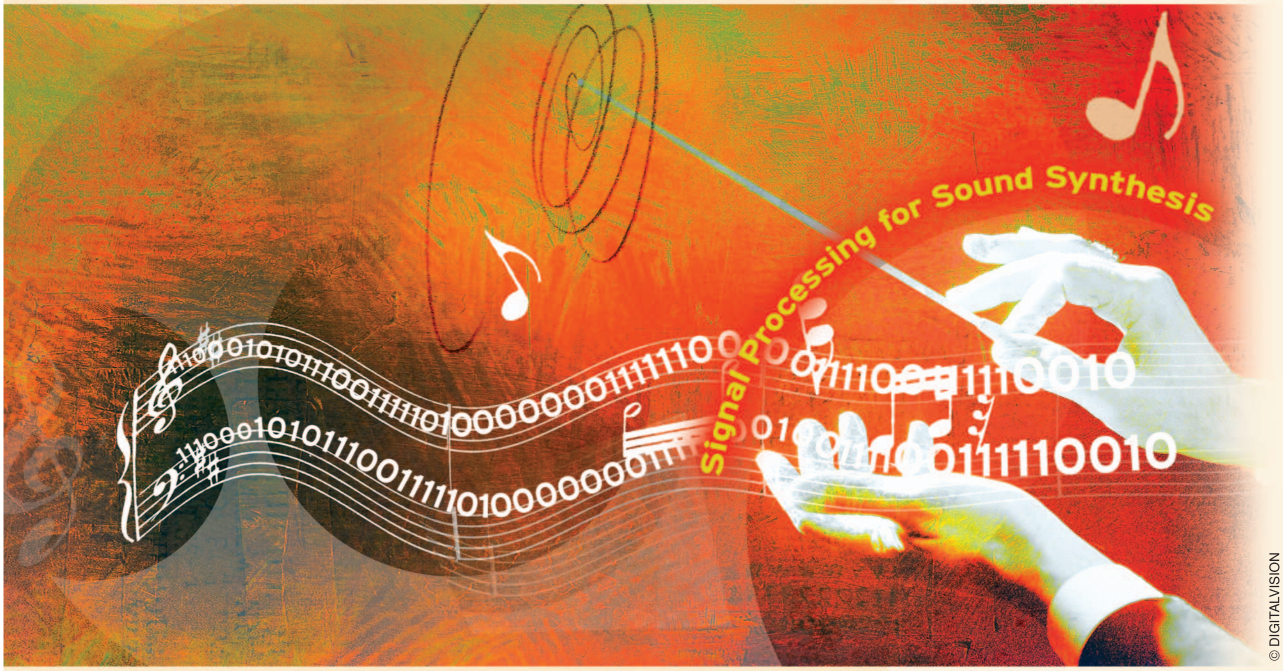
[Pietro Polotti and Gianpaolo Evangelista]



© DIGITALVISION

Fractal Additive Synthesis

[A deterministic/stochastic model for sound synthesis by analysis]

Gabor expansions or wavelets? A vivid debate characterized the last decade of the past century when wavelets were born and were being raised by an ever-growing scientific community. Wavelets were innovative and fun to play with [1]. In addition, a negative result known as the Balian-Low theorem [2], [3] showed that it is impossible to obtain an optimally compact Gabor representation [4] from time-frequency atoms whose uncertainty product is finite. However, while wavelets enchanted, among others, the image processing community, the audio processing community remained highly skeptical. Yet, the unprecedented nonuniform time-frequency resolution enjoyed by the wavelet representation, reminiscent of the functioning of the cochlea and at the very basis of our auditory perception, should have appealed to a wider group of researchers. Probably one reason is to be ascribed to the fact that the simplest wavelets are obtained by limiting the frequency resolution to one octave, too poor for audio applications. Moreover, when we think of sound, we often think of pitch, harmonics, or partials, and we are carried into the realm of Fourier analysis. A vast chunk of audio applications ranging from sinusoidal models to perceptual coders is in fact based on variations on the phase vocoder theme, another synonym of Gabor expansions and short-time Fourier transforms (STFTs). Results on Wilson or local cosine bases [5] also show that time-frequency representations with a sinusoidal character can overcome the Balian-low limitations.

But wavelets also possess nice properties that make them suitable to represent transients and noise at small or large scale. Wait a minute! Aren't transients and noise the spice blending music together, the salt and pepper of any sound? Could sounds sound natural and sharp when deprived of these components? But these are precisely the objects that are harder to model with sinusoids.

The wavelet song is hard to be sung unless a tuning fork can be drawn from the hat. Here we try to show that sinusoidal and wavelet methods can coexist and reinforce each other. Our tuning fork is drawn from a pitch-synchronous representation of pseudoperiodic signals. Pseudo because nothing is perfect, not even mother nature's periodicity. Our sound synthesis by analysis method is based on the harmonic band wavelet transform (HBWT) obtained by hybridation of local cosine and wavelet bases. From the friction of the bow on the string to the breath flowing in the mouthpiece, noise stimulates music and noise will stimulate our discussion. The frequency spectra of pitched

1053-5888/07/$25.00©2007IEEE

sounds show harmonically related peaks, at least approximately. However, a portion of the energy is also found in between the peaks. It is reasonable to try to represent harmonics in terms of a periodic trend plus fluctuations at several scales. The power spectral distribution of voiced sounds seems to decay as an inverse power of the distance in frequency from each peak. It is reasonable to try to represent harmonics in terms of a periodic trend plus fluctuations at several scales. These considerations inspired us to model sounds with pseudoperiodic noise built out of an ensemble of modulated self-similar $1/f$-like stochastic processes. Since the signal is generated by a superposition of partials, our technique falls in the category of additive synthesis, a classical method in sound synthesis. However, our partials are not elementary signals as they have the shape of pseudoharmonic components obtained by frequency modulating fractal $1/f$ noise sources. This justifies the name fractal additive synthesis (FAS). We argue that noise and periodicity blend perfectly well in the HBWT representation. Well, almost!

The goal of this article is to present, in an informal discussion, the methods on which FAS rests on. We will illustrate the coding aspects of FAS as well as the potentiality of the method for what concerns sound transformations, such as time stretching and pitch shifting. Additionally, since the model of the harmonic components is independent from the model of the noise, expression and amplification of the excitation noise can be incorporated.

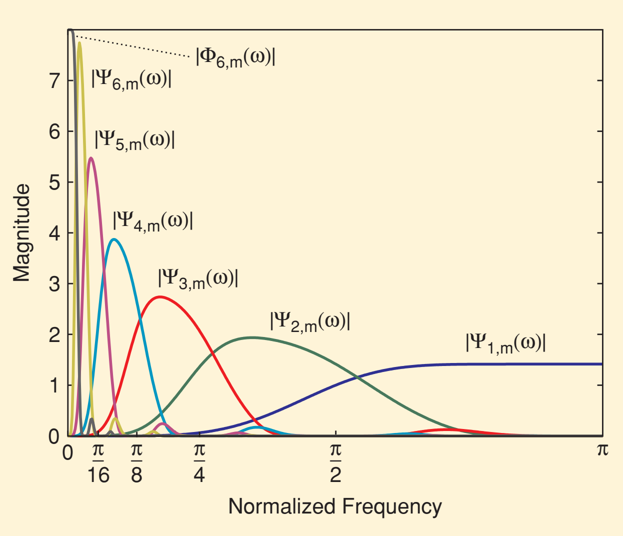
WAVELETS

Together with local cosine bases and $1/f$ noise, wavelets are ingredients of our synthesis method. Excellent books (for example [1], [6]–[8]) on the subject have been published. For further in-depth studies, please refer to the voluminous literature produced in the past two decades, which is not possible to thoroughly cite here in finite space and time. We will only attempt to recall some basic facts about wavelets. The integral wavelet transform allows us to represent sound signals as organized clouds of grains in a time-scale plane [1]. The basic elements of the representation are obtained by scaling and shifting a unique wavelet function $\psi(t)$ in the following fashion:

$$\psi_{\alpha,\tau}(t) = \frac{1}{\sqrt{\alpha}}\psi\left(\frac{t-\tau}{\alpha}\right), \qquad (1)$$

where $\alpha > 0$ is the scaling factor and τ is the time shift. Wavelets are chosen among bandpass functions that have the property to be localized in both time and frequency and have finite uncertainty product. By scaling, both the center frequency and the bandwidth of the wavelet are multiplied by the same factor so that the the Q factor, defined as the ratio of these quantities, remains constant, and the same is true for the uncertainty product. Since the center frequency of the wavelets moves according to the inverse of scale, the organized cloud of wavelets can also be considered as a time-frequency representation in terms of grains of constant uncertainty. At lower frequencies (large scale), we obtain sharper frequency resolution at the expense of lower time resolution. At higher frequencies (small scale), the time resolution is sharper and the frequency resolution is lower. This should be compared to the uniform time-frequency resolution of the STFT [4].

