NEW SOUND DECOMPOSITION METHOD APPLIED TO GRANULAR SYNTHESIS

Charles Bascou and Laurent Pottier

GMEM
Centre National de Creation Musicale
15, rue de Cassis
13008 MARSEILLE FRANCE
www.gmem.org
charles.bascou@free.fr dvlpt@gmem.org

ABSTRACT

In the field of granular decomposition of sound, the Matching Pursuit algorithm is particularly well suited in representing signals with simple sonic entities localized in time and frequency. Our main goal here is to extend this method towards a sound decomposition on a set of arbitrary microsounds leading to a more adaptive framework.

1. INTRODUCTION

In 1946, Dennis Gabor introduced the idea that any sound could be decomposed into a set of simple acoustical events called quantum [1]. This granular model of sound was perceptually validated according to the limitations of the auditory system dealing with time-frequency discrimination. Each of these acoustical events corresponds to a local Time-Frequency component of the sound and thus can represent a large variety of signal structures from transients to pitched sustained parts.

This theory of signal and perception introduced a new approach to synthesize sounds called Granular Synthesis. The main principle of this technique is the accumulation of a large amount of basic parametric sonic events called grains. Iannis Xenakis was one of the first music composers who used the grain as the basic symbolic component of some of his pieces (mostly instrumental) and thus broke the wall between micro and macro musical structure. Since the 1970s, Curtis Roads has explored many aspects and applications of this synthesis technique from real-time pitch shifting of sounds to complex textures generation [2]. He particularly studied the perception effects of the different synthesis parameters and proposed an exhaustive categorization of the diverse applications according to the constraints/relations applied to these parameters. Thus, lots of high-level control strategies have been introduced but with an empirical background.

From here comes our interest to deduce granular synthesis parameters from previously analyzed sound, that is to say to design a granular analysis/synthesis tool. This idea comes with the application of such techniques to natural noisy sounds in mind. It points to sounds that can be defined as the accumulation of more or less complex sonic grains, with their proper temporal and spectral variability. For example, the sound of rice falling onto a metal plate is composed of thousands of elementary “ticks”; the rain produces, in the same way, the accumulation of a large amount of water droplet microsounds...

In fact, in the real world, when multiple realizations of a same event, of a same phenomenon occur, we can expect these types of sounds. Our goal [3] is thus to analyze natural sounds in order to extract the temporal and spectral distribution of those grain streams, and to model these evolutions/fluctuations by correlated statistical laws. We expect, *in fine*, the possible synthesis of sounds perceptually ascribable to the class of the analyzed sound. This might lead us to the classification of granular sounds according to their characteristics. Thus we have to detect the grain parameters of the analyzed sound and this paper will focus on this task. The first part presents briefly the Matching Pursuit algorithm and its limitations for our application. Then we propose an extension of this algorithm to the spectral domain that we have called Spectral Matching Pursuit.

2. THE MATCHING PURSUIT DECOMPOSITION

Since the 1990s, new signal analysis techniques have found to be directly related to the granular representation of sounds. Unlike the classical Fourier Transform massively used for audio analysis purposes, the Matching Pursuit (MP) algorithm proposed by Mallat and Zhang [4] in 1993 is particularly well suited to decompose sounds into elementary sonic entities. In fact, the backgrounds of Fourier Transform and MP are not so different: these methods both project the analyzed signal on a set of elementary waveforms. The main difference is that the set of functions used in MP is redundant and, unlike in Fourier Transform, doesn’t constitute a basis in the signal space. This feature allows the representation of a large variety of signal struc-

tures in a compact manner, i.e. with the smallest set of non-zero coefficients. Indeed, in the Fourier case, a signal well localized in frequency but not in time will have a compact representation. Whereas, a signal well localized in time, like an impulse, will have a non-compact representation, the method characterizing this simple signal structure in a less efficient manner.

The idea of MP is to use an overcomplete dictionary of functions to represent in a compact way a wide range of signal time-frequency behaviors. The traditionally used dictionary is a set of symmetric Gabor atoms indexed by their frequency and duration [4]. Some work has introduced the use of asymmetric functions such as damped sinusoids [5] resolving pre-echo artifacts on transients.

The decomposition process is a greedy algorithm so the quality of the representation, in terms of approximation error, is directly related to the number of iterations of the algorithm. At each step i, it chooses the atom $g_{m(i)}$ in the dictionary D that best approximates the residue $r_i[n]$ of the analyzed signal $x[n]$. Then, this atom contribution is subtracted from the residue and the same scheme is applied again to the new residue $r_{i+1}[n]$. The algorithm stops either when a specific number of atoms has been detected or when the norm of the residue is below a user defined threshold. The algorithm starts with $r_1[n] = x[n]$ and at each iteration i it computes :

$$r_{i+1} = r_i - \langle g_{m(i)}, r_i \rangle g_{m(i)}[n] \tag{1}$$

The Matching Pursuit, in its basic form, uses the inner product for the correlation calculation between the analyzed signal and the dictionary atoms. More complex distance functions have been introduced in the HRMP algorithm to reduce the pre-echo artifact avoiding creating energy where there was none [6].

