AN ITERATED RATIONAL FILTER BANK FOR AUDIO CODING

Thierry BLU

France Télécom–CNET PAB/STC/SGV
38-40 rue du Général Leclerc
92131 Issy-les-Moulineaux, France

e-mail : blu@issy.cnet.fr

ABSTRACT

This paper proposes a regular third-of-an-octave filter bank for high fidelity audio coding. The originality here is twofold:

- first, the filter bank is an iterated orthonormal rational filter banks for which the generating filters have been designed so that its outputs closely approximate a wavelet transform. This is different from the known coding algorithms which all use an integer filter bank, and most often a uniform one
- second, the masking procedure itself is modelized with the help of a wavelet transform unlike the classical procedure in which a short time spectrum is computed and which gives rise to unwanted preecho effects. The masking procedure is then made equivalent to a quantization procedure.

A simple non-optimized algorithm has been worked out in order to show the benefits of such a structure, especially in terms of preecho (which is perceptually inaudible), and the disadvantages, especially as far as delay is concerned.

1. INTRODUCTION

Among the various algorithms which aim at compressing high fidelity audio sounds, the most efficient [7,4,5,11] make use of one of the psychoacoustic characteristics of the ear: the frequency masking. One of them has been adopted in the normalization as part of the MPEG recommendations [10].

However, all these algorithms make use of a masking procedure which is perfectly adapted to stationary signals, but induce unwanted preecho effects when applied to highly non-stationary inputs, such as percussive sounds. Indeed, these algorithms are mainly differenciated from each other by the kind of analysis-synthesis filter bank involved, whereas the masking procedure is always based on a short time fourier spectrum, whose window can be adapted to follow the non-stationarities.

Yet, the analysis of the early stage of sound processing in the internal ear suggests that a wavelet transform is more closely adapted to human perception than a uniform transform, at least for frequencies higher than 500 Hz. The goal with such a transform would thus be to decompose an audio sound into critical bands, which imposes a scale factor of roughly 6/5 in the discrete wavelet transform.

This paper shows the first example of real implementation of a wavelet transform with a fractional scale factor in a coding algorithm. With this goal in sight, it has been necessary to design [2] a two-band perfect reconstruction rational filter bank [1] whose iterations lead to a selective third-of-octave filter bank. This application is part of a larger more theoretical work on iterated rational filter banks, whose purpose was to extend the well-known results from the dyadic case (octave iterated filter banks which implement a discrete wavelet transform for scale factor 2 [8,9]) to the case of fractional samplers. The results have taken the form of a PhD thesis [3], held recently.

2. ITERATED RATIONAL FILTER BANKS

A filter bank with fractional rate changes is constituted of branches of the form given in figure 1

Figure 1: rational branch

where p and q are integers. Such an operator ensures a rate change of q/p, and for $q > p$, G being lowpass, its iterations generate limit (under conditions explored in [1]) functions φ_n which are not integer translates one from another [1], unlike the classical "dyadic" situation for which $q = 2$ and $p = 1$. The idea of iterating a rational branch was originally emitted in [6] but these authors then concluded that the iterations did not converge for FIR filters, since they were expecting only one single shifted function.

As a consequence of the existence of limit functions, an iterated analysis rational filter bank such as the one depicted in figure 2 (note that here G is supposed to be lowpass and H high-pass) could be re-interpreted as a discrete time-scale transform [1]. Due to the lack of shift-invariance of the limit functions, this transform can never be exactly a wavelet transform —at least for FIR generating filters—. However a result of [1] much more developed in [3] indicated that the error between the two transforms can be made very small for an adequate choice of the generating filter.

It is shown in [3] how to effectively compute the shift error associated to the limit functions: this error is closely

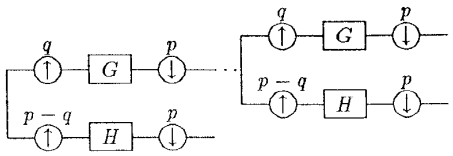

Figure 2: iterated analysis rational filter bank

linked to the selectivity of the iterated filter bank, and also, to some extent, to the selectivity of the generating filter itself.

Finally, an efficient design algorithm for FIR orthogonal filter banks has been described in [2,3] which opens the possibility to implement a transform with reasonable characteristics. From now to the end of this paper, it will be assumed $p = 6$, $q = 5$ (which corresponds to a critical band analysis, slightly finer than the third of octave analysis) and that the iterations generate a discrete wavelet transform up to an (square norm) error smaller than 48 dB.

3. MASKING EFFECT

In this section we will show that, for stationary input signals, the usual computation of the masking effect can be made equal to the computation of the power of a wavelet transform.

