INHARMONIC SOUND SPECTRAL MODELING BY MEANS OF FRACTAL ADDITIVE SYNTHESIS

Pietro Polotti, Fritz Menzer

Audio Visual Communications Laboratory – LCAV
École Polytech. Féd. de Lausanne, Switzerland
pietro.polotti@epfl.ch,
fritz.menzer@epfl.ch

Gianpaolo Evangelista

Dept. of Physical Sciences

Univ. "Federico II", Naples, Italy
gianpaolo.evangelista
@na.infn.it

ABSTRACT

In previous editions of the DAFX [1, 2] we presented a method for the analysis and the resynthesis of voiced sounds, i.e., of sounds with well defined pitch and harmonic-peak spectra. In a following paper [3] we called the method Fractal Additive Synthesis (FAS). The main point of the FAS is to provide two different models for representing the deterministic and the stochastic components of voiced-sounds, respectively. This allows one to represent and reproduce voiced-sounds without loosing the noisy components and stochastic elements present in real-life sounds. These components are important in order to perceive a synthetic sound as a natural one.

The topic of this paper is the extension of the technique to inharmonic sounds. We can apply the method to sounds produced by percussion instruments as gongs, tympani or tubular bells, as well as to sounds with expanded quasi-harmonic spectrum as piano sounds.

1. INTRODUCTION

The FAS method allows one to separate the stochastic components of voiced sounds from the deterministic ones and to resynthesize both of them separately. Both these components are important from a perceptual point of view and contribute in different ways to a high-fidelity resynthesis. The deterministic part contains the structure of the sound. The stochastic part contains the real-life flavor, i.e. all the information relative to the random deviations with respect to a strictly deterministic representation of sounds. This part is necessary in order to avoid an “electronic-like” synthetic sound.

In [4] we demonstrated that the wavelet transform in its harmonic extension, i.e., the Harmonic-Band Wavelet Transform (HBWT) is a "natural" tool to separate, decompose and resynthesize both the deterministic components and the noisy sidebands of the harmonic spectral peaks. This decomposition allows one to represent the different components of the sound by means of a restricted set of parameters. These parameters, different for the deterministic and the stochastic parts, correspond to the two models that make the FAS an interesting method both for sound synthesis/processing and for data compression in the context of Structured Audio.

The new result that we present in this paper is the extension of the method to the inharmonic case. Our previous HBWT model was confined to the harmonic spectrum case. The time-frequency plane tiling was strictly harmonic. This is a major limitation and makes the method not usable for a large class of sounds, for instance all the sounds produced by percussive instruments. The spectra of many of these instruments show relevant peaks (see Figure 1). These peaks are the partials or deterministic components of the sound and can be sinusoidally modeled. These partials also show an approximately 1/*f* spectral behavior around the peak as in the harmonic case. These 1/*f*-shaped, spectral sidebands are the stochastic components. Thus the same stochastic model used in the harmonic case can be employed. It is therefore reasonable to find a way to extend the FAS method to sounds with spectra of the kind of Figure 1.

The main problem addressed in this paper is to provide a more flexible analysis/synthesis structure, which extends the FAS model to inharmonic sounds. In order to do this we abandon the Perfect Reconstruction (PR) structure provided by the HBWT and resort to a non-PR scheme, which is able to deal with aperiodic spectra like the one in Figure 1. A non-PR structure leads to aliasing problems and artifacts in the resynthesis. These artifacts are minimized by the filter design procedure and optimization. This part is discussed in detail in Section 3.