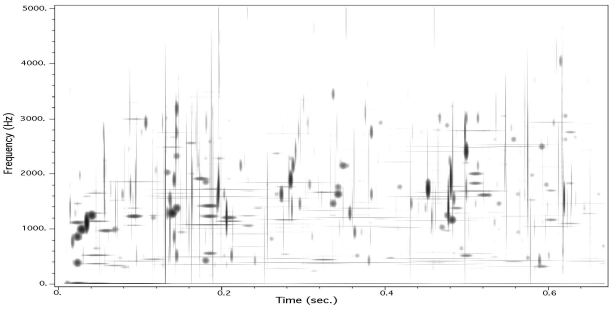

Figure 1. The representation of a water sound Matching Pursuit decomposition.

Thanks to the *LastWave* software [7], We have successfully tested this method on a variety of complex sounds. It has been found to be particularly well suited for sound *a priori* composed by simple sinusoidal grains, like water sounds. An example of decomposition of this type of sound is represented in figure 1. This figure shows grains according to their Wigner Time-Frequency distribution.

In theory, we can represent any type of granular sound with this method but we have to pay particular attention to keep the inherent granular structure of the natural analyzed signal. Indeed, lots of granular sound classes in the real world are composed of grains with a complex frequency structure. We can cite "scratching" or "cracking" sounds made by the accumulation of thousands of complex microsounds not necessarily deterministic. Those can be decomposed by the algorithm but in breaking the complex inherent grain into a large set of simple atoms. This decomposition behavior is not desired here. The grains have to keep a physical sense and to be directly related with the analyzed signal. We extend the Gabor's sonic quanta theory to an arbitrary set of complex grains. Thus, it leads us to a granular decomposition method onto a set of arbitrary microsounds. The temporal domain signal model, used in MP, shows its limitations in this purpose especially with grains containing a significant stochastic part. From here comes the idea to work in the frequency domain for the decomposition process. For that purpose, we propose in next section to extend the MP algorithm to the spectral domain.

3. SPECTRAL MATCHING PURSUIT

3.1. Principle

The principle of adaptive granular decomposition of the MP is retained. The main idea is, rather than working with the temporal signal, to use its spectrogram for signal/atoms distance calculation. Spectrograms provide an accurate representation of various signal behaviors. With spectral smoothing techniques, we can model non deterministic signals giving an approximation of their spectral power density. This method can be compared to dictionary based spectral form recognition. Moreover, the use of dictionaries offers a great advantage especially because of their adaptivity to a given analyzed signal. Indeed, it gives the user the ability to configure the set of atoms he wants the signal to be represented with. For our application to natural noisy sounds, the dictionary construction is a non-trivial problem. The parameters chosen to characterize atoms must be compatible with the granular synthesis engine. As we will see later, we propose to generate the dictionary by transformations of one or multiple "characteristic atoms" directly picked up from the analyzed signal.

3.2. Algorithm

Let x be the analyzed signal and g_k the dictionary atoms. The spectrogram $|X(t,f)|$ of the signal x, with the reduced frequency $f = f_{hz}/F_s$, is defined as :

$$|X(t,f)| = |\sum_{n=1}^{N-1} w[n]x[t+n]e^{2j\pi fn}| \tag{2}$$

We start from the assumption that $|X(t,f)|$ can be decomposed as a weighted sum of $|G_k(t,f)|$:

$$|X(t,f)| = \sum_{i=1}^{M} \alpha_i |G_{m(i)}(t,f)| \tag{3}$$

We now work with matrices and no longer with temporal vectors. To simplify notations and the method structure, we introduce a vectorial form of the spectrograms $X[p_{t,f}]$ defined as :

$$X[p_{t,f}] = |\sum_{n=0}^{N-1} w[n]x[t \times d + n]e^{2j\pi fn}| \quad (4)$$

$$\text{with the index} \quad p_{t,f} = tN + fN \quad (5)$$

where N size of the STFT and d the step size in order to tune their overlap.