The redundancy of the integral transform is reduced or removed by sampling on a nonuniform grid in the time-scale plane (τ, α). The most common grid is dyadic: $\alpha_n = 2^n$; $\tau_{n,m} = 2^n m$; $n, m \in \mathbf{Z}$. The computation of the expansion coefficients can be performed by iterating the analysis section of a two-channel critically sampled filter bank, consisting of a quadrature-mirror filter (QMF) pair cascaded by factor 2 downsamplers [1]. As part of our synthesis scheme, the wavelet transform structure and its inverse are respectively shown in Figure 4(a) and (b). The QMF filters $H(z)$ and $G(z)$ are half-band lowpass and high-pass, respectively. The impulse responses of the analysis filters $\tilde{H}(z)$ and $\tilde{G}(z)$ are simply obtained by time-reversal of the impulse responses of the synthesis filters. The discrete-time bandpass wavelets $\psi_{n,m}(k)$, useful for the expansion of sampled signals, are defined as the impulse responses of the lower branches of the synthesis filter bank, to which we adjoin the impulse response of the upper branch referred to as the lowpass scaling sequence $\varphi_{N,m}(k)$. The structure implements a finite-scale wavelet expansion in which the signal is decomposed into a low-frequency trend given by the scaling component and into deviations from the trend over several scales given by the wavelet components. The frequency domain character of discrete-time wavelets and scaling sequence is displayed in Figure 1.



[FIG1] Magnitude Fourier transform of discrete-time wavelets and scaling function ($N = 6$).

WAVELETS WITH A PITCH

The applications of wavelet expansions in audio are limited by several factors. Dyadic wavelet sets have the same resolving power as an octave band equalizer, generally too poor for processing and feature extraction. Higher-frequency resolution wavelet sets have been proposed, which are based on rational sampling grids [9] or frequency warping [10], [11] to match the wavelet decomposition to critical bands in a perceptual scale. The price to pay is a very constrained design in the first case or a higher computational cost for their evaluation in the second case.

In this article, we focus on another aspect that is missing in the classical wavelet framework and that plays an equally important role in audio processing. A large class of audio signal components are periodic or pseudoperiodic. The class includes the sounds from melodic musical instruments, singing and voiced segments of speech. These signals are characterized by a defined and stable pitch, possibly altered by slow vibrato or glissando. While pitched signals are compactly represented by harmonic components or bands in Fourier domain, there is no counterpart in classical wavelet analysis. Microfluctuations due to temperature drift, excitation noise and human factors make these signals deviate from strict periodicity even when the pitch is stable. In Figure 2, one can appreciate the fluctuation of the waveform in the stationary part of a constant pitch violin sound. Any signal in which the waveshape of the periods is constant sounds definitely synthetic and unpleasant to our ears. This continues to be true even if the initial and final transients (attack and decay) are exactly reproduced, as in wavetable synthesis. An important quality factor in the synthesis of pitched sounds is therefore the capability to suitably represent period-to-period fluctuations. This can be done in the time domain by comparing waveshapes or in time-frequency by comparing the short-time spectra. Respectively, these two approaches lead to the pitch-synchronous (PSWT) and harmonic band wavelet transforms (HBWT).

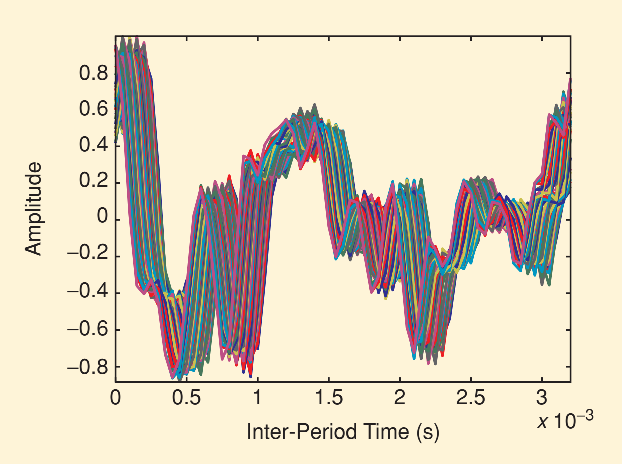
THE PITCH-SYNCHRONOUS WAVELET TRANSFORM

The tuning fork of our method is drawn from a pitch-synchronous representation of signals. With reference to Figure 3(a), a signal can be partitioned into a sequence of overlapping blocks by applying a sliding window $w(k)$. If the window w has finite length, then the subsignals $x_m(k) = x(k)w(k - mP)$ also have finite length and can be processed, e.g., transformed, independently using a finite number of operations. If desired, each finitely transformed signal can undergo the inverse transformation. If $w(k)$ satisfies the property:

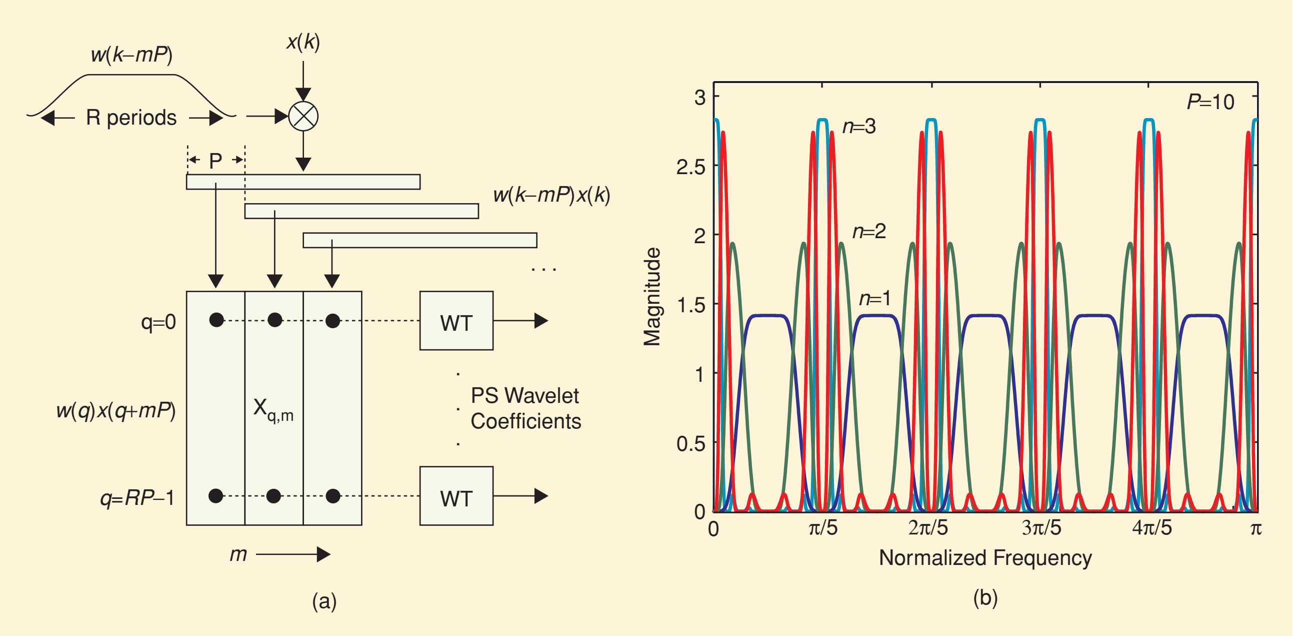
$$\sum_m w^2(k - mP) = 1 \quad \text{for any } k, \tag{2}$$

where P is an integer, then the original signal can be recovered by overlap add (OLA) methods, i.e., by multiplying the blocks by the window and by summing the overlapping repositioned blocks.

The interest, in our context, is that when the signal $x(k)$ has average period P then each block subsignal contains a certain number of periods, with the oldest period getting out of



[FIG2] Varieties of period waveforms in a violin sound.