2. FRACTAL ADDITIVE SYNTHESIS (FAS) OVERVIEW

From experimental evidence we know that the spectra of voiced sounds are composed by harmonic peaks and sidebands with an approximately 1/*f* behavior around the harmonic peaks. In [4] we introduced and defined a new class of stochastic processes that we called the pseudo-periodic 1/*f* noise. The main idea of the FAS is to model voiced sounds by means of pseudo-periodic 1/*f* processes. This is also called the 1/*f* pseudo-periodic model. More precisely, the 1/*f* pseudo-periodic model represents the harmonic peaks f_k of a voiced sound and their sidebands as approximately $1/|f - f_k|$ segments in the neighborhood of each f_k. In the discrete case, k=1,…, P/2, where P is equal to the pitch in samples of the discrete time voiced sound. These segments reproduce the 1/*f*-like behavior of the sidebands of the harmonics and the harmonics themselves. In the FAS method each spectral

segment is processed separately and decomposed by means of a Wavelet Transform (WT). The idea of using the WT for the analysis and synthesis of the sidebands of the harmonic peaks comes from the existing analogies between the wavelet frequency domain dyadic subdivision and the 1/*f*-like spectral behavior of the sidebands [5, 6]. Also, the name fractal additive synthesis comes from the selfsimilarity properties of both the wavelets and the 1/*f* noise (see [7,8] and [9,10], respectively).

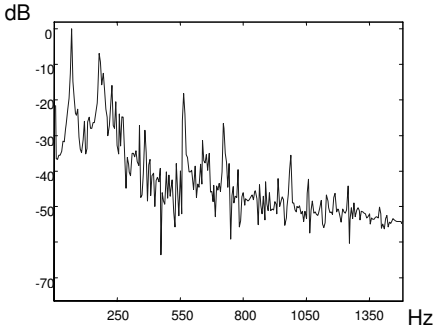

Figure 1: *Magnitude Fourier transform of a gong.*

The separation of the $1/|f-f_k|$-like segments, i.e., of the sidebands, is achieved by means of a Modified Discrete Cosine Transform (MDCT) [11]. The whole structure given by the MDCT followed by the WT forms the Harmonic-Band Wavelet Transform (HBWT). The HBWT is the tool for the analysis, processing and synthesis of the pseudo-periodic 1/*f* noise, i.e., of our model for the representation of voiced sounds. The HBWT was defined in [1] and [4].

The MDCT is implemented by means of the filters:

$$g_{p,r}^{P}(l) = g_{p,0}^{P}(l - rP), \qquad \begin{matrix} p = 0,....,P-1; \\ l = 0,....,2P-1 \end{matrix} \qquad r \in \mathbf{Z} \tag{1}$$

and

$$g_{p,0}^{P}(l) = W(l)\cos\left(\frac{2p+1}{2P}\left(l - 2P + \tfrac{1}{2}\right)\pi - (-1)^{p}\tfrac{\pi}{4}\right),$$

where the length *2P* lowpass prototype impulse response $W(l)$ satisfies specific conditions [12]. In

Figure 2 and 3 the analysis and synthesis filter banks implementing the HBWT are shown. The terms $G_p^P(z)$ and $G_p^P(z^{-1})$ are the Fourier transform of (1) and of its anticasual version, respectively. In the FAS the number of channels *P* is "tuned" to the pitch of the sound that one wants to analyze and/or synthesize. In this way the passbands of the filters $G_p^P(z)$ correspond to the sidebands of the harmonic peaks of the analyzed/synthesized sound and at the output of each channel *p* of the MDCT we obtain a downsampled version of one particular sideband (for the details, see [4]). The following WT subdivides each harmonic sideband as shown in Figure 4. This subdivision follows in a "natural" way the 1/*f*-like behavior of the sidebands and allows to distinguish in a straightforward way between the spectral peak corresponding to the harmonic, i.e., to the deterministic component and the wavelet subbands corresponding to the noise present in real-life sounds, i.e., to the stochastic components.