3.1. Masking formulation

This psychoacoustic effect is dealt with details in [12]. The initial experiment questioned the audibility of a pure tone in the presence of a narrow band noise: a threshold curve could thus be drawn, giving the intensity under which a pure tone is not heared, is masked.

We shall make some assumptions which are frequently admitted in the perceptual coders

- the masking curve is proportional to the intensity of the masking noise
- the masking curve has a constant shape when frequency is given in Bark scale (not in Herz), that is to say if $B(f_0, f)$ is the masking curve induced by a narrow band noise at frequency f_0, then

$$B(f_0, f) = \beta(u(f) - u(f_0)) \tag{1}$$

  where $u(f)$ is the relation between Bark scale and Herz scale
- the masking curve of a complex sound is a weighted (with the specific intensities) sum of the elementary curves

and we shall add another one which can be verified to be roughly correct for frequencies higher than $f_c = 500$ Hz [12]: the transfer equation between Herz (variable f) and Bark (variable u) scales is logarithmic (up to an additive constant)

$$u = \log_{\frac{6}{5}}(f/f_c) \tag{2}$$

Thus, the masking curve $M_x(f)$ of a stationary signal $x(t)$ is given by the following equation

$$M_x(f) = \theta \int D(f_0) B(f_0, f) df_0 \tag{3}$$

where $D(f_0)$ is the power spectral density of x at frequency f_0 and θ a constant, usually taken as 10^{-3}, in order to account for the masking of a pure tone by another pure tone (and not a narrow band noise).

3.2. Mask wavelet

We shall now show that (3) takes the form of the average power of a wavelet transform for a particular wavelet μ. Let

$$m(u,t) = \int x(\tau) \mu \left(\left(\tfrac{6}{5}\right)^u (\tau - t) \right) d\tau$$

An easy computation which makes use of (1) and (2) leads to

$$M_x(f) = \alpha \left(\frac{6}{5}\right)^{2u} < m(u(f), t)^2 > \tag{4}$$

where α is a constant, and provided we choose the wavelet so that its fourier transform $\hat{\mu}$ obeys

$$\left| \hat{\mu} \left(\left(\tfrac{5}{6}\right)^u \right) \right|^2 = \beta(u) \tag{5}$$

Of course, this formulation is true only for frequencies higher than 500 Hz.

It is indeed of great importance to replace the computation of a fourier spectrum by a time-frequency transform, since the latter is naturally extended to non-stationary signals. In that process, the only thing to take care of is the phase of the fourier transform of the wavelet, which is not accounted for in (5).

3.3. Masking procedure

The usual way to use the masking curve in coding algorithms is to consider that this curve gives a noise level for every frequency and that two signals whose difference is less than that noise is not audible: the masking curve is used so that it provides a linear quantization step. It is thus necessary to have an estimation of the spectrum of the signal. The wavelet formulation of the masking curve and some other psychoacoustic reasons (critical band, i.e third-of-octave analysis of the auditory system) suggests that we decompose the signal in critical bands

$$y(u,t) = \int x(\tau) \psi \left(\left(\tfrac{6}{5}\right)^u (\tau - t) \right) d\tau$$

where, now, the analysis wavelet ψ has good frequency separation properties, unlike the mask wavelet μ. Consequently, our reformulation of the masking reduces to comparing $< y(u,t)^2 >$ and $< m(u,t)^2 >$.

As far as we know, the auditory system is much more sensitive to pure tones than to noises. A particular treatment will be thus needed, since a third-of-octave analysis is otherwise not precise enough for the detection of pure sounds. If, for stationary signals, we want to be coherent with what is done in more frequency selective coders [7]

it can be shown that the proportionality constant linking $\sqrt{< m(u,t)^2 >}$ and the quantization step for $y(u,t)$ is different for tone or non-tone signals. In the implementation that now follows, this constant has been computed taking into account the values given in [7].

4. CODING ALGORITHM

We have devised an architecture and a simple method for the coding of 32 kHz audio sounds: this is by no means an optimized algorithm —the author did not have the hardware necessary to find the best values for all the parameters— , and is just meant to show the feasability of the approach. Besides, the extension to a sampling rate of 44,1 or 48 kHz is straightforward.

As has been shown above, it will be necessary to implement two discrete wavelet transforms with fractional scale factor, here $\frac{6}{5}$. The choice was to retain the iterated structure described in [1] and studied in details in [3].

4.1. Implementation of the wavelet transforms

Instead of building two independent filter banks, it has been preferred to base the two wavelet transforms on the same iterated low-pass branch as shown in figure 3. This has a slight implication on the masking effect as it is implemented: the masking of low-frequencies by the high frequencies is underestimated.