[FIG3] (a) Timeline of the computation of the PSWT. (b) Magnitude Fourier transform of PS wavelets and scaling sequence, with rectangular window w and $P = 10$.

the picture and a new period popping in as the window is shifted in time exactly by one period. If, as is often the case for voiced sounds, the waveshape of one period is very similar to the waveshape of the next period, then it may be interesting to compare them, e.g., by averaging and differencing. The average is a mean period waveshape while the difference gives us the fluctuations of the waveshape. To realize this idea it is convenient to time shift the subsignals so that their nonzero samples are all aligned and start from the origin. We can define the pitch-synchronous matrix

$$X_{q,m} = x(q + mP)w(q), \tag{3}$$

with $M = RP$ rows ($q = 0, 1, \ldots, M - 1$) and theoretically infinite columns, each containing R periods of the signal weighted by the window samples. Looking through the rows of the matrix one can track the time evolution of a specific sample in the period.

As pointed out previously, the discrete-time finite scale wavelet expansions decompose the signal into a trend plus fluctuations at several scales. If we apply an array of wavelet transforms along the rows of the pitch-synchronous representation matrix (3), then we are able to represent the signal periods in terms of a locally periodic trend plus fluctuations from the periodic trend. The result can be neatly expressed into a signal expansion referred to as the pitch-synchronous wavelet transform (PSWT).

The timeline of the PSWT computation is shown in Figure 3(a). In the exceptional case $R = 1$, where the window w is constrained to be rectangular, the pitch-synchronous representation is equivalent to the polyphase representation of the signal and we obtain an orthogonal basis [12].

The discrete time Fourier transform (DTFT) of the pitch-synchronous wavelets and scaling sequences are comb shaped and tuned to the pitch of the signal [12], as in the diagram in Figure 3(b). The combs are quite peculiar: while the peaks of the scaling sequences are adjusted to the harmonics of the signal, the peaks of the wavelets form sidebands of the harmonics. The larger the scale, the narrower the sidebands and the closer the sidebands approach the harmonics without ever reaching them. Hence, the projection of the signal onto the space spanned by the scaling sequences obtains a more harmonic signal, i.e., devoid of most of the frequency content in between the harmonics. On the contrary, by projecting the signal onto the space spanned by the wavelets, the harmonics are canceled and noise and transients are revealed. Moreover, the fast fluctuations mostly contribute to energy in wavelet subspaces at small scales, which occupy frequency bands further away from the harmonics. The slow fluctuations occupy frequency bands closer to the harmonics and are mostly represented by large scale wavelets. Thus, the PSWT is useful for separating the noisy components of the signal from the harmonic content. For example, one can extract the bow noise in a violin sound by projecting the signal onto the pitch-synchronous wavelet subspaces [12].

The HBWT generalize the PSWT by treating each harmonic component separately. The PSWT can also be generalized to accommodate time-varying pitch [12]. One way is to extend the number of rows of the pitch-synchronous matrix $X_{q,m}$ to an integer multiple of the longest period of the signal expressed in number of samples. The shorter periods will produce gaps in the matrix that can be filled in by arbitrary values. The extra values produced are discarded in the synthesis.

COSINE MODULATED FILTER BANKS

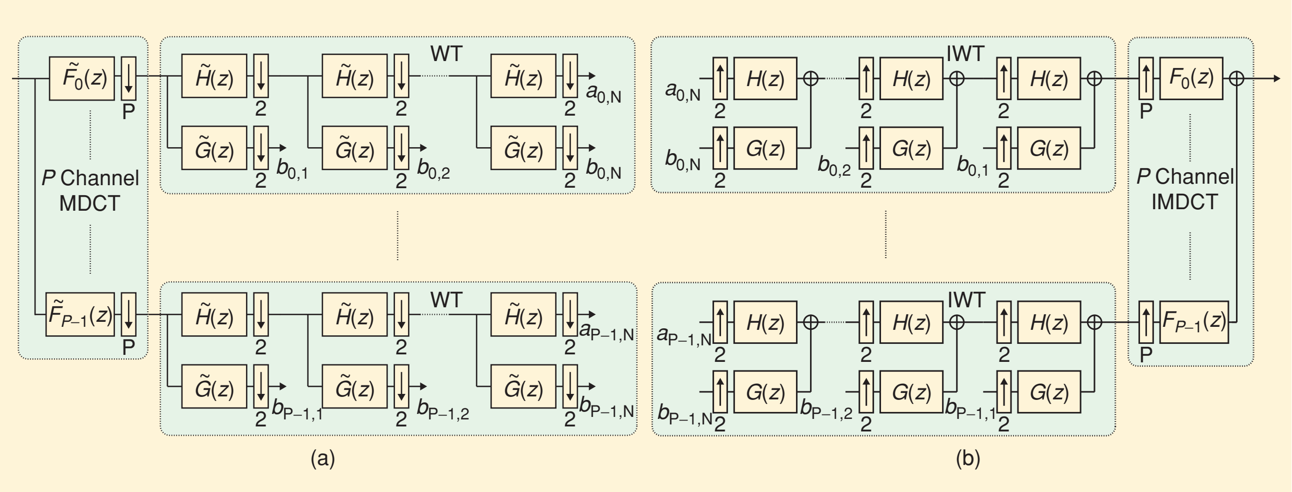
Local cosine bases fulfill Gabor's dream of compactly covering the time-frequency plane with uniform resolution elements of finite uncertainty: the logons. Time-frequency efficient orthonormal bases are obtained by modulating a window function by means of real sinusoids characterized by two peaks in the frequency domain [5].

The type of modulation needed can be expressed in terms of discrete cosine transforms (DCTs) or discrete sine transforms (DSTs) [13]. These representations have been invented and reinvented in the context of multirate filter banks and audio and image coding, including the MPEG standards [14]. Several terms to denote them have been coined. Among the others, modified DCT (MDCT) [15], and cosine modulated filter banks (CMFB) [16]. In the case of DCT-IV type of modulation, the discrete-time signal $x(k)$ is expanded in terms of a set of sequences obtained by shifting a cosine modulated window [15]. A popular choice is the sine window $w(k) = \sin(\frac{\pi}{2P}(k + \frac{1}{2}))$; $k = 0, 1, \ldots, 2P - 1$. Other choices are the modulated sine window used in Vorbis [17] or the Derived Kaiser Bessel window used in Dolby ac-3 [18] and MPEG-4 AAC [14], [19].

The signal expansion can be computed by means of a P channel multirate filter bank with synthesis impulse responses $f_p(k)$ and resampling factor P. The impulse responses $\tilde{f}_p(k)$ of the analysis filters are obtained by time reversal.

THE HARMONIC BAND WAVELET REPRESENTATION

The HBWT [20] is the spectral counterpart of the PSWT. The idea is that if the waveforms of adjacent periods of a pseudoperiodic signal are similar, then their short-time spectra are also similar. Therefore, instead of representing averages and differences in the time domain, one can represent similar quantities in the frequency domain. The advantage is the reduced sensitivity of the representation to the perfect synchronization of the period waveforms. Since the short-time Fourier spectra are complex quantities of which magnitude and phase have physical meaning, it is more convenient to compare real spectra obtained through DCT. This is obtained by synchronizing the number of channels of the CMFB to the number of samples in a signal's period. In turn, this form of pitch-synchronous MDCT can be regarded as the pitch-synchronous representation (3) cascaded by a frequency sampled DCT transformation of each windowed signal segment. Averages and differences at several scales are well represented by wavelets, as we have discussed in the PSWT case. Hence, the harmonic band wavelets are obtained by cascading a local cosine representation with a wavelet representation. The analysis and synthesis structures for computing the HBWT and its inverse are shown in Figure 4.



[FIG4] (a) The HBWT filter bank as a cascade of a P channel analysis CMFB with P wavelet filter banks. (a) The analysis structure. (b) The synthesis structure. The filters in the analysis section $\tilde{F}_p(z)$, $\tilde{H}(z)$ and $\tilde{G}(z)$ are time-reversed versions of the synthesis filters $F_p(z)$, $H(z)$ and $G(z)$, respectively.

The harmonic band (HB) wavelets $\hat{\psi}_{p,n,m}(k)$ are the impulse responses of the cascade of the wavelet synthesis filter bank with the synthesis CMFB. The frequency domain relationship between HB wavelets and ordinary wavelets is the following:

$$\hat{\Psi}_{p,n,m}(\omega) = F_p(\omega)\Psi_{n,m}(P\omega), \qquad (4)$$

where $F_p(\omega)$ are the frequency responses of the synthesis CMFB. It is easy to show that the HB wavelets are proportional to filtered versions of the PS wavelets.