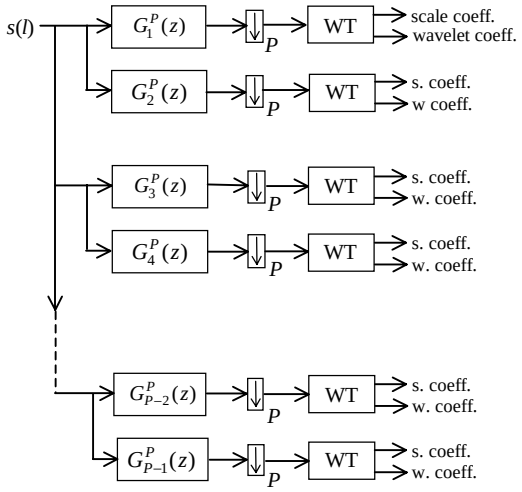

Figure 2 : *HBWT analysis filter bank. The filters* $G_p^P(z)$ *implement the P-channel MDCT. A WT is applied to the output of each MDCT channel. At the final output of each channel p we obtain the wavelet and scale coefficients of the decomposition of the p^th sideband corresponding to the k^th harmonic with* $k = \lceil p/2 \rceil$.

In [1] and [2] we presented two different methods for modeling the HBWT coefficients by means of a limited number of perceptually meaningful parameters. The first method is aimed to modeling the wavelet coefficients that is the coefficients corresponding to the stochastic components. In [1] we introduced a periodic version of the results concerning the analysis and the synthesis of 1/*f* noise by means of the WT by employing properly energy scaled white noise as coefficients [5, 6]. Our stochastic model consists in extracting the energy envelopes of the wavelet coefficients for each subband. The white noise resynthesis coefficients are then modulated in amplitude by means of these envelopes. Furthermore, from the analysis of the HBWT coefficients and from listening to the resynthesis results, it appears clear that approximating the coefficients with rough white noise is not sufficient. There is a small degree of autocorrelation in the wavelet coefficients, which is hearable in the resynthesis. In order to reproduce this autocorrelation we introduced a further modeling step, performing an LPC analysis of the wavelet coefficients. The white noise resynthesis coefficients are then 'colored' by means of the resulting AR filters.

The second model is based on similar principles as those underlying the sinusoidal models [13, 14]. The idea is to exploit the smoothness of the curves formed by the scale coefficients coming out from the analysis of voiced sounds [2]. We consider pairs of scale coefficient sets corresponding to the two sidebands of one harmonic and use these pairs of coefficients to generate complex numbers (the left sideband coefficients provide the real part and the right sideband coefficients provide the imaginary

part of the new complex coefficients). In this way we are able to represent these coefficients in terms of an envelope and a phase. Since the original scale coefficients form smooth curves, both the envelope and the phase are also smooth curves. In particular the phase is nearly linear. The same models are applied to the resynthesis coefficients of the non-harmonic case.

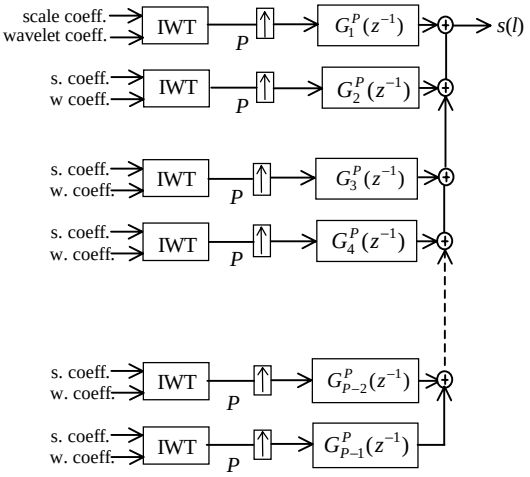

Figure 3 : *Inverse HBWT filter bank. The same notation of Figure 2 holds. The IWT blocks represent the Inverse Wavelet Transform.*

3. EXTENDING FAS

The method reviewed in the previous section is now extended to the analysis and synthesis of inharmonic sounds, i.e., sounds with an aperiodic waveform but with relevant spectral peaks. Such peaks are the partials of the sound and correspond to deterministic components, i.e., for instance, to vibrational modes of membranes, metallic or wooden surfaces and bars occurring in many instruments as gongs, tam-tams, tympani, bells, vibraphones, marimbas and many other percussive instruments. Also, these peaks are not harmonically spaced but they show an approximately $1/|f - f_n|$ shape as in the harmonic case. Here f_n denotes the frequency of the n^{th} partial.

