frequency). Regarding the frequency resolution, this window was sufficiently long for synthesizing the required 32 ERB subbands.

Figures 5, 6 and 7 illustrate the analysis/synthesis process of air swishing, wind and wave sounds by using the 1024-tap synthesis window. Both temporal and spectral properties of the original sounds were accurately reproduced with the chosen synthesis window. Sound examples are available at [40].

Finally, the spatialized synthesizer allowed to easily extend sound sources around the listener based on the decorrelation technique described in Section 3.3. The perceptual evaluation of the extension effect was effectuated by conducting a formal listening test [31]. The test was performed on several items representative of the main categories of environmental sounds, in particular, sea wave and wind sounds using two reproduction systems (headphones and a standard 5.0 loudspeaker setup). Results showed that the extension control implemented in the synthesizer accurately modified the perceived width of these noisy sound sources. Sounds are available at [41].

4.2 Limitation of the IFFT synthesis

A limitation of the IFFT synthesis remains the inherent trade-off between time and frequency resolutions set by the length of the synthesis window. In particular, narrowband noise synthesis requires long windows (for example, more than 1024 taps; longer the window, narrower the bandwidth) with which short transient signals cannot be synthesized. Freed proposed a variant version of the IFFT synthesis algorithm [24] to reproduce noisy signals with narrow frequency bands [42, 43]. His algorithm synthesized short-time sinusoids whose phase was randomly distributed between 0 and 2π at each frame. The resulting signal was perceived as a narrow-band noise. This approach is equivalent to the stochastic synthesis method described in section 2.2 and considering

a spectral envelope that is non-zero only for one frequency bin (the non-zero bin corresponding to the desired center frequency). The resulting noisy signal has the narrowest bandwidth that can be generated by IFFT synthesis. The multiplication by the synthesis window being equivalent to a convolution in the frequency domain, this narrowest bandwidth is equal to the bandwidth of the synthesis window. Choosing a long synthesis window guarantees a narrow bandwidth, but a bad resolution in the time domain. On the contrary, choosing a short synthesis window guarantees a good time resolution, but a wide bandwidth in the frequency domain. Consequently, IFFT synthesis may not accurately generate noisy sounds presenting both short transients and narrowband components.

The fire sound is a good example to illustrate the limitation of the IFFT synthesis since it is constituted of both crackling (noisy transients impacts), hissing (narrowband noises) and combustion. A way to overcome this trade-off consisted in using different windows according to the different contributions of the sound. In particular, we chose 128 taps for cracklings, 1024 for hissing and for combustion noises. However, it is desirable to have a single synthesis window, otherwise the all architecture described in Section 3.2 (including the spatialization modules) must be duplicated for each synthesis window.

We recently proposed an alternative to the IFFT synthesis for synthesizing narrowband noises with short transients using a single synthesis window. The proposed method is based on the so-called subband technique to generate time-frequency noise with an auto-correlation function so that the resulting output PSD fits any arbitrary spectral envelope at the expense of extra computations [44]. In particular, the bandwidth of the generated noise can be narrower than the one of the synthesis window. We are currently investigating the possibilities offered by this method.

4.3 Sound transformations

Compared to classical wavetable synthesis, the spatialized synthesizer allows parametric transformations of the generated signals. In the context of analysis/synthesis, the analysis leads to the determination of the synthesis parameters for reconstructing the original sound. Modifying synthesis parameters allows the creation of new sounds (see Figure 8).

In [45] the authors used the analysis/transformation/synthesis framework for generating complex scenes from a small set of environmental recordings. In particular, they decompose the original sounds into deterministic and stochastic parts. They apply different parametric transformations to these components to generate new sound sequences avoiding unnatural repetitions.

In the synthesizer, classical signal transformations such as pitch-shifting or time-stretching can be easily applied on the spectral envelopes. Stretching or warping the envelopes in time by simple interpolation techniques results in a temporal stretching/warping of the reconstruded sound without pitch alterations. Also interpolating and shifting the spectral envelopes in the frequency domain produces transpositions without temporal alterations. These transformation processes are illustrated on Figure 9. Sound morphing can also be realized to go from one set of spectral envelopes to another. The simplest morphing is a simple linear interpolation between the two sets of envelope.

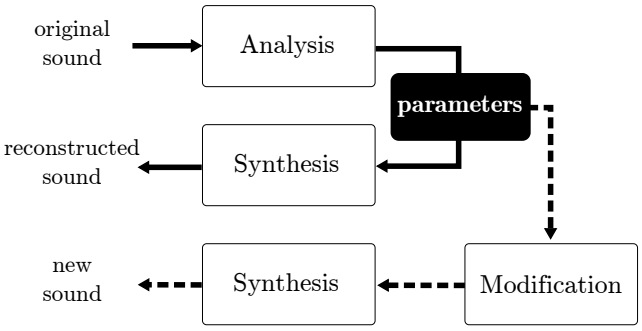

Fig. 8. Analysis/transformation/synthesis framework: the synthesis parameters extracted from the analysis of an original sound are used to resynthesize the original one or to create new sounds by signal transformations (e.g., by pitch-shifting, time-stretching, equalization or morphing).