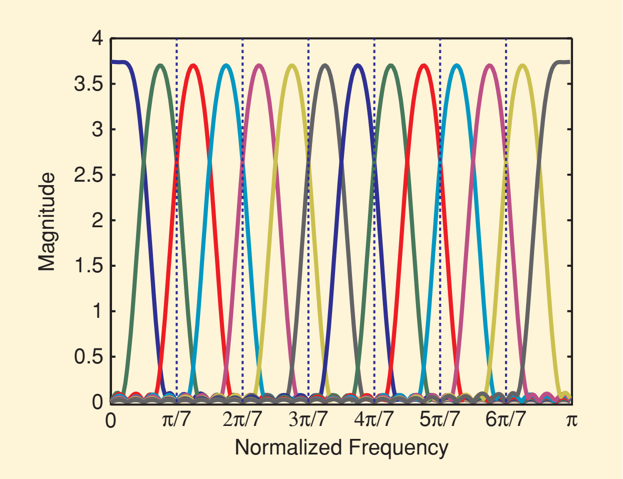
In the frequency domain, the cosine modulated sequences peak at frequencies $(p + (1/2))(\pi/P)$. As shown in Figure 5, if P equals the number of samples in a signal's period, then each harmonic peak at frequency $2\pi r/P$ is essentially represented by two MDCT components, those with $p = 2r - 1$ and $p = 2r$, for $r = 1, 2, \ldots, \lfloor (P - 1/2) \rfloor$. Moreover, the dc component is represented by the channel $p = 0$ and, when P is an even integer, the highest harmonic component at frequency π is solely represented by the channel $p = P - 1$. Apart from these exceptions, each harmonic term is represented by two sidebands.

In principle, each sideband of the harmonics is further split into several subsidebands by the wavelet decomposition. However, since the filters of the CMFB are not ideal, the subsidebands on the two sides of the harmonics overlap in the frequency domain. This result is reported in Figure 6, where the Fourier transforms of the HB wavelets pertaining to a single harmonic term are individually diagrammed for the lower and upper sidebands. Further overlap occurs for the upper sideband (lowest scale) of a harmonic and the lower sideband (lowest scale) pertaining to the next harmonic. As in the PSWT representation, the sidebands further away from the harmonics represent fast fluctuations, while those lying close to the harmonics represent slow fluctuations. However, in the HBWT we are able to separately analyze each harmonic band. In other words, in the frequency domain each HB wavelet corresponds to a single two-sided tooth of the PS wavelets' comb.

FRACTAL ADDITIVE SYNTHESIS

In our sound synthesis method, we drive the synthesis by means of the results of the analysis of any given sound that we would like to reproduce. Since the HBWT and its inverse form a perfect reconstruction scheme we could just feed the synthesis channels with the coefficients computed by the analysis. However, for sound transformation purposes, we also need a sufficiently flexible model that allows us to reproduce, with the least amount of parameters, pitch shifted as well as longer or shorter versions of the same sound, with the possibility to add dynamics and expression.

In the FAS scheme, the harmonic components are synthesized via the HBWT scaling sequences which, in the frequency domain, are shaped as overlapping narrow-band sidebands of the harmonics. Each pair of adjacent harmonic sidebands can be seen as a generalized sinusoidal term. These pseudoharmonic components are completed by a set of wider sidebands covering



[FIG5] Magnitude Fourier transform of the cosine modulated basis sequences (dotted lines) and harmonics (solid lines) of a P-periodic signal ($P = 14$).

the spectrum portion that lies in between the harmonics. The bandwidth of these sidebands is nonuniform, becoming wider and wider as we move away from the harmonics. These sidebands are in charge of reproducing noise and stochastic fluctuations present in instrumental sounds.

The method we propose is somehow related to the consolidated tradition of a wide number of STFT-based models [21]. In the more general context of synthesis by analysis methods, these models recently enjoyed a considerable success. Actually, the STFT by itself is a rather nonflexible and inefficient tool from the audio compression point of view. However, different models have been proposed as an evolution of the STFT, with the aim of exploiting the computational efficiency of the STFT while adapting the inner organization of data to the object of the analysis, i.e., a sound with a spectrum presenting both relevant peaks and noisy bands [22], [23]. One of these models is given by the generalized Sinusoidal Model introduced in [24]. To obtain a sinusoidal representation of the sound, an accurate detection of the spectral peak tracks in the STFT domain is performed [25]. Finally the peaks and their phases are organized as time-varying sinusoidal tracks. An evolution of the sinusoidal model is the sinusoidal plus residual model or spectral modeling synthesis (SMS) [26]. The introduction of a residual and of a distinction between a deterministic component and a stochastic component of sound makes the model much more flexible and efficient with respect to the previous ones, while maintaining good sound fidelity. In this sense, the SMS represents a consolidated system for audio synthesis and coding, including time-varying scenarios. The weak point of SMS is that the stochastic component is defined simply as the difference between the original signal and the sinusoidal resynthesis. The residue is synthesized as white noise whose spectrum is shaped by means of a filter obtained from an approximation of the whole spectral behavior. This approximation is derived by means of a linear interpolation of the magnitude spectrum of the residue itself. The approach presents some limitations in the synthetic sound quality.

The FAS technique contains an explicit model for the noisy components of sound, which represents an evolution with respect to all of the STFT-based models and it can be seen as a significant improvement for high quality sound reproduction. In this section, we illustrate the FAS model in its components: noise and pseudoharmonics. Our tour starts from the $1/f$ noise, whose characteristics inspired the construction of our model. Next, the synthesis of the stochastic components by means of the pseudoperiodic $1/f$ model is described. A method for the synthesis of the pseudoharmonics in terms of envelopes and phase functions in the HBWT domain is illustrated. Reduction of the synthesis parameters via perceptual criteria and coding is then discussed.

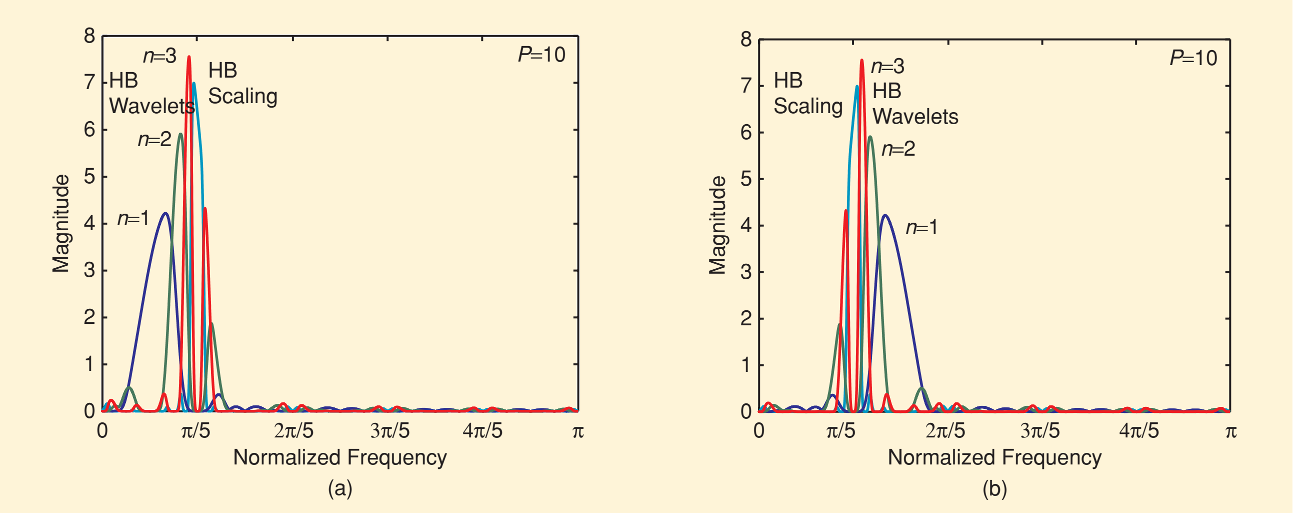
PINK NOISE

Flicker noise or $1/f$ noise, also known in the audio community as pink noise, is a very particular type of random signal with a long history in the tradition of experimental sciences [27]. In fact, a long list of natural phenomena or human activities contain or are governed by a $1/f$ behavior. The term $1/f$ originates from the observation of time series $x(k)$ by means of the periodogram $|\hat{X}_N(2\pi f)|^2$, obtained by taking the magnitude square of the DFT of a finite length N sample [28]. When plotted on log-log axes, the spectrum estimate shows a linear behavior, i.e., the periodogram decays as an inverse power of frequency:

$$\mathrm{E}\left\{|\hat{X}_N(2\pi f)|^2\right\} \approx \frac{\sigma_x^2}{|f|^\gamma}\,, \tag{5}$$

for some exponent γ and nonzero finite constant σ_x, where the symbol E denotes expectation. However, no wide-sense stationary process can have a $1/f^\gamma$ power spectrum density since the spectrum would not be integrable, leading to the infrared catastrophe for $\gamma \geq 1$ or to the ultraviolet catastrophe for $\gamma \leq 1$.

