Fractal Additive Synthesis:  
A Pitch-Synchronous Extension of the Method for the 
Analysis and Synthesis of Natural Voiced-Sounds  

Pietro Polotti  
EPFL-INR-LCAV, Ecublens  
1015, Lausanne, Switzerland  

Abstract  
Real-life musical sounds contain both deterministic and stochastic components. The deterministic part provides the main timbre characteristics of a sound. It is in a sense the timbre archetype. In case of voiced-sounds this archetype consist of the harmonic partials and their relative amplitude. This model can be enriched by many different kinds of time modulations. By means of these modulations we are able to provide a more realistic synthetic sound. Nevertheless the inner dynamic of sounds, i.e., all the natural fluctuations with respect to an artificial not naturally evolving sound are contained in the stochastic components. Furthermore these components contain the noise due to the physical excitation system. All these noisy components are important in order to perceive a sound as a natural one.  

In some previous works [7]-[10] we developed a synthesis by analysis method called fractal additive synthesis. The fractal additive synthesis method allows one to analyze and synthesize the deterministic and the stochastic components of sounds separately. The synthesis process is driven by a set of perceptually meaningful parameters, which can be deduced from the analysis of real-life sounds. The most attractive feature of the fractal additive synthesis is that we are able to model not only the deterministic part of sounds but also the stochastic one at a very refined perceptual level and with a minimum amount of parameters controlling the synthesis process.  

The subject of this paper is an extension of the fractal additive synthesis method, in order to include the case of voiced-sounds with time-varying pitch. This will provide a model able to deal with the slight detunings that occur in natural voiced-sounds, as well as to deal, for instance, with vibrato effects. In order to do this we need to design a perfect reconstruction P-channel filter bank, whose number of channels P can change period by period according to the variation of the pitch of a voiced-sound. This filter design technique and its adaptation to the fractal additive scheme are the main subjects of this paper. The result is a sound synthesis technique able to provide synthetic sounds with natural pitch and timbre dynamics.  

Introduction  
The goal of this paper is a method for the analysis and the synthesis of voiced-sounds, which includes the case of voiced-sounds with varying pitch.  

One of the most challenging aspects of the analysis and the synthesis of voiced-sounds is the definition of a good model for the noisy part of sounds. Musical signals contain both deterministic and stochastic components. The deterministic part provides the pitch and the fundamental structure of a sound. This fundamental structure can be easily represented by means of a very restricted set of parameters. The stochastic part contains the "life of a sound", that is all the micro-fluctuations with respect to an electronic-like/non-evolving sound as well as noise due to the physical excitation system. The latter is of fundamental importance to feel a sound like a natural one.  

In a previous edition of the ICMC [10] we presented a method for the analysis and resynthesis of voiced-sounds, i.e., of sounds with harmonic-peak spectra. In order to implement this method, we defined a new analysis/synthesis tool: the Harmonic-Band Wavelet Transform (HBWT). Our technique allows one to separate the stochastic components of voiced-sounds from the deterministic components and to resynthesize the stochastic components independently from the deterministic components by means of a restricted set of perceptually relevant parameters. Our work was based on the assumption, experimentally verified, that the energy distribution of the sidebands of voiced-sounds is approximately shaped as powers of the inverse of the distance from the closest partial. In other words their behavior in the neighborhood of the partial is of the kind $1/(f/f_n)$, where $f_n$ is the $n^{th}$ harmonic. As a consequence we demonstrated that wavelet transforms are a "natural" tool to extract, decompose into subbands and resynthesize the noisy sidebands of the harmonic
spectral peaks [9]. This approximately 1/f behavior around the harmonics is due to the natural fluctuations of voiced-sound with respect to a pure deterministic "artificial sounding" sound as well as to the noise due to the excitation system. The fact that the sidebands have an approximately 1/f behavior around the harmonic peak is related to self-similarity properties, i.e., something related to a fractal behavior [5]. This is the reason where the name fractal additive synthesis comes from [8]. Also this is the reason why the wavelet transform, thanks to its self-similar properties, apply to this kind of spectra in a natural way [6].