At each step i, the algorithm must choose in the dictionary D the atom $G_{m(i)}$ which minimizes the two-norm of the residue R_{i+1} defined by :

$$R_{i+1} = R_i - \alpha_i G_{m(i)} \quad (6)$$

with $m(i)$ the index of the atom chosen at step i. The algorithm starts with $R_1 = X$. We have to find $G_{m(i)}$ as :

$$G_{m(i)} = \arg\min_{G_{m(i)} \in D} \|R_{i+1}\|^2 \quad (7)$$

Moreover the orthogonality principle gives :

$$\langle R_{i+1}, G_{m(i)} \rangle = \langle R_i, G_{m(i)} \rangle - \alpha_i \langle G_{m(i)}, G_{m(i)} \rangle = 0 \quad (8)$$

Hence :

$$\alpha_i = \frac{\langle R_i, G_{m(i)} \rangle}{\langle G_{m(i)}, G_{m(i)} \rangle} = \langle R_i, G_{m(i)} \rangle \quad (9)$$

implying the G_k to be unit-norm. We can deduce the two-norm of the residual signal $\|R_{i+1}\|$:

$$\|R_{i+1}\|^2 = \|R_i\|^2 - |\alpha_i|^2 \quad (10)$$

Minimizing $\|R_{i+1}\|^2$ implies maximizing $|\alpha_i|^2$. We thus have to choose at each step i the atom which has the greater correlation coefficient $|\alpha_i| = |\langle R_i, G_{m(i)} \rangle|$.

3.3. Dictionary construction

The main point here is to keep the dictionary structure related to granular synthesis parameters. The analysis process gives us results as a list of amplitudes, times and indexes of the chosen atoms. In an underlying way, this index thus point to a set of parameter used to create these atoms. The synthesis engine developed at GMEM in the Max/MSP environment currently support sinusoidal grains and buffer based grains (*i.e.* recorded sound granulation) with the control of their frequency/transposition. Our first experiment to construct the dictionary is thus to pick one or more grains which seem to be the more characteristic in the original signal. These grains are then directly used in the synthesis engine buffers. According to the synthesis features available, we experimented generating the dictionary atoms by transposing the "characteristic grains" on a given scale, thus modifying their pitch, length and spectral envelope. More transformations can be considered but with paying a particular attention to the synthesis capabilities.

4. DECOMPOSITION EXAMPLES

Here is a concrete example of gravel sound decomposition with such techniques. We have especially focused on the hard task of dictionary construction, which greatly influences the analysis results. The spectrogram of this sound is represented in figure 2. It is a quite dense granular sound composed of complex grains, made by rocks knocking against one another.

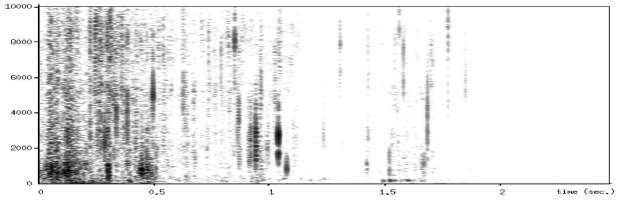

Figure 2. Gravel sound spectrogram from 0 to 10000Hz.

We will give 3 examples of decomposition according to different dictionary construction processes. This step of the method is critical in terms of analysis result reliability and usability. As we saw previously, the main idea in order to adapt the analysis method to the synthesis engine is to take one or more grains in the analyzed signal and to transform them. Here, only transposition in temporal domain has been used.

A very simple first example is to put just one grain, previously picked up from the sound, in the dictionary. The spectrogram of the sound and the atom has been calculated using 4096 as FFT size and 128 as hop size. We have directly specified 1500 as the number of atoms we wanted to represent the signal (corresponding to 1500 iterations of the algorithm). The algorithm gives us times of occurrence and amplitude of the detected atoms. The spectrogram of the reconstructed signal is represented in figure 3.

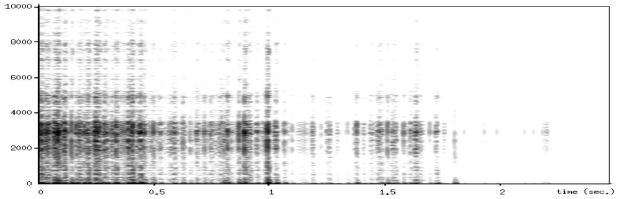

Figure 3. Spectrogram from 0 to 10000 Hz of the reconstructed sound with a single atom dictionary.

As expected, the resulting sound is very monotone. Indeed we ignored the spectral variability of the grains in the analyzed signal. However, it provides a good representation of its temporal structure.

To improve the representation of the spectral variability, we constructed, as a second example, a dictionary by transposing the previously chosen "characteristic grain" on a given scale. It has been done by resampling the initial grain, in the temporal domain, by a factor going from 0.5 to 1.5 with a 0.1 step. This scale has been chosen according to an estimation of the spectral variability of the

analyzed signal. The dictionary now contains 10 grains. The number of iteration steps is still set to 1500.