Methods to define and generate $1/f^\gamma$ processes have been devised based on fractional integration of zero-mean Gaussian white noise [29], which, with a proper correction term [30],



[FIG6] Harmonic Band wavelets pertaining to the first harmonic of period $P = 10$ samples. (a) Lower sidebands ($p = 1$). (b) Upper sidebands ($p = 2$).

yields a stationary increment process termed fractional Brownian motion (fBm).

A fundamental property of $1/f$ noise is its fractal behavior. The sample paths of an fBm process $B_\gamma(t)$ can be shown [30], [31] to satisfy

$$B_\gamma(\alpha t) \overset{\mathrm{D}}{=} \alpha^{(\gamma-1)/2} B_\gamma(t), \tag{6}$$

where the equality is in distribution. This means that, except for an amplitude factor, the statistical properties are scale-invariant, i.e., the process is statistically self-similar with fractal dimension $D = 2 - (\gamma - 1)/2$.

The nonstationary and self-similar characteristics of $1/f$ noise call for proper analysis and synthesis methods in which time and scale are variables of the representation [31]. The results in [32] show that the coefficients obtained by expanding fBm on an orthogonal dyadic wavelet basis are zero-mean and approximately uncorrelated both along and across scale, provided that a sufficiently regular wavelet is employed in the analysis. A wavelet based synthesis model was proposed for the $1/f$ processes [33], which avoids the use of long convolutions in generating fBm by fractional integration. One simply adopts uncorrelated noise sources as wavelet coefficients with scale dependent variances. The result is a nonstationary process whose time-average spectrum is approximately $1/f^\gamma$.

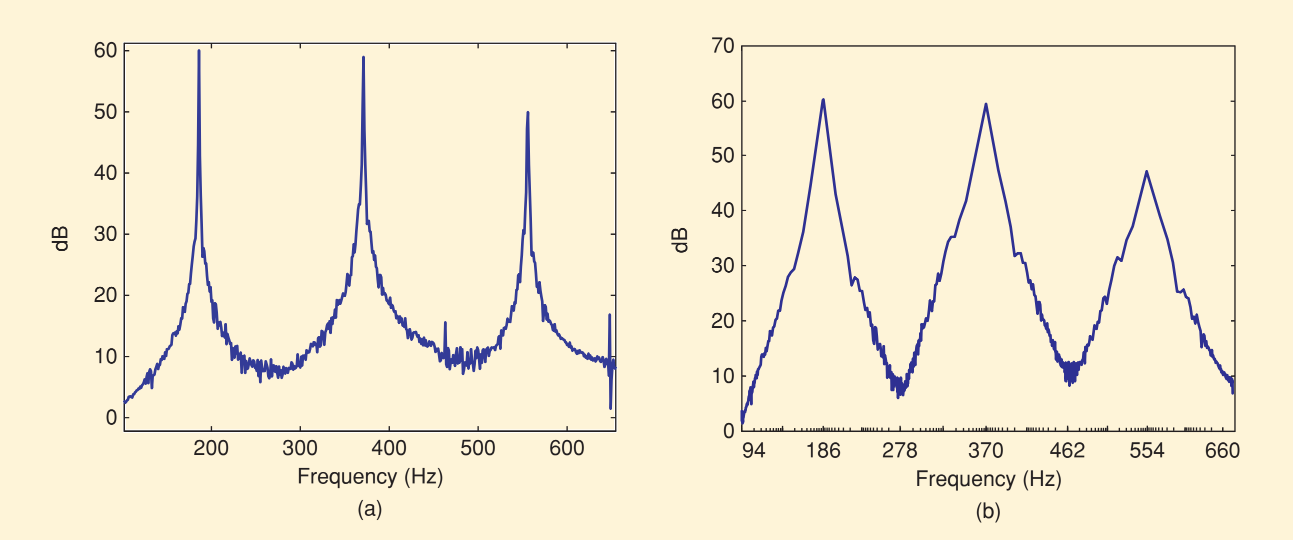
In principle, the sole parameter γ is sufficient to control the slope of the $1/f$-shaped power spectrum of the synthetic signal or to determine the variances of the synthesis coefficients for each wavelet subband. The $1/f$-like behavior of the wavelet subbands is visible in Figure 1.

SINGING FRACTAL NOISE

As already discussed, real-life voiced sounds presenting an approximately periodic structure show many fluctuations from period to period with respect to a strict reiteration of some archetype waveform. In other words, these sounds are pseudoperiodic signals. The presence of micro-fluctuations with respect to a pure periodic behavior results in spectra of the kind of Figure 7(a), where power decays somehow regularly as we move away from the harmonic peaks. In Figure 7(b), the same spectrum is plotted using a locally logarithmic frequency scale arranged on the harmonic grid. There, it is clearly visible how the distribution of the energy has an approximately $1/f$ behavior around each harmonic peak: the curve looks fairly linear in log-log axes. Our idea is that these $1/f$-like spectral segments describe the stochastic long-term correlated components present in voiced sounds in speech and music, in a similar way as $1/f$ processes well represent chaotic systems that are strongly influenced by their past behavior. The question is how to define a model for these chunks of $1/f$ spectral segments arranged on a harmonic grid. In other words, the challenge is to find a proper method to manipulate ordinary $1/f$ processes to fit the different $1/f$-like sidebands of the harmonics. The synthetic $1/f$-like spectral segments will become the building block for assembling any kind of pseudoperiodic signal.

The main idea is to adopt collections of properly modulated finite band $1/f$ processes. This is the philosophy of the $1/f$ pseudoperiodic model introduced in [20], which generalizes a model based on the PSWT presented in [34]. By controlling the energy and the slope of each $1/f$ sideband, it is possible to adapt the spectra of these signals to that of any given real-life voiced-sound, containing not only harmonic components (the peaks) but also stochastic components (the peak sidebands). The HBWT is the natural tool for this construction.

From the analysis point of view, when the number of channels of the MDCT blocks of the HBWT structure in Figure 4(a) is tuned to the pitch of a voiced sound, the passband of each channel corresponds to a specific sideband of a harmonic peak. Thus, it becomes natural to analyze the different $1/f$ processes by means of wavelet expansions as described previously.



[FIG7] First three harmonics in the frequency spectrum of a French horn: (a) Linear frequency scale; (b) Locally log-log scale: the frequency axis is logarithmic around each harmonic.

The FAS model takes in consideration additional information extracted from the analysis given by the spectral contents of the HBWT coefficients. While in first approximation the coefficients can be considered as uncorrelated, the HBWT analysis of real-life voiced-sounds does show some correlation. Accordingly, in our model we color the synthetic white noise that we feed into the HBWT scheme by means of linear predictive coding (LPC). Moreover, to model deviations from a static and strict local $1/f$ behavior, the noise energy is modeled in a time-dependent and scale-dependent way, with a slight increase in the number of parameters.

The lower part of Figure 8(a) summarizes the procedure for the extraction of the set of parameters describing the behavior of the wavelet coefficients. Conversely, the lower part of Figure 8(b) represents the generation of the synthetic coefficients controlled by means of the parameters derived from the analysis.