In a more recent work [7] we also developed a model for the deterministic HBWT analysis coefficients. This model is quite similar to sinusoidal modeling techniques, where instead of modeling the amplitude and the phase of the harmonics, we model the amplitude and the phase of the HBWT analysis coefficients relative to the harmonic peaks. Our method is thus a complete model for the representation of voiced-sounds, i.e., for the representation of both their deterministic and stochastic components.

The topic of this paper is an extension of the technique in order to be able to follow the natural evolution of the pitch of real life voiced-sounds, as well as more significant pitch modulation as in case of vibrato effects. The main challenge of this task is the design of a perfect reconstruction filter bank with time-varying number of channels $P$, where $P$ is tuned to the sound pitch period by period. Our previous harmonic-band wavelet model was constrained to the case of fixed harmonic spectra. In the HBWT analysis $P$ was tuned to the average pitch of the analyzed voiced-sound. In order to set our method free from this limit we introduce a perfect reconstruction pitch synchronous $P$-channel filter bank design technique, obtained by modifying an already existing filter design method [3] and adapting this filter bank design to our fractal synthesis scheme. The filter design procedure that we are going to introduce is based on the polyphase representation of $P$-channel filter banks [1, 2] and on a simple factorization of the polyphase matrix into a product of elementary matrices.

The output of the time-varying $P$-channel filter banks will be processed in a similar way as in the previous fractal additive synthesis scheme. I.e., every channel output will be wavelet transformed and the analysis coefficient will be modeled back by means of set of parameters with the same meaning as in the fixed-pitch version of the method.

1. Fractal Additive Synthesis: a Review

The Harmonic-Band Wavelet Transform (HBWT) is the mathematical "backbone" of fractal additive synthesis. In order to implement the discrete-time HBWT we need a set of pass-band filters and a set of two quadrature mirror filters (QMF) organized in cascade at the output of each filter of the pass-band filter bank (see Fig. 1). The pass-band filters separate the sidebands of the harmonic peaks one by one. The cascades of QMF filters separate the deterministic and stochastic components of each sideband and analyze them from a multiscale (fractal) point of view. More precisely each cascade of QMF filter implements an ordinary wavelet transform. The pass-band filter bank is a cosine-modulated filter bank that means that the whole set of pass-band filters is obtained by means of a cosine modulation of a single low-pass prototype [2]. Figure 1 represents the filter
schemes implementing the HBWT and its inverse. \( P \) is "tuned" to the length in samples of the average period of the analyzed voiced-sound, i.e., to the pitch of the sound. The coefficients \( \alpha_{p,m} \) are the analysis coefficients relative to the deterministic components, while the coefficients \( b_{p,m} \) are the coefficients relative to the noisy components. The all structure is perfect reconstruction, while providing critical sampling.

The fractal additive synthesis method consists in a parametrical representation of the analysis and resynthesis (or simply synthesis by itself) coefficients. This representation is perceptually meaningful and intuitive. In other words we define a set of parameters for each harmonic peak and sideband approximating the behavior of the coefficients corresponding to the deterministic component and controlling the statistical behavior of the coefficients of the noisy components. More in details we need to compute:

1. The time envelopes of the synthesis coefficients of the deterministic components.
2. The phase of the synthesis coefficients of the deterministic components.
3. The time envelopes of the energy of the coefficients of the noisy components.
4. A set of LPC filters controlling the spectral behavior of the synthesis coefficients of the noisy components.

The parameter extraction process is shown in figure 2. In figure 3 we show the fractal additive synthesis scheme. For more details see [7, 9].

Fig. 3: The fractal additive synthesis scheme. 4 set of parameters control the generation of the HBWT synthesis coefficients \( \alpha_{p,m} \) and \( w_{p,n,m} \) corresponding to the deterministic and stochastic components of the synthetic sound, respectively.