The idea is thus to use the same two models as in the harmonic case in order to control the resynthesis coefficients of the partials peaks and of their sidebands, respectively. The principle of the wavelet subband subdivision is therefore maintained. What we need to change is the MDCT section of the method, which is limited to a uniform (harmonic) segmentation of the frequency domain.

In order to free the method from its harmonic-grid limit, we design a non-uniform cosine-modulated filter bank (CMFB), where the bandpass filters are adaptively tuned to the non-equally-spaced spectral peaks of an inharmonic sound. The system is not PR, in the sense that only the spectral regions corresponding to the main peaks are analyzed and the overlap of the filter passbands is chosen empirically according to the spectral peak distribution. As a result of the analysis we get sets of analysis coefficients plus some more or less relevant residue. The residue energy can be arbitrarily reduced at an expense of an increasing computational time. As already said, the 'main body' of the sound, including both the partial peaks and their noisy spectral sidebands, is analyzed in a similar way as in the case of the HBWT.

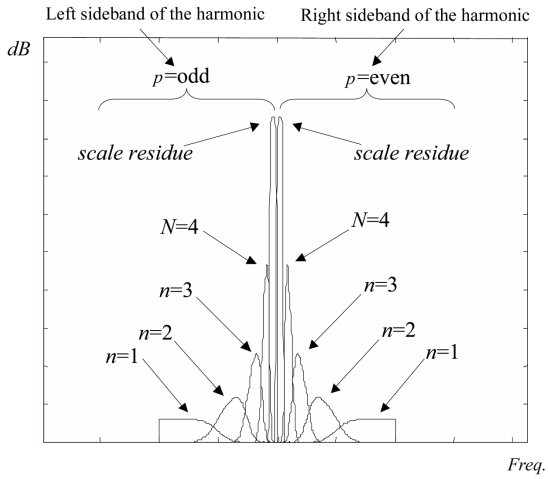

Figure 4 : *Magnitude Fourier transforms of the HBWT subband decomposition of a single harmonic. Left and right sidebands. In this case N=4 is the* (*arbitrary*) *maximum wavelet scale. The two residual bands* (*scale residue*) *containing the harmonic peak, i.e., the deterministic component, correspond, in the wavelet subdivision, to the scale function.*

3.1. Peak detection

A preliminary and fundamental step before the design of the inharmonic CMFB is the implementation of a good spectral peak estimation algorithm, in order to find the frequencies of the partials of the sound and consequently design the relative filters. With respect to a normal peak detector it is also necessary to define the optimal bandwidth of the filters that will subdivide the spectral range, taking into account not only the partial position but also the position of its two neighbors (see Figure 5). The goal of the partial detection algorithm is to find all the 'significant' peaks in the magnitude Fourier transform (FT) of a sound, where to be significant or not finally depends only on perceptual criteria.

As a first step we consider the average of the spectrogram frames of the sound. This is an easy and effective way to make the partials become more distinct and to get rid of the noise in the spectrum (Blackman-Tukey method [15]). In this 'cleaned' spectrum we perform the pick detection. The basic principle of the algorithm used in this work consists in comparing the magnitude of the candidate peak to a linear combination of the mean and standard deviation of a certain region R of the magnitude FT. The region R is chosen in different ways (Figure 6) in the neighborhood of the candidate peak itself. In order to make the algorithm more robust, different values of the coefficients of the linear combination and different criteria of

definition of the region R are considered and compared before designating a peak.