THE HARMONIC COMPONENT

To make our method into a complete technique for the synthesis by analysis of voiced-sounds we also need to provide a model for the deterministic components. In the HBWT domain, the harmonic components of sound are represented by the HB residual scale coefficients $a_{p,N}(m)$ of Figure 4(a) and (b). To model these sequences we resort to a complexification of the HB scale coefficients in pairs of adjacent channels, corresponding to the two sidebands of each harmonic. We form the complex sequences:

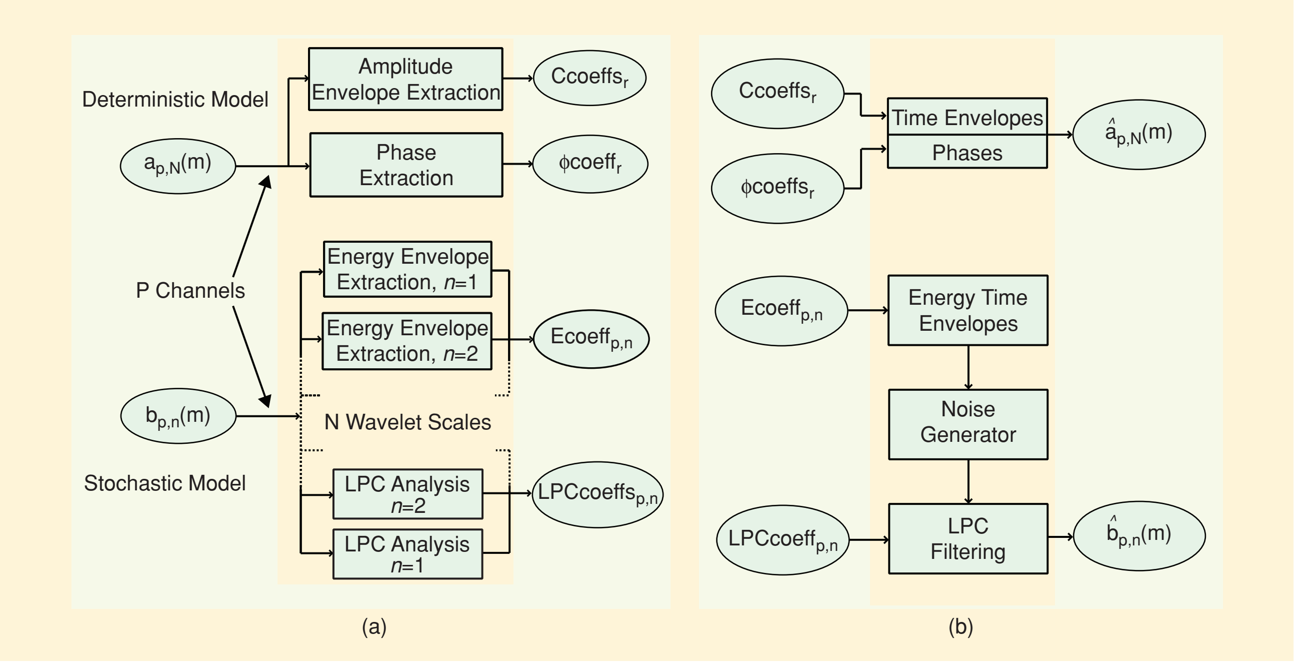
$$c_{r,N}(m) = a_{2r-1,N}(m) + ja_{2r,N}(m), \qquad (7)$$

where $a_{2r-1,N}(m)$ and $a_{2r,N}(m)$ are the HBWT scaling coefficients at scale level N, respectively, of the lower and upper sideband of the r^{th} harmonic, i.e., the outputs of the last low-pass filters in the p^{th} and $(p+1)^{th}$ channels of Figure 4(a), with $p = 2r - 1$. In polar form, we have:

$$c_{r,N}(m) = C_{r,N}(m)e^{j\varphi_{r,N}(m)}, \qquad (8)$$

where $C_{r,N}(m)$ is the amplitude and $\varphi_{r,N}(m)$ is the phase of the complexified coefficients.

Notice that these sequences could be equivalently extracted from the analysis filter bank in which each pair of adjacent channels of the CMFB is combined into a single complex filter. It is possible to show that the complex filter acts as an approximate Hilbert transformer on the harmonic bands: it combines the two sidebands into a single band centered on the harmonic and is essentially nonzero only for the positive frequencies. The result is further sharpened by the presence of the scaling comb in cascade with the CMFB in the HBWT analysis structure. One can show [35] that an input sinusoid of frequency differing by a small amount $\Delta\omega$ from the analysis frequency $2\pi r/P$ results into sinusoidal scaling coefficients with frequency $2^N P\Delta\omega$. Therefore, the phase of the complex coefficients has a linear trend, which is constant if the input is perfectly tuned with the frequency $2\pi r/P$. Furthermore, if the input sinusoid has a sufficiently smooth envelope, the magnitude of the complex coefficients is proportional to the envelope via a scaling factor that depends itself on $\Delta\omega$. Therefore, the magnitude of the complex coefficients (7) is closely related to a downsampled version of the envelope of the input. These results hold approximately true for pseudoperiodic signals containing several harmonic partials with arbitrary amplitude envelopes. As a consequence, the amplitudes $C_{r,N}(m)$ and the phases $\varphi_{r,N}(m)$ in (8) form smooth curves and nearly linear curves, respectively. These curves can be easily and



[FIG8] (a) FAS resynthesis parameter extraction from HBWT analysis. (b) Parametric resynthesis coefficient generation.

efficiently approximated by means of polynomial interpolation. In particular, we adopted linear splines as interpolating elements.

The results in the example of a clarinet note are shown in Figure 9(a) and (b), where spline polynomials of order 2 interpolate the amplitudes with nine knots and the phases with 11 knots. The polynomial approximation (solid line) is sufficient to make the synthetic sound undistinguishable from the original one. The behavior of the phase curves is reasonably linear in the stationary part. A slight temporary detuning is noticeable between coefficients 15 and 30. The nonlinearity found at the beginning and at the end of the curves corresponds to the attack and the decay transients, respectively.

In our model, the transients are perfectly reconstructed from the original analysis coefficients. The analysis-resynthesis procedure is performed only to be able to deal with the deterministic and the stochastic components separately. An extension of the method, analogue to transient modeling synthesis (TMS) [36] in SMS, is not available yet. Since both TMS and FAS operate in the DCT domain, an investigation in a similar direction could be worthy. An interesting by-product of FAS is, however, the possibility to perform transient detection directly on the HBWT analysis coefficients [37].

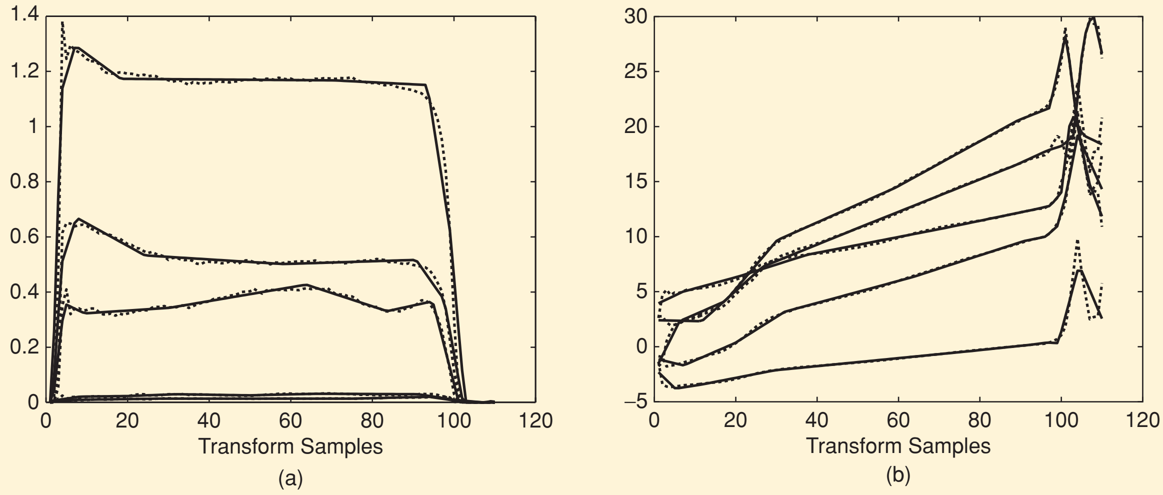
The method was tested on several notes of different musical instruments like clarinet, oboe, bassoon, trumpet, French horn, violin, viola and cello. All of them gave results comparable to the case of the clarinet, as it is possible to appreciate from the audio examples available at [38]. Figure 8(a) represents the extraction of the resynthesis parameters from the HBWT coefficients $a_{p,N}(m)$ and $b_{p,n}(m)$. The parameters $C\text{coeff}_r$ and φcoeff_r are the coefficients and knots of the polynomial interpolation of the complexified HB scale coefficients of the r^{th} harmonic. The parameters $E\text{coeff}_{p,n}$ are the interpolation coefficients of the energy envelopes of the HBWT coefficients $b_{p,n}(m)$. The parameters $LPC\text{coeff}_{p,n}$ are the filter coefficients resulting from the LPC analysis of the $b_{p,n}(m)$. Figure 8(b) illustrates the parametrically controlled generation of the synthetic HBWT coefficients $\widehat{a}_{p,N}(m)$ and $\widehat{b}_{p,n}(m)$.