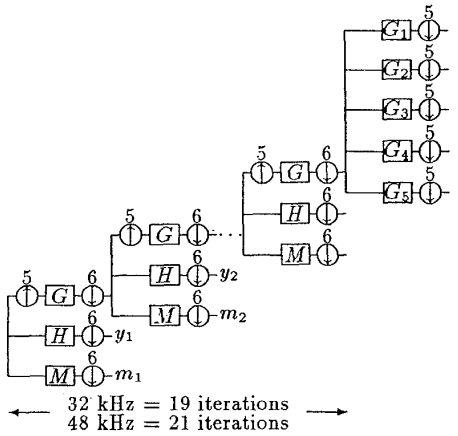

Figure 3: Iterated structure used for audio coding

The orthonormal filter design algorithm used has been published in [2] and provided a perfect reconstruction pair G,H of respective length 204 and 44: their frequency response is given in figure 4. The third filter which leads to the mask wavelet is not constrained by perfect reconstruction, and was thus easily designed (filter length = 40) so that its logarithmic slope toward low frequencies (corresponding to the masking of high frequencies by low frequencies) constant at 10 dB per Bark. In order to determine the phase of $\hat{\mu}$, M has been "synchronized" with H , that is to say the highest sample of the impulse response of H coincides with the one of M.

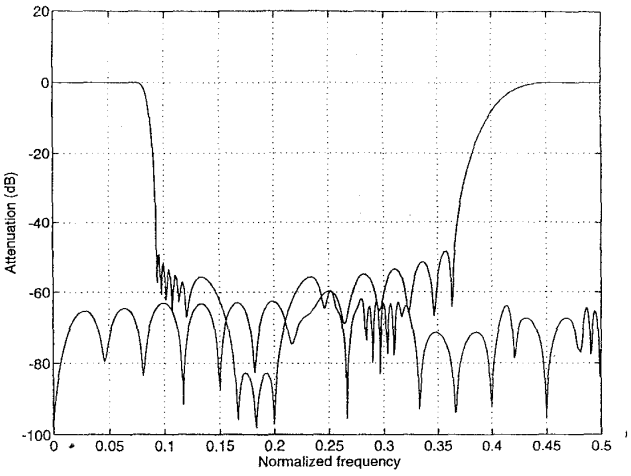

Figure 4: Low and high-pass filters G,H

4.2. Quantization

The samples of the masked (m_j) and non-masked (y_j) outputs at iteration j are put into frames of length 20. This size has been chosen in order to ensure a preecho period less than 4 milliseconds for the first critical band (iteration $j = 1$). This value is consistent with what is otherwise known about the temporal masking in auditory experiments.

The average power of m_j is computed by $< m_j[n]^2 >= \frac{1}{20}\sum_{k=-9}^{10} m_j[20n+k]^2$. Through the multiplication of a "mask constant" $\theta_j[n]$, this power gives a quantization step for the series of samples $(y_j[20n+k])_{k=-9..10}$. The whole trick is thus concentrated in the computation of the mask constant. From the value given in [7], it can be shown [3] (for stationary signals) that if y_j is tonal then $\theta_j[n] = 10^{-3}$ and on the contrary, if the spectrum of y_j is flat, then $\theta_j[n] = \frac{5^{j-1}}{6^j}$.

In order to determine the presence of tonality, we have simply used a linear prediction model which can be sketched as follows: find the best parameter $a_j[n]$, such that the average power of the signal y_j filtered by $P(z) = z^{-1} - 2a_j[n] + z$ is minimized over the 20 samples corresponding to frame n. If this residual power is less than, say half the power of y_j then the band is tonal, otherwise it is not. This is indeed a very rustic procedure for which much optimization can be done: as it was claimed sooner, the algorithm presented here is just meant to show the feasability of the wavelet (and iterated rational filter bank) approach.

All this being done, we end up with three kinds of data to encode: the quantized values of y_j at the nominal rate of each critical band (i.e $\frac{5^{j-1}}{6^j}$ times the sampling frequency); the masked powers $< m_j[n]^2 >$ that we choose to quantize non linearly (according to data given in [12]) over 8 bits, and transmitted at the nominal rate of the critical band, divided by 20; the tone parameters $a_j[n]$ when necessary, quantized linearly over 8 bits, and transmitted up to the twentieth of the nominal critical band rate.

4.3. Absolute hearing threshold

Especially for high frequencies, it is useful to take the absolute hearing threshold into account for the computation of the quantization step. When the masking value is smaller than this threshold, the latter is then preferred. One must however take into consideration the fact that this procedure makes the whole process nonlinear, unlike the masking procedure alone. This might have unwanted consequences when one wants to modify the listening sound level...