More elaborate solutions can be found in [46] for morphing between spectral envelopes of speech signals. Subband equalization of the reconstructed signal is also possible by weighting the spectral envelope with a set of coefficients before the synthesis process. It provides an efficient and intuitive control of the spectral shape of the reconstructed sound.

In addition, since synthesis and spatialization are processed at the same level of sound generation, sound transformations can be achieved on timbre parameters with respect to the source position, as shown in Figure 9.

5 Conclusion

In this paper, we focused on noisy environmental sounds. We proposed an overview of the existing stochastic models generally developed in the context of additive signal model. We described an efficient frequency-domain synthesis technique based on IFFT algorithm. The use of stochastic modeling in the frequency domain allowed us to propose a one-stage architecture where spatialization techniques operate directly at the synthesis stage. Listening tests have shown that our technique produced realistic extended environmental sound sources such as sea waves and wind.

References

1. P. R. Cook. *Real Sound Synthesis for Interactive Applications*. A. K Peters Ltd., 2002.
2. D. Rocchesso and F. Fontana. *The Sounding Object*. http://www.soundobject.org/, 2003.
3. W. W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.
4. W. W. Gaver. How do we hear in the world? explorations in ecological acoustics. *Ecological Psychology*, 5(4):285–313, 1993.
5. K. van den Doel, P. G. Kry, and D. K. Pai. Foleyautomatic: physically-based sound effects for interactive simulation and animation. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 537–544, 2001.

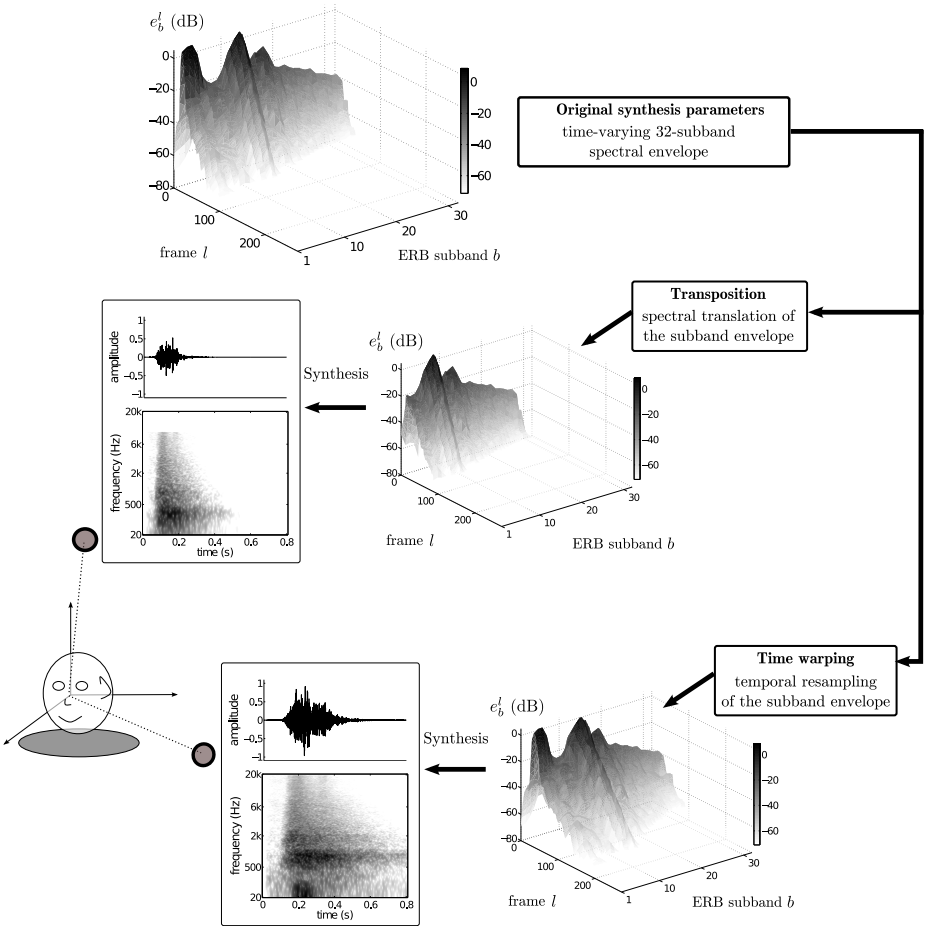

Fig. 9. Examples of transformations by modifying the original synthesis parameters of a whoosh sound produced by a moving stick (Fig. 5). Transposition to lower frequencies leads to the perception of a ticker stick. Time warping (resampling by a factor 1.5 here) leads to the perception of a slower motion. Sound examples are available at [40]. The transformations can be achieved with respect to the source position.