Dealing with pitch evolving sounds, as in case of vibrato or glissando effects, appears to be a harder task in HBWT than in PSWT. Although the representation remains exact independently of mistuning, leakage of the harmonic components into the inharmonic bands can lead to inefficient and inaccurate FAS modeling. In [35] and [39], an approach based on a time-varying filter bank and on zero-padding of the pitch-synchronous components was presented, which is computationally very efficient but not totally satisfactory in terms of perceptual results. In [35], an alternate approach based on dynamic frequency warping techniques [40] is explored, which achieves extremely good perceptual results but is quite heavy from a computational point of view. A different approach based on dynamic time warping as in [25] is currently under study. Frequency warping also allows us to extend the FAS technique to inharmonic sounds [35].

FAS AS A DIGITAL AUDIO EFFECT

Pitch shifting by FAS is a straightforward process. It is sufficient to feed an appropriate synthesis filter bank tuned to the desired pitch with the analysis coefficients of the original sound. An example of implementation of FAS developed in Pure Data, a visual language for sound synthesis and processing [41], can be downloaded from [42]. The example of the viola gives the idea of the potentiality of FAS pitch shifting in terms of sound quality. It is sufficient to analyze four samples (one for each string) to synthesize the whole range.

Time scaling is also an immediate operation if one considers constant pitch sounds or even sounds with a regular vibrato. As far as the transient is preserved, then the amplitude and phase envelopes of the harmonic components and the amplitude envelopes of the noisy components can be arbitrarily rescaled and modified. As already discussed, the general case of pitch evolving sounds does not have a solution comparable to that provided by the SMS method, for example [26], which nowadays forms a versatile and complete corpus of sound analysis and



[FIG9] (a) Amplitude $C_{r,N}(m)$ and (b) phases $\varphi_{r,N}(m)$ of the complex HB scale coefficients (dotted line) for $r = 1, \ldots, 5$ and their spline interpolation (solid line) in a clarinet sound.

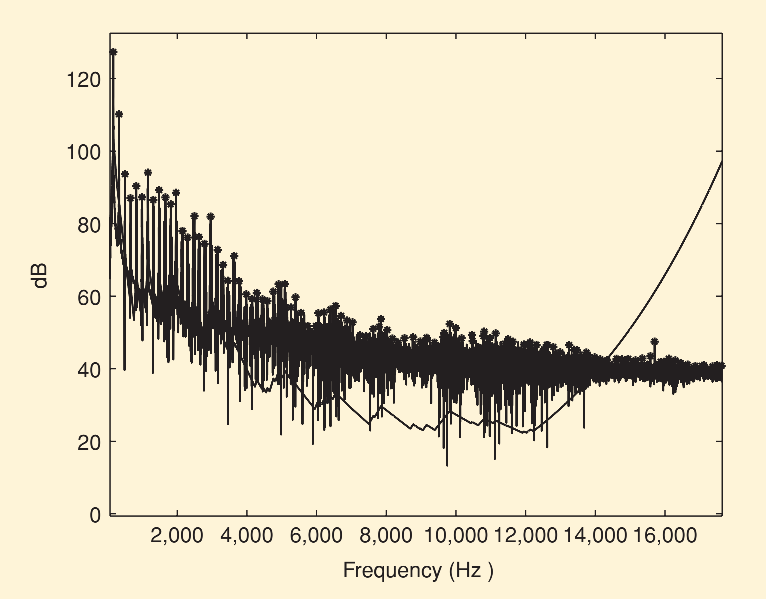
synthesis tools usable in many different applications [43]. Up to now, FAS is a complete model developed at core level.

As already underlined, the model of the harmonic components is independent from the model of the noise. Thus, expression by means of dynamic amplification of the excitation noise can be incorporated. Additionally, since FAS allows to analyze and resynthesize each harmonic independently, spectral modification connected to expression is easily implemented.

EMBEDDING PERCEPTUAL MODELS. FAS AS A TOOL FOR DATA COMPRESSION

Psychoacoustic criteria can be taken into account to further reduce the number of synthesis parameters for data compression purposes. Absolute threshold of hearing, localization theory and masking effects are the main principles on which various compression algorithms are based, like the ones included in the MPEG-1 standard. The localization theory describes how the inner ear and, more precisely, the cochlea works as a selective filter with nonuniform resolution.

From a perceptual point of view, the frequency bands into which the cochlea filter bank is subdivided correspond to the so called critical bands [44]. A sound pertaining to a critical band can annihilate the effect of other sounds belonging to the same critical band, provided that the sound pressure level of the latter is below a certain measurable threshold. The energy gap between a tone and its masking threshold is the signal to mask ratio (SMR). In addition, the masking effect extends itself to the neighbor critical bands in an asymmetric way, following a certain decay law called spreading of the masking threshold [44]. In the case of complex sounds, each sufficiently outstanding spectral peak works as a simple tone, generating a masking effect in the near frequency range. Thus, to evaluate the global masking effect of a complex sound one has to consider the masking contribution of all of the spectral peaks present within a certain time interval.



[FIG10] Psychoacoustic mask for an E2 legato cello note.

In our case, the definition of a global masking threshold over the whole frequency range allows us to discard all the HBWT coefficients that produce perceptually irrelevant sound components. The global masking threshold is obtained from the superposition of the absolute hearing threshold and of SMR with spreading, pertaining to all the relevant partials of the sound.

As an example, Figure 10 represents the global mask for a cello sound. All of the HBWT coefficients corresponding to the spectral subbands whose levels are smaller than the masking threshold can be discarded. In the case of the cello sound of Figure 10, the noisy components in the range from 3000 Hz to 14000 Hz are above the masking threshold. To achieve high quality reproduction one needs to reproduce this part of the sound. In our experiments, we considered 1–4 s samples at a sampling rate of 44.1 KHz. For each sample, we computed an order 10 LPC analysis for the stochastic component of each harmonic r and scale n. For all of the sound samples that we tested we used ten equispaced coefficients to approximate the energy envelope of the random generated stochastic coefficients $\widehat{b}_{p,n}(m)$. Taking into account the number of knots and coefficients of the spline interpolation of the complex deterministic coefficients $c_{r,N}(m)$, the compression rate that we obtain is approximately of 20:1. Similar results are obtained with all of the traditional musical instruments mentioned in the previous section. These results are extremely appealing compared to the MPEG coders, where at a digital audio level only psychoacoustic criteria are considered. An exception is given by the MPEG-4 harmonic and individual lines plus noise (HILN) coding scheme. In the context of structured audio coders, the FAS method can be seen as a transparent and effective tool for the encoding and the compression of the stationary part of voiced-sounds, including a highly efficient and hi-fi model for the noisy components.

A complete codec algorithm could be obtained by adding dynamic bit allocation and entropy coding stages to the FAS scheme. By joining the capabilities of FAS in terms of both pitch shifting, noise component modeling and data compression, a straightforward application would be to provide a super sampler. Transfer or store huge amount of samples of any kind of pitched instrument with the most various expressive palette becomes much less demanding with FAS. An example is given by the viola case mentioned previously.

CONCLUSION

In this article, we outlined time-frequency techniques that, when brought together, form the basis for FAS. The method is based on an exact orthogonal transform, the HBWT. The FAS model can be seen as an intelligent spectrogram, i.e., as a spectrogram where the frequency bins (the HBWT subbands) are adapted to the spectrum of the analyzed sound by tuning the number of channels P to the period of the sound. Local cosines replace the classical windowed complex exponentials, whose main property is that the basis elements form sidebands of the harmonics rather than being centered on the harmonics themselves. The wavelet transform nonuniform frequency subdivision characteristic is exploited to synthesize each sideband of the harmonic peaks by means of colored noise, generating an approximate pseudoperiodic $1/f$ behavior. The pseudoharmonics are modeled by narrow sidebands whose weights can be generated by means of amplitude envelopes and phase functions in a complexified HBWT domain. We also showed how perceptual criteria can be employed to reduce the number of synthesis parameters. The possibility of independently controlling the noisy components and the harmonic terms and the definition of parameters such as volumes and envelopes for all of the components independently provide powerful tools for processing voiced-sound for sound design purposes.