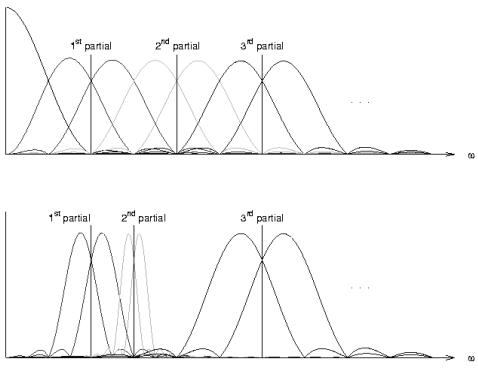

Figure 5 : *CMFB design*. a) *The harmonic case*. b) *The inharmonic case*.

As a second step we need to define a preliminary evaluation of the bandwidth. The considered estimators are the distance of the peak from the left and right neighbor peak positions (d_{left} and d_{right}, respectively) and the approximately $1/|f - f_n|$ shape of the sidebands of the n^{th} partial. As a simple evaluation criteria of the $1/|f - f_n|$ shape we take the length $d_{lthreshold}$ of an interval around the peak where the magnitude spectrum is above a certain threshold depending on the spectral characteristics of the sound (see Figure 7). The chosen bandwidth is the minimum among $d_{left}/2$, d_{right} /2 and $d_{lthreshold}$ /2. The parameter $d_{lthreshold}$ is important in order to maintain the method in the frame of the pseudoperiodic 1/*f* model, especially in the case of isolated spectral peaks.

3.2. Filter design

The design of a inharmonic CMFB requires the definition of the most appropriate 'hypothetic-pitch' for each detected partial peak. If the first partial of the inharmonic sound could correspond to a certain 'harmonic' k_1 of a 'hypothetical-pitch' P_1, we implement a whole P_1-channel MDCT filter bank by means of the set of filters $g_{p,r}^{P_1}(l)$ as given in (1) and we keep only the two filters with $p=2k_1-1$ and $p=2k_1$. As said in the previous subsection, the definition of the parameters P_1 and k_1 depends not only on the position of the partial but also on the position of the neighbor peaks. The shape of the sidebands is taken into account as well. This provides a preliminary estimate of the bandwidth π/P_1 of the P_1 -channel filter bank. But these are not the only criteria for the choice of P_1 and k_1. It is worthwhile to consider also an optimization procedure aimed to reduce as much as possible the aliasing occurring at the analysis of each partial. In general it can be shown that when a sinusoid at frequency $k\pi/P$ is analyzed by a *P*-channel cosine modulated filter bank, only the outputs of the $2k^{th}$ and $(2k\text{-}1)^{th}$ channels will be different from zero. Therefore, due to the fact that the filter bank is a perfect reconstruction filter bank, this sinusoid can be reconstructed without aliasing using only these two bands.

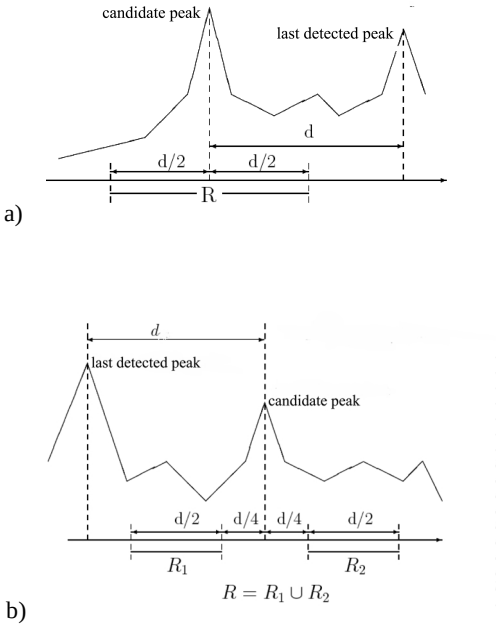

Figure 6 : *Two examples of the different criteria adopted for defining the region R , in which we search a peak.*