2. Efficient Cosine Modulated Filter Banks

We will now introduce a method for designing a time variant \( P_{\text{max}} \)-band filter bank able to switch to an arbitrary number of bands \( P_m \leq P_{\text{max}} \) while maintaining perfect reconstruction and critical sampling during the transitions. More in general we will be able to switch to a number of bands \( P_m \leq P_{\text{max}} \) at any time \( t_m \) with \( t_m = \sum_{m=0}^M P_m \) for some sequence \( P_0, P_1, \ldots, P_M \leq P_{\text{max}} \). The \( P_m \)'s are the time sequence of number of \( m = 0 \) bands of the time-varying filter bank. Our goal is to obtain a pitch synchronous version of the HBWT scheme that is a scheme able to deal with voiced-sounds with varying pitch as for instance the one shown in figure 5. The sequence of channels \( P_0, P_1, \ldots, P_M \leq P_{\text{max}} \) will be tuned to the pitch variations of the kind of those of figure 6.

In order to do that we consider a new kind of cosine-modulated filter bank [3]. The whole design method is based on the polyphase representation of filter banks [1, 2]. The design technique consists in a factorization of the polyphase matrices representing the filter bank. This factorization decomposes the polyphase matrix into a set of elementary sparse matrices, a diagonal matrix and a cosine-modulation matrix.

The great and general advantage of modulated filter banks is that the all set of filters can be designed by means of a simple modulation of the FIR baseband prototype. From a computational point of view this mean that we need to run an optimization only of the prototype frequency response, in order to approach as much as possible the case of an ideal low-band filter. More than that the design method that we are going to introduce allows one to deduce all the filters banks for any \( P_m \leq P_{\text{max}} \) number of channels from the single prototype of the \( P_{\text{max}} \) case. Furthermore the filter banks we are going to design are biorthogonal. Biorthogonality, with respect to the orthogonal case, allows one to choose a partially arbitrary length of the impulse response. This is an evident advantage in...
terms of the optimization of the filter frequency response. We start with the case of time-invariant number of bands and we will introduce the time-varying case in the next section. We first subdivide the input sound $s[l]$ into length-$P$ vectors, where $P$ is the number of channels of the analysis and synthesis filter bank. In this way we obtain a polyphase representation of $s[l]$:

$$s[m] = (s_0[m], \cdots, s_{P-1}[m])^T$$  \hfill (1)

with

$$s_i[m] = s[mP + i]$$  \hfill (2)

In the z-domain this can be written as:

$$S(z) = [S_0(z), \cdots, S_{P-1}(z)]^T$$  \hfill (3)

where $S(z)$ is the z-transform of the $s_i[m]$.

Then we consider a type-2 polyphase representation of an analysis $P$-channel filter bank. That is we design a matrix $A(z)$ whose elements are given by:

$$[A(z)]_{p,l} = \sum_{m=0}^{\infty} k_p(mP + P - 1 - l)z^{-m},$$

$$p,l = 0, \cdots, P-1$$  \hfill (4)

where the $k_p$ are the impulse responses of the $p^{th}$ filter of the $P$-channel filter bank. A type-1 polyphase representation provides the inverse $P$-channel cosine-modulated filter bank and we denote the corresponding polyphase matrix as $R(z)$. The analysis/resynthesis polyphase matrices $A(z)$ and $R(z)$ satisfy the perfect reconstruction relationship:

$$R(z) \cdot A(z) = I$$  \hfill (5)

The great advantage of this formulation is the extreme simplicity of the design consisting, as we will see, of a simple product of elementary matrices. The computational efficiency of the polyphase representation of the filter banks is a further advantage. Also, as we will see, the formalism and implementation of filter banks with time-varying number of channels will be extremely easy and straightforward.