4.4. Coding

This part has not really been implemented: the only thing that is done is to compute the statistical entropy of each data to transmit (y_j,$< m_j[n]^2 >$ and $a_j[n]$) for every coded sound, in order to get an idea of the necessary transmit rate. Apart from the three kinds of data to transmit, we have felt the need to transmit two more bits every 20 samples frame in order to indicate the presence (or absence) of "tonality", and to indicate if the quantized values in the frame are all zeros (in which case, of course, these values are not transmitted and the frame reduces to 2 bits), which is a very elementary run length coding method.

It must be added that the final uniform transform shown in figure 3 is not really implemented and that we have only estimated its rate: it is however very low-pass ([0 – 500 Hz]) as compared to the whole bandwidth considered (here [0 – 16 kHz]) and that its contribution to the final rate must be, either 16 kbps if we decide to quantize its samples over 16 bits, or 8 kbps if we choose a selective enough 5-band transform for which a nonlinear 8 bits per pixel quantization is achievable. Of course, no further entropy reduction is taken into account in this basic estimation.

4.5. Synthesis

To reconstruct the signal from the coded samples, an iterated rational synthesis filter bank is applied to the inverse quantized samples. This is exactly a mirror structure of the one shown in figure 2 for which the low and high-pass filters are, respectively, $G(z^{-1})$ and $H(z^{-1})$ (orthonormal filters).

4.6. Characteristics

Due to the length of the low-pass filter, the delay is very important: 200 ms. However, this structure does not imply any audible preecho, unlike classical algorithms. This is not surprising for high frequencies since in that case the computed value is less than 4 ms; this is however more unexpected for low frequencies where the computed value is close to 100 ms: such a result must be understood as a consequence of the wavelet-like analysis of the sound by the internal ear.

5. CONCLUSION

We have presented both a new frequency masking model and structure to implement it, in order to compress high fidelity sounds. This dynamical model has proven efficient as far as preecho is concerned: as a matter of fact, preecho is not audible. Its worst feature is, for the time being, a high delay: this is because we made the choice of very selective filters, but this constraint might be to strong. We did not give rate values, mainly because of an initially bad absolute hearing threshold, lately modified, which has prevented the coding algorithm to be transparent for all tested sounds: the obtained rates where then between 70 and 110 kbps for very good quality sounds.

We think indeed that this algorithm is just an example of what can be done with a wavelet transform for both the analysis-synthesis part and the masking part, and we wanted to stress its advantages. The drawbacks can with no doubt be minimized with further work.

6. REFERENCES

[1] T. Blu , "Iterated Filter Banks with Rational Rate Changes – Connection with Discrete Wavelet Transforms", *IEEE Trans. SP Special Issue on Wavelets*, Vol. 41, No. 12, pp. 3232-3244, Dec. 1993

[2] T. Blu , "Lossless Filter Design in Two-Band Rational Filter Banks: A New Algorithm", *Proceedings of the GRETSI*, Vol. 1, pp. 69-72, Juan-les-Pins, France, 1993

[3] T. Blu, "Bancs de filtres itérés en fraction d'octave — Application au codage de son" (in french), *PhD thesis*, ENST Paris, 1996

[4] Y.F. Dehery, M. Lever and J.B. Rault, "Une norme de codage sonore de haute qualité pour la diffusion, les télécommunications et les systèmes mulimédias" (in french), *L'écho des recherches*, No. 151, pp. 17-28, 1st trimester 1993

[5] J. D. Johnston and K. Brandenburg, "Wideband Coding – Perceptual Considerations for Speech and Music", in *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi ed., New York 1991

[6] J.Kovačević and M. Vetterli , "Perfect Reconstruction Filter Banks with Rational Sampling Rate Changes", in *Proc. ICASSP* May 1991, Vol. 3, pp. 1785-1788, Toronto, Canada

[7] Y. Mahieux, "High Quality Audio Transform Coding at 64 kbit/s", *Ann. Téléc.*, Vol. 47, No. 3-4, pp. 95-106, March-April 1992

[8] S. Mallat , "A Theory for Multiresolution Signal Decomposition: The Wavelet Decomposition", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 11, No. 7, pp. 674-693, July 1989

[9] Y. Meyer , "Ondelettes" (in french), Hermann ed., Paris 1990

[10] "Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5 MBit/s", *International norm MPEG1*, ISO No. 11172,1992

[11] D. Sinha et A.H. Tewfik, "Low Bit Rate Transparent Audio Compression Using Adapted Wavelets", *IEEE Trans. SP*, Vol. 41, No. 12, pp. 3463-3479, Dec. 1993

[12] E. Zwicker and R. Feldtkeller, "Psychoacoustique" (in french), *CTST*, Masson 1981