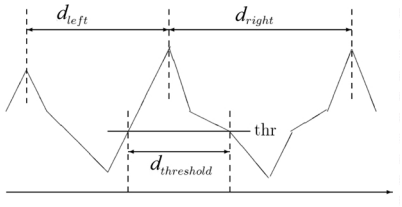

Figure 7 : *The parameters for the definition of the first estimate of the bandwidth (before the optimization) of the filters relative to one partial peak.*

This means that if the crossover frequency of the two filters passbands is well centered around the peak we can achieve a nearly aliasing-free reconstruction of the deterministic part of the sound, i.e. of the part where the aliasing effects are more relevant. Increasing the parameter *P* provides filters with narrower passbands. This obviously means a higher resolution in terms of distribution of the cross over frequencies of the filters and the possibility of getting arbitrarily close to the partial frequency. Then the goal of the optimization algorithm is a trade-off between the 'tuning' of the filters around the partial peak and a bandwidth large enough to include the sideband of the partial. In order to do this the following parameters are considered: the frequency of the partial, the preliminary estimate of the bandwidth, an interval for the variation of the bandwidth and an upper bound for the deviation from the frequency of the partial. The algorithm first calculates all the filters with bandwidths in

the given interval and takes the one, which differs the least from the desired frequency. If the difference does not fulfill the upper bound condition, then the program reduces the bandwidth until the condition is fulfilled. The same criteria are applied to define the parameters P_n and k_n of the other couple of filters, corresponding to the other partials.

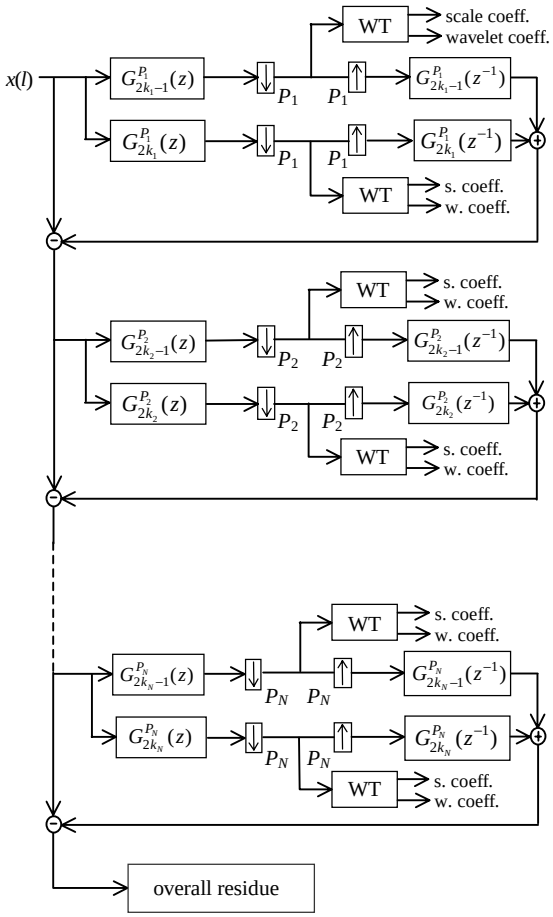

Figure 8 : *Analysis scheme. The index P_n refers to the P_n-channel filter bank chosen to analyze the n^{th} partial, n=1,….,N. The indexes k_n refers to the couple of filters selected from the P_n-channel FB 'surrounding' the partial peak. 'WT' denotes a wavelet transform.*

Once all the filters are defined, the inharmonic CMFB is implemented as in Figure 8. The structure of Figure 8 has the advantage of being PR at the condition of keeping the overall residue and adding it back to the reconstructed sound. More in detail, the filters $G_{2k_n-1}^{P_n}(z)$ separate the sidebands of the n^{th} partial. In other words the n^{th} partial is processed as the k^{th} harmonic of a hypothetic voiced sound with pitch P_n. Then each sideband is wavelet transformed and divided into subbands as in the harmonic case. The meaning of the upsampling of order P_n and of the filters $G_{2k_n-1}^{P_n}\left(z^{-1}\right)$ is to reconstruct both the n^{th} partial and the aliasing due to the downsampling of order P_n. In this way we keep track of the aliasing through the following partial analysis steps. Then, when we reconstruct the partials and add them up altogether with the overall residue, we are able to achieve time domain aliasing cancellation. As already mentioned, the overall residue can be arbitrarily reduced in energy by means of a recursive analysis at the cost of an increasing number of parameters.