AUTHORS

Pietro Polotti (polotti@sci.univr.it) is with the Video Image-Processing and Sound group (VIPS) at the University of Verona, Italy. He is also professor of electronic music at the Conservatoire Giuseppe Verdi of Como, Italy. From 1998 to 2002, he worked at the Ecole Polytechnique Fédérale de Lausanne (EPFL) in Switzerland. From the EPFL, he obtained a Ph.D. in Communication Systems with a thesis on audio coding based on the wavelet transforms. Recently, his research interests moved towards auditory display and sound design. With VIPS, he collaborates with different European research projects, as the Coordination Action Sound to Sense, Sense to Sound (S2S^2) and the project Closing the Loop of Sound Evaluation and Design (CLOSED).

Gianpaolo Evangelista (giaev@itn.liu.se) received the laurea in physics (summa cum laude) from Federico II University of Naples, Italy, in 1984 and the M.Sc. and Ph.D. degrees in electrical engineering from the University of California, Irvine, in 1987 and 1990, respectively. He has been with the Centre d'Etudes de Mathématique et Acoustique Musicale (CEMAMu/CNET), Paris, France; the Microgravity Advanced Research and Support (MARS) Center, Naples; Federico the ACoustics and Electronics group at Federico II University of Naples, and the Laboratory of Audiovisual Communications at the Swiss Federal Institute of Technology (EPFL), Lausanne. Since 2005 he has been a professor at the Linköping University, Sweden, where he heads the Sound Technology research group. He is the author or coauthor of more than 80 journal or conference papers and book chapters. He has been a recipient of the Fulbright fellowship. His interests include audio, music, and image processing; coding; wavelets; and multirate signal processing.

REFERENCES

[1] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1997.

[2] R. Balian, "Un principe d'incertitude fort en théorie du signal ou en mécanique quantique," *C.R. Acad. Sci.*, vol. 292, no. 20, pp. 1357–1362, 1981.

[3] F. Low, "Complete sets of wave packets," in *A Passion for Physics—Essay in Honor of Geoffrey Chew*, C. DeTar, Ed. Singapore: World Scientific, 1985, pp. 17–22.

[4] D. Gabor, "Theory of communication," *J. IEE*, vol. 93, no. 26, pp. 429–457, 1946.

[5] I. Daubechies, S. Jaffard, and J.L. Journé, "A simple Wilson orthonormal basis with exponential decay," *SIAM J. Math. Anal.*, vol. 22, no. 2, pp. 554–572, 1991.

[6] I. Daubechies, *Ten Lectures on Wavelets*, (CBMS-NSF Conf. Series in Applied Math, vol. 61). Philadelphia: SIAM, 1992.

[7] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[8] F. Keinert, *Wavelets and Multiwavelets*. London: Chapman & Hall, CRC, 2003.

[9] T. Blu, "Iterated filter banks with rational rate changes connection with discrete wavelet transform," *IEEE Trans. Signal Processing*, vol. 41, pp. 3232–3244, Dec. 1993.

[10] G. Evangelista, "Dyadic warped wavelets," *Adv. Imaging Electron Phys.*, vol. 117, pp. 73–171, Apr. 2001.

[11] G. Evangelista and S. Cavaliere, "Discrete frequency warped wavelets: Theory and applications," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 874–885, Apr. 1998.

[12] G. Evangelista, "Pitch synchronous wavelet representations of speech and music signals," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3313–3330, Dec. 1993.

[13] K.R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. New York: Academic, 1990.

[14] Moving Picture Experts Group [Online]. Available: http://www.chiariglione.org/mpeg/

[15] H.S. Malvar, *Signal Processing with Lapped Transforms*. Norwood, MA: Artech House, 1992.

[16] R.D. Koilpillai and P.P. Vaidyanathan, "Cosine-modulated FIR filter banks satisfying perfect reconstruction," *IEEE Trans. Signal Processing*, vol. 40, pp. 770–783, Apr. 1992.

[17] Vorbis [Online]. Available: http://xiph.org/vorbis/

[18] Advanced Television Systems Committee [Online]. Available: http://www.atsc.org/standards.html

[19] M. Bosi and R.E. Goldberg, *Introduction to Digital Audio Coding and Standards*. Norwell, MA: Kluwer, 2003.

[20] P. Polotti and G. Evangelista, "Analysis and synthesis of pseudo-periodic 1/f-like noise by means of wavelets with applications to digital audio," *EURASIP J. Appl. Signal Processing*, vol. 2001, no. 1, pp. 1–14, Mar. 2001.

[21] M. Portnoff, "Time-frequency representations of digital signals and systems based on short time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 1, pp. 55–69, Feb. 1980.

[22] X. Rodet and P. Depalle, "Spectral envelopes and inverse FFT synthesis," in *Proc. 93-rd Conv. Audio Engineering Society*, San Francisco, CA, Oct. 1992.

[23] M. Goodwin and M. Vetterli, "Time-frequency signal models for music analysis, transformation, and synthesis," in *Proc. IEEE Int. Symp. Time-Frequency Time-Scale Analysis, (TFTS'96)*, Paris, France, 1996, pp. 133–136.

[24] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[25] T. Abe and M. Honda, "Sinusoidal model based on instantaneous frequency attractors," *IEEE Trans. Audio, Speech Language Processing*, vol. 14, no. 4, pp. 1292–1300, 2006.

[26] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, "Spectral processing," in *DAFX—Digital Audio Effects.* New York: Wiley, 2002, pp. 373–438.

[27] M. Keshner, "1/f noise," *Proc. IEEE*, vol. 70, no. 3, pp. 212–218, Mar. 1982.

[28] V. Solo, "Intrinsic random functions and the paradox of 1/f noise," *SIAM J. Appl. Math.*, vol. 52, no. 1, pp. 270–291, 1992.

[29] J.A. Barnes and D.W. Allan, "A statistical model of flicker noise," *Proc. IEEE*, vol. 54, pp. 176–178, Feb. 1966.

[30] B.B. Mandelbrot and H.W.V. Ness, "Fractional Brownian motion, fractional noises and applications," *SIAM Rev.*, vol. 10, pp. 422–436, Oct. 1968.

[31] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 910–917, Mar. 1992.

[32] A. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 904–909, Mar. 1992.

[33] G. Wornell, *Signal Processing with Fractals: A Wavelet-Based Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[34] G. Evangelista, "Pseudo-periodic 1/f-like noise," in *Proc. IEEE Int. Symp. Time-Frequency Time-Scale Analysis (TFTS'96)*, Paris, France, June 1996, pp. 121–124.

[35] P. Polotti, "Fractal additive synthesis," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2003 [Online]. Available: http://library.epfl.ch/theses/?nr=2711

[36] T.S. Verma and T.H.Y. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Comput. Music J.*, vol. 24, no. 2, pp. 47–59, 2000.

[37] P. Polotti and G. Evangelista, "Multiresolution sinusoidal/stochastic model for voiced-sounds," in *Proc. Digital Audio Effects Conf. (DAFx-01)*, Limerick, Ireland, Dec. 2001, pp. 120–124.

[38] "FAS sound examples" [Online]. Available: http://staffwww.itn.liu.se/~giaev/FAS_examples.html

[39] P. Polotti, "Fractal additive synthesis: A pitch-synchronous extension of the method for the analysis and synthesis of natural voiced-sounds," in *Proc. Int. Computer Music Conf. (ICMC-02)*, Göteborg, Sweden, Sept. 2002, pp. 387–392.

[40] G. Evangelista and S. Cavaliere, "Audio effects based on biorthogonal time-varying frequency warping," *EURASIP J. Appl. Signal Processing*, vol. 2001, no. 1, pp. 27–35, Mar. 2001.

[41] M. Puckette, "Pure data," in *Proc. Int. Computer Music Conf. (ICMC'96)*, Hong Kong, Aug. 1996, pp. 269–272.

[42] "FAS implementation in pure data" [Online]. Available: http://www.s2s2.org/docman/task,cat,view/gid,91/Itemid,65/

[43] CLAM [Online]. Available: http://www.clam.iua.upf.edu

[44] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer-Verlag, 2006.