6. 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*, pages 175–181, 2002.
7. N. Raghuvanshi and M. C. Lin. Interactive sound synthesis for large scale environments. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 101–108, 2006.
8. Y. Dobashi, T. Yamamoto, and T. Nishita. Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics. *ACM Transactions on Graphics (Proc. SIGGRAPH 2003)*, 22(3):732–740, 2003.
9. Y. Dobashi, T. Yamamoto, and T. Nishita. Synthesizing sound from turbulent field using sound textures for interactive fluid simulation. *EUROGRAPHICS*, 23(3):539–546, 2004.
10. 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.
11. N. E. Miner and T. P. Caudell. Using wavelets to synthesize stochastic-based sounds for immersive virtual environments. In *Proceedings of of ICAD 97-The fourth International Conference on Auditory Display*, 1997.
12. S. Conversy. Ad-hoc Synthesis of auditory icons. In *Proceedings of of ICAD 98-The fifth International Conference on Auditory Display*, 1998.
13. M. Goodwin. *Adaptive Signal Models: Theory, Algorithms and Audio Applications*. PhD thesis, University of California, Berkeley, 1997.
14. R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis system based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4), 1986.
15. X. Serra and J. O. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990.
16. P. Hanna and M. Desainte-Catherine. A statistical and spectral model for representing noisy sounds with short-time sinusoids. *EURASIP Journal on Applied Signal Processing*, 5(12):1794–1806, 2005.
17. K. Fitz and L. Haken. Bandwidth enhanced sinusoidal modeling in lemur. In *Proceedings of the International Computer Music Conference*, 1995.
18. K. Fitz, L. Haken, and P. Christensen. Transient preservation under transformation in an additive sound model. In *Proceedings of the International Computer Music Conference*, 2000.
19. K. Fitz and L. Haken. On the use of time-frequency reassignment in additive sound modeling. *JAES*, 50(11):879–893, 2002.
20. M. Goodwin. Residual modeling in music analysis-synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996.
21. H. Dudley. The vocoder. *Bell Labs Record*, 17:122–126, 1939.
22. B. Gold and C. M. Rader. The channel vocoder. *IEEE Transactions on Audio and Electroacoustics*, 15(4):148–161, 1967.
23. J. O. Smith. *Spectral Audio Signal Processing, October 2008 Draft*. http://ccrma.stanford.edu/∼jos/sasp/, 2008. online book.
24. X. Rodet and P. Depalle. Spectral envelopes and inverse fft synthesis. In *Proceedings of the 93rd AES Convention*, 1992.
25. William Hartmann. *Signal, Sound and Sensation*. American Institute of Physics, 2004.
26. X. Amatriain, J. Bonada, A. Loscos, and X. Serra. *DAFX: Digital Audio Effects*, chapter Spectral Processing. John Wiley & Sons Publishers, 2002.
27. T. Verma, S. Bilbao, and T. H.Y. Meng. The digital prolate spheroidal window. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996.

28. D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty. V- The discrete case. *Bell System Technical Journal*, 57:1371–1430, 1978.
29. P. Hanna and M. Desainte-Catherine. Adapting the overlap-add method to the synthesis of noise. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFX'02)*, 2002.
30. C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone. Spatialized additive synthesis of environmental sounds. In *Proceedings of the 125th AES Convention*, 2008.
31. C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone. A 3d immersive synthesizer for environmental sounds. *accepted to IEEE Transactions on Audio, Speech, and Language Processing*.
32. 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.
33. M. R. Schroeder. An artificial stereophonic effect obtained from a single audio signal. *JAES*, 6(2), 1958.
34. R. Orban. A rational technique for synthesizing pseudo-stereo from monophonic sources. *JAES*, 18(2), 1970.
35. M. A. Gerzon. Signal processing for simulating realistic stereo images. In *AES Convention 93*, 1992.
36. G. Kendall. The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19(4):71–87, 1995.
37. A. Sibbald. Method of synthesizing an audio signal. United State Patent No. US 6498857 B1, december 2002.
38. G. Potard and I. Burnett. Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays. In *Proc. Int. Conf. on Digital Audio Effects (DAFx'04)*, 2004.
39. J.-M. Jot, M. Walsh, and A. Philp. Binaural simulation of complex acoustic scene for interactive audio. In *Proceedings of the 121th AES Convention*, 2006.
40. *www.lma.cnrs-mrs.fr/∼kronland/spatsynthIcad09/index.html*.
41. *www.lma.cnrs-mrs.fr/∼kronland/spatsynthIEEE/index.html*.
42. A. Freed. Real-time inverse transform additive synthesis for additive and pitch synchronous noise and sound spatialization. In *Proceedings of the 104th AES Convention*, 1998.
43. Adrian Freed. Spectral line broadening with transform domain additive synthesis. In *Proceedings of the International Computer Music Conference*, 1999.
44. D. Marelli, M. Aramaki, R. Kronland-Martinet, and C. Verron. Time-frequency synthesis of noisy sounds with narrow spectral components. *accepted to IEEE Transactions on Audio, Speech, and Language Processing*.
45. A. Misra, P. R. Cook, and G. Wang. A new paradigm for sound design. In *Proc. Int. Conf. on Digital Audio Effects (DAFx06)*, 2006.
46. X. Rodet and D. Schwarz. *Analysis, Synthesis, and Perception of Musical Sounds: Sound of Music*, chapter Spectral Envelopes and Additive + Residual Analysis/Synthesis, pages 175–227. Springer, 2007.