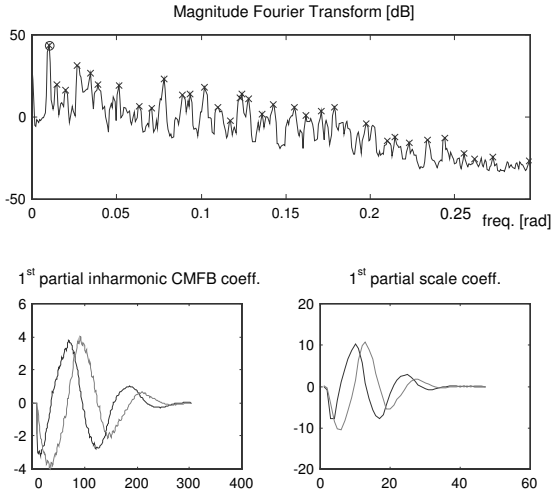

Figure 9 : a) *Magnitude Fourier transform of a gong. The 'x's denote the detected partials.* b) *The output of the two channels of the inharmonic CMFB corresponding to the first partials* (*the circled peak of figure* a). c) *The scale coefficients resulting from the wavelet analysis of the coefficients of figure* b.

Figure 9 represents some result of the peak detection applied to a gong (a) and the resulting scale coefficients of the analysis of the first partial (c). From the latter figure it is evident how the scale coefficients form smooth and slowly oscillating curves as in the harmonic case. The pseudo-sinusoidal model can thus be applied successfully also in the inharmonic case. The stochastic model for the 1/*f*-shaped sidebands of the partials holds as well both from a numerical and from a listening point of view.

4. EXPERIMENTAL RESULTS AND APPLICATIONS

We applied our method successfully to instruments with different degree of inharmonicity, ranging from very inharmonic sounds to quasi-harmonic sounds: gongs, tympani, tubular bells, and a piano. All sounds have been both partially reconstructed (without the residue) by means of the analysis coefficients and resynthesized by means of parametrically controlled coefficients. In some cases (gong and tubular bells) the synthetic sounds are hardly distinguishable from the original ones. Also the deterministic part and the different wavelet scale (stochastic) components were synthesized separately.

The method can also be viewed as a new synthesis technique. It is a sort of augmented additive synthesis. We can add an arbitrary number of partials, arbitrarily distributed in the frequency range. Furthermore we can control parametrically the shape of the partial sidebands, i.e., we can control the amount of

timbre dynamics and noisiness. A possible 'parameters scenario' for a FAS modulus could be to have a couple of parameters per partial: one for the amplitude and one for the 'noisiness', where the latter parameter would control the slope of the spectral sidebands of the partial. Inner parameters (also editable) could be the amplitude envelopes, the phase of the complex scale coefficients, the central frequencies and bandwidths of the inharmonic CMFB.

Also digital audio effects as pitch shifting, time stretching, noise-to-harmonic component ratio modifications are easily obtainable by means of interpolation and modulations of the parameters controlling the synthesis coefficients generation. The most interesting feature is that the two independent models for the stochastic components and for the deterministic components allow one to process them separately. The time-stretched and pitch-shifted samples of the gong and the tubular bell sound very realistic. Another effect we implemented was to transform a inharmonic sound into a harmonic one, i.e., to keep the partials and their natural behavior and resynthesize them by means of an artificial harmonic CMFB. We realized different versions of a harmonized tympani and gong, with different noisy sidebands widths. In general the method can be seen as a flexible sound processor allowing one to manipulate the spectrum of a sound in a perceptually meaningful way.