We first write $A(z)$ as:

$$A(z) = C \cdot F(z)$$  \hfill (6)

where $C$ is the $P \times P$ Discrete Cosine Transform (DCT) type IV matrix, whose elements are given by:

$$C_{p,l} = \cos \left( \frac{\pi}{P} (p + 0.5)(l + 0.5) \right),$$

$$0 \leq p, l \leq P-1$$  \hfill (7)

and $F(z)$ is a matrix with a sparse "bi-diagonal form", containing the polyphase components of the prototype:

$$F = \begin{bmatrix} f_{0,0} & 0 & f_{0,p} \\ 0 & \ddots & 0 \\ f_{P,0} & 0 & f_{P,P} \end{bmatrix}$$  \hfill (8)

The $F(z)$ can be factorized into a product in the following way:

$$F(z) = \prod_{i=1}^{0} L_i(z) \cdot D,$$  \hfill (9)

where the matrices $L_d(z,m)$ have the following form:

$$L_i(z) = J + \text{diag} \left( l_0, \cdots, l_{P/2-1}^*, 0, \cdots, 0 \right) \cdot z^{-1}$$  \hfill (10)

and $D$ is a diagonal matrix:

$$D = \text{diag} \left( d_0, \cdots, d_{P-1} \right).$$  \hfill (11)

The number of matrices $v$ is directly related to the impulse response length of the filters. It is thus related to the number of parameters at disposal to optimize the frequency responses of the low-pass prototype filter in order to have a good approximation of the ideal case.

The inverse or resynthesis polyphase matrix is given by:

$$R(z) = F^{-1}(z) \cdot C^{-1},$$  \hfill (12)

where $C^{-1}$ is the inverse of the $P \times P$ DCT type IV matrix (7) and $F^{-1}$ can be easily computed as:

$$F^{-1}(z) = D^{-1} \cdot \prod_{i=1}^0 L_i^{-1}(z),$$  \hfill (13)

where

$$L_i^{-1}(z) = J - \text{diag} \left( 0, \cdots, 0, l_{P/2-1}^*, \cdots, l_0^* \right) \cdot z^{-1}$$  \hfill (14)

and $D^{-1}$ is the inverse of (11).

In this way we obtain a PR analysis/resynthesis scheme, where the analysis and resynthesis operations are given by:

$$Y(z) = A(z) \cdot S(z) \quad \text{and} \quad \hat{S}(z) = R(z) \cdot Y(z),$$  \hfill (15)

respectively.

3. The Time-Varying Case

The main goal of this work is to provide a pitch synchronous extension of the fractal additive
synthesis method. This means the introduction of a new class of wavelets: the Pitch-Synchronous HBWT (PSHBWT). The PSHBWT can be viewed as an extension of the pitch-synchronous wavelet transform [4]. We first need to design a perfect reconstruction time-varying cosine modulated filter bank and then we will adapt it to the structure of fractal additive synthesis. An extension of the P-channel filter design of the previous section to the case of filter banks with time-varying number of bands is possible and it will provide the necessary tool for the implementation of the PSHBWT. In order to do this we have first to modify the DCT matrix in order to make it suitable for modulating a matrix \( \mathbf{F}(z) \) of size \( P \times P \), while generating a number \( P' \) of bands. We can obtain this by splitting into two symmetric parts the modulation matrix (horizontally the analysis and vertically the synthesis one) and inserting in between the two parts a zeros matrix of size \( P' \times P' \) in the analysis case and of size \( P' \times P' \) in the synthesis case. We obtain the following matrices:

\[
\mathbf{C}_a(m) = \begin{bmatrix} \mathbf{C}_a(m) & 0 \\ 0 & \mathbf{C}_a(m) \end{bmatrix}
\]

for the analysis case and

\[
\mathbf{C}_r(m) = \begin{bmatrix} \mathbf{C}_r(m) \\ 0 \\ \mathbf{C}_r(m) \end{bmatrix}
\]

for the synthesis case, with

\[
\mathbf{C}_a(m) \cdot \mathbf{C}_r(m) = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{bmatrix}
\]

The matrix \( \mathbf{F}(z) \) has also to be adapted to the \( P \) case. In order to do this we need to change the matrices \( \mathbf{L}(z) \) in the following way:

\[
\mathbf{L}(z) = \mathbf{J} + \text{diag} \left( \mathbf{I}_{P}, \ldots, \mathbf{I}_{m/2} \right) \cdot z^{-1}
\]

Notice that the matrices \( \mathbf{L}(z) \) (and thus the resulting matrices \( \mathbf{F}(z) \)) are of size \( P \times P \) also for the \( P \) case.