The reconstructed sound is more interesting than the first example. Indeed, it covers a wider range of the spectral space. Although its spectral content is still quite far from the original one. Moreover, the unique use of transposition to generate grains can give the reconstructed sound an undesired synthetic color. The result is obviously highly correlated with the "characteristic grain" choice we made. The choice of an optimal grain would have certainly led us to a more reliable result. A solution to this problem is the use of multiple "characteristic grains" to construct dictionaries.

The third example shows a decomposition on a dictionary constructed with multiple sound grains. Every one of the 4 "characteristic grains" has been chosen by hand to be the most distant, "spectrally" speaking, from each other, in order to better represent the grain variability of the analyzed sound. The dictionary has been generated by transposing each grain by a factor going from 0.8 to 1.2 with a 0.1 step.

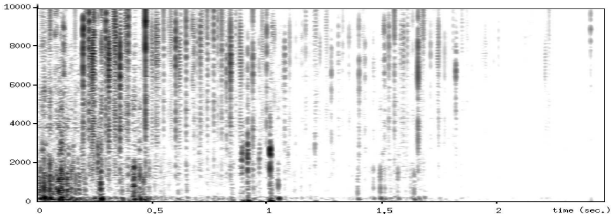

Figure 4. Spectrogram from 0 to 10000 Hz of the reconstructed sound with the 4 transposed grains dictionary.

We can see in figure 4 that this decomposition handles more various frequency structures according to those seen in the original signal. This remark is reinforced when we look at the energy decay of the residual spectrum of the 3 decompositions represented in figure 5. We can see that the use of multiple characteristic atoms, corresponding to the thick black line, gives a much more reliable approximation of the sound. Indeed, in the first two examples, there are parts of the signal the grains can't cover, the residual energy remaining quite high after the 1500 iterations.

The last decomposition result sounds much more natural and, although no formal test has been made, it can be perceptually recognized as a gravel sound. Indeed, the transposition artifact we had in the second example is greatly reduced by the spectral diversity brought by the 4 "characteristic grains".

5. IMPROVEMENTS AND FUTURE WORK

This analysis tool is an important part of the "GMU" project at GMEM [3]. Although it is still in its early stages of design and development, it provides promising results in the domain of granular decomposition. The sound synthesis can produce realistic sounds sometimes very close

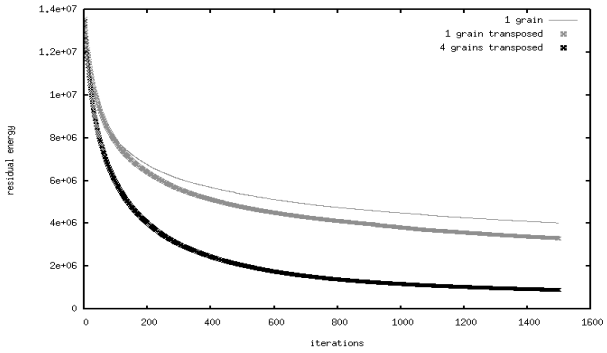

Figure 5. Decay of the residual spectrum energy (first example is thin grey, second is thick grey and last is thick black).

of the original. However, there is still lot of work especially in methods to construct the dictionary. This part is really critical in terms of decomposition reliability. We have especially to evaluate the feasibility of an automatic choice method of "characteristic grains". An other approach would be, when the production of the analyzed signal is reproducible, to record one (or several) grain token as an elementary waveform for dictionary construction. At a higher level, the next steps we have to accomplish are to study the parameter fluctuation laws and the correlation between them for more flexibility in synthesis.

6. REFERENCES

[1] D. Gabor. "Acoustical Quanta and the Theory of Hearing", *Nature*, 159(4044):591-594, 1947.

[2] C. Roads. *Microsound*, Cambridge, Massachusetts: MIT Press, 2002.

[3] L. Pottier. "GMU - An Integrated Microsound Synthesis System" *Proceedings of the Computer Music Modeling and Retrieval Conference*, Montpellier (France), 2003.

[4] S. Mallat and Z. Zhang. "Matching pursuits with time-frequency dictionaries", *IEEE Transactions on Signal Processing*,vol 41, no 12, p 3397–3415, 2004.

[5] M. Goodwin and M. Vetterli. "Matching Pursuit and Atomic Signal Models Based on Recursive Filter Banks" *IEEE Transactions on Signal Processing*, 1998.

[6] R. Gribonval, X. Rodet, P. Depalle, E. Bacry and S. Mallat. "Sound signal decomposition using a high resolution matching pursuit" *Proceedings of ICMC'96*, Clear Water Bay, Hong-Kong, 1996.

[7] R. Gribonval, E. Bacry. LastWave 2 Software `http://www.cmap.polytechnique.fr/~Bacry/LastWave/`, 2003.