The method can also be viewed in terms of audio data compression. When taking into account psychoacoustic criteria, compression ratios of the order of 1/30 can be obtained just in terms of parametric representation, i.e., before any quantization and coding optimization.

5. CONCLUSIONS

An extension of the FAS to the case of inharmonic sounds has been introduced. The new method is not a perfect reconstruction method. Nevertheless a nearly aliasing free reconstruction of sounds can be achieved.

The analysis and resynthesis by means of the inharmonic FAS produces good results from a perceptual point of view. Not only the deterministic components of the sound are modeled and reproduced, but also the noisy components, which are important in order to perceive a sound as realistic. The results were good even for sounds, which are generally difficult to model, such as the case of a gong.

The method provides great flexibility in terms of sound processing. Experiments on the modifications of the synthesis parameters show that there are interesting applications in the fields of sound synthesis and digital audio effects for electronic music.

6. REFERENCES

[1] P. Polotti, G. Evangelista. 1999, "Dynamic Models of Pseudo-Periodicity", *Proceedings of the 99 Digital Audio Effects Workshop*, pp. 147-150, Trondheim, Norway.

[2] P. Polotti, G. Evangelista, "Multiresolution Sinusoidal/Stochastic Model for Voiced-Sounds", *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-01)*, Limerick, Ireland, Dec. 2001.

[3] P. Polotti, G. Evangelista, "Fractal Additive Synthesis by means of Harmonic-Band Wavelets", *Computer Music Journal*, 25(3), pp. 22-37, Fall 2001.

[4] Polotti, P. and G. Evangelista. 2001. "Analysis and Synthesis of Pseudo-Periodic 1/*f*-like Noise by means of Wavelets with Applications to Digital Audio", *EURASIP Journal on Applied Signal Processing*, Hindawi Publishing Corporation, pp. 1-14.

[5] G. W. Wornell and A. V. Oppenheim, "Wavelet-Based Representations for a Class of Self-Similar Signals with Applications to Fractal Modulation," *IEEE Trans. Inform. Theory*, Vol. 38, No. 2, pp. 785-800, March 1992.

[6] G. W. Wornell, "Wavelet-Based Representations for the 1/*f* Family of Fractal Processes," *Proc. IEEE*, Vol. 81, No. 10, pp. 1428-1450, Oct. 1993.

[7] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice Hall, 1995.

[8] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley –Cambridge Press, 1996.

[9] M. S. Keshner, "1/*f* Noise," *Proc. IEEE*, Vol. 70, No 3, pp. 212-218, March 1982.

[10] D. T. Gillespie, "The Mathematics of Brownian Motion and Johnson Noise," *Am. J. Phys.*, Vol. 64, pp. 225-240, March 1996.

[11] P. Vaidyanathan, *Multirate Systems and Filter Banks*. Upper Saddle River NJ: Prentice Hall Signal Processing Series, 1993.

[12] T. Q. Nguyen and R. D. Koilpillai, "The Theory and Design of Arbitrary-Length Cosine-Modulated Filter Banks and Wavelets, Satisfying Perfect Reconstruction", *IEEE Trans. on Signal Processing*, Vol. 44, No. 3, pp. 473-483, March 1996.

[13] Serra, X. and J. O. Smith. 1990. "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition.", *Computer Music Journal* 14(4): pp. 14-24.

[14] Serra, X. "Musical sound modeling with sinusoids plus noise", in *Musical Signal Processing*, A. Piccialli, G. De Poli, C. Roads and S. T. Pope. Swets & Zeitlinger, Amsterdam, Holland.

[15] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice Hall, Upper Saddle River, New Jersey, 1997.