Finally we modify the input sound in order to subdivide it into vectors of length \( m/2 \) in order to fit the \( \mathbf{F}(z) \) matrices, while using only \( m \) input samples at each period \( m \). In order to do this we simply insert \( m/2 \) zeros in the middle of the time-varying polyphase vector of the input sound \( s(t) \), i.e., we build the following vectors:

\[
s[m] = (s_0[m], \ldots, s_{m/2}[m], 0, \ldots, 0, s_{m/2}[m], \ldots, s_{m-1}[m])
\]

with

\[
s_0[m] = \sum_{j=0}^{m-1} P_j + 1
\]

Thanks to the sparse bi-diagonal structure of the matrices the zeros inserted into the input vectors arrive together at the modulation matrix \( \mathbf{C}(m) \) and in a position corresponding to the position of the inserted zeros in the \( \mathbf{C}(m) \) itself.

The output vectors \( y[m] \) of the analysis filter bank \( \mathbf{A}(z) \) (see Fig. 4) have length \( m \) and can be written as:

\[
y[m] = (y_0[m], \ldots, y_{m-1}[m])
\]

\[
\text{Cosine-modulated filter bank}
\]

\[
\text{Wavelet transform}
\]

\[
\text{Inv. wavelet transform}
\]

\[
\text{Inverse cosine-modulated filter bank}
\]

\[
\text{Fig. 4 Polyphase representation of a cosine-modulated filter bank with time-varying number of channels: a) the analysis bank, b) the synthesis bank. The scheme includes also the wavelet transformation of each channel. The whole structure implements the Pitch Synchronous Harmonic-Band Wavelet Transform (PSHBWT).}
\]

391
By means of a simple zero-padding we can make all of them of the same length \( P_{\text{max}} \). In this way the \( P_{\text{max}} \) sequences of analysis coefficients can be injected into the wavelet filter banks as in the normal HBWT analysis. Finally the same parameter extraction as in the fractal additive synthesis can be performed. Due to the pitch-synchronous filter bank, the PSHBWT analysis coefficients corresponding to the deterministic part now represent the period by period time-varying harmonic peaks of the analyzed sound. In a similar way the PSHBWT analysis coefficients corresponding to the stochastic components represent the time-varying noisy sidebands of the harmonic peaks.

The advantage of the pitch-synchronous version of our method is evident and confirmed by the experimental results. The introduction of the PSHBWT, as already said, allows one to deal with sounds with varying pitch of the kind of the flute with vibrato of figure 5, whose pitch variations are shown in figure 6. More experiments have been done with other instruments, confirming the efficacy of the method.

Fractal additive synthesis allows one to control and reproduce the micro fluctuations present in real-life sounds so important from a perceptual point of view in order to hear a sound as a natural one. The method is a sort of additive synthesis where one adds not only the deterministic components of sounds but also the 1/f-like sidebands of the harmonic peaks reproducing the above mentioned stochastic micro-fluctuations. By means of the filter design technique introduced in this paper we extended the fractal additive synthesis method to the most common case of voiced-sounds with time-varying pitch. More precisely we adapted a new type of time-varying cosine-modulated filter banks to the HBWT scheme, obtaining what we called the Pitch Synchronous Harmonic-Band Wavelet Transform (PSHBWT). We showed how the fractal additive synthesis parametric model can be extended to PSHBWT analysis/synthesis scheme. The most appealing feature of this method is that it allows one to reproduce the stochastic fluctuations of voiced sounds with varying pitch by means of a very restricted number of parameters.

References