Sound Signals Decomposition Using a High Resolution Matching Pursuit

R. Gribonval* Ph. Depalle† X. Rodet** E. Bacry†† S. Mallat‡‡

IRCAM, Analysis/Synthesis Department
1 place Igor-Stravinsky,
F-75004 PARIS, FRANCE
fax : (33)-1-42772947

CMAP,
Ecole Polytechnique
F-91128 PALAISEAU CEDEX, FRANCE

Abstract

Sound recordings include transients and sustained parts. Their analysis with a basis expansion is not rich enough to represent efficiently all such components. Pursuit algorithms choose the decomposition vectors depending upon the signal properties. The dictionary among which these vectors are selected is much larger than a basis. Matching Pursuit is fast to compute, but can provide coarse representations. Basis Pursuit gives a better representation but is very expensive in terms of calculation time. This paper develops a High Resolution Matching Pursuit : it is a fast, high time-resolution, time-frequency analysis algorithm, that makes it likely to be used for musical applications.

1 Introduction

The complexity of structures encountered in sound signals requires the development of adaptive low-level representations in order to provide meaningful analysis. Usual time-frequency analysis methods, such as Wavelet [KM88] or Short Time Fourier Transform [RS78] [Moo78], perform a decomposition of the signal according to a given fixed basis. Although such a decomposition entirely characterizes the signal, a basis is a minimal set of vectors that is not rich enough to represent efficiently all components. Indeed, some signal structures may be diffused across many basis elements : such an expansion could then become difficult to correlate with perceptual entities.

Actually, sound signals include transients that are well represented by short waveforms, and sustained parts that are more efficiently decomposed over long waveforms with short frequency support.

Lately, new adaptive approaches have been developed in order to choose the decomposition vectors depending upon the signal properties : for example, Coifman and Wickerhauser [CW92] [BCG94] introduced an adaptive "best basis" selection, but such an analysis algorithm cannot distinguish overlapping features, such as a "click" and a sine wave together, as far as it chooses an *orthogonal* basis (among a given family of such bases) to decompose the sound.

Pursuit algorithms, such as Matching Pursuit (MP) [MZ93] or Basis Pursuit (BP) [CD95] were designed to overcome these problems. The decomposition vectors are selected among a redundant family of elementary waveforms, both well-localized in time and frequency. This family of *time-frequency atoms*, which is much larger than a basis, is called a *dictionary*.

MP is fast to compute, but can provide coarse representations. BP gives a better representation but is very expensive in terms of calculation time. The High Resolution Matching Pursuit (HRMP) developed in this paper (see also [GBM+96]) is a fast pursuit algorithm providing high time-resolution time-frequency representations. It is designed to comply with a good time-localization constraint, thus eliminating pre-echo effects that MP introduced. Therefore it is likely to give much better results for musical applications.

2 Matching Pursuits

A *dictionary* is a family of vectors $\mathcal{D} = (g_\gamma)_{\gamma\in\Gamma}$ included in a Hilbert space H, with a unit norm $\|g_\gamma\| = 1$. A *matching pursuit* is an iterative algorithm that decomposes the signal over dictionary vectors as follows.

Let $R^0 f = f$. We suppose that we have computed the n^{th} order *residue* $R^n f$, for $n \geq 0$. We then choose an element $g_{\gamma_n} \in \mathcal{D}$ which "closely"

*IRCAM, gribonva@clipper.ens.fr
†IRCAM, phd@ircam.fr
**IRCAM, rod@ircam.fr
††CMAP, bacry@cmapx.polytechnique.fr
‡‡CMAP, mallat@cmapx.polytechnique.fr

matches the residue $R^n f$, in the sense that

$$|C(R^n f, g_{\gamma_n})| = \sup_{\gamma \in \Gamma} |C(R^n f, g_\gamma)|, \quad (1)$$

where $C(f, g_\gamma)$ is a *correlation function* that measures the similarity between f and g_γ.

The residue $R^n f$ is then sub-decomposed into

$$R^n f = C(R^n f, g_{\gamma_n}) g_{\gamma_n} + R^{n+1} f, \quad (2)$$

which defines the residue at the order n+1. In the Matching Pursuit (MP), initially introduced by Mallat and Zhang [MZ93], the correlation function that is used is the inner product $C(f, g_\gamma) = < f, g_\gamma >$. Other correlation functions can be used : indeed, the object of this paper is to introduce a correlation function that is adapted to the requirements of audio signals representation. This correlation function will be given by Eq. (8) and will lead to our new pursuit algorithm, High Resolution Matching Pursuit (HRMP).

With either of those correlation functions, the energy of the error $\|R^n f\|^2$ is proved to decay to zero. Thus by iterating Eq. (2) we obtain the *atomic decomposition* of the signal

$$f = \sum_{n=0}^{+\infty} C(R^n f, g_{\gamma_n}) g_{\gamma_n}. \quad (3)$$

The structure of MP enables it to be implemented with a fast algorithm.

3 Gabor Dictionary

To analyze time and frequency localization properties of one-dimensional signals, such as speech or music recordings, we use a large dictionary of *time-frequency atoms*.

Let $g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right)$ be a Gaussian function of unit norm. For any scale $s > 0$, modulation frequency ξ and translation u, we denote $\gamma = (s, u, \xi)$ and define

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{i\xi t}. \quad (4)$$

The index γ is an element of the set $\Gamma = R^+ \times R^2$. The function $g_\gamma(t)$ is centered at the abscissa u and its energy is concentrated in a neighborhood of u, whose size is proportional to s. Its Fourier transform is centered at the frequency $\omega = \xi$, and its energy is concentrated in a neighborhood of ξ, whose size is proportional to $1/s$.

Short scale atoms almost correspond to "clicks", whereas large scale atoms are nearly pure sine waves. This dictionary is thus likely to comply with the representation of transients structures as well as of stationary features.

4 Energy Distributions

The time-frequency energy distribution of $f(t)$ is then defined by

$$Ef(t, \omega) = \sum_{n=0}^{+\infty} |C(R^n f, g_{\gamma_n})|^2 W g_{\gamma_n}(t, \omega). \quad (5)$$

where $W g_{\gamma_n}(t, \omega)$ is the Wigner distribution of g_{γ_n}, *i.e.* a two-dimensional Gaussian "blob" in the time-frequency plane. Figure 1, 2-MP and 2-HRMP display such time-frequency energy distributions.

For example, when applied to a medium pitch (*e.g.* G5 sharp) piano sound, a matching pursuit provides a time-frequency representation (Figure 1) that displays simultaneously structures of very different scales. At first, one can see the quasi-harmonic structure of the note. It is displayed by horizontal lines, corresponding to large scale, well-localized in frequency atoms. Then, below 100 Hz, shorter horizontal lines display the partials of the quasi-harmonic resonance of the piano's sounding board, at a fundamental frequency of approximately 20 Hz. Lastly, vertical features, corresponding to fine-scale transitory structures describe both the attack at the beginning of the note, and the fall back of the piano's damper on the string at its end.

5 High Resolution Matching Pursuit

The Matching Pursuit is a greedy algorithm in that it optimizes at each step the amount of the signal energy it grasps. This often leads to a choice of features which globally fits the signal structures but is not best adapted to its local structures.

Indeed, for instance, a signal composed of two bumps modulated by a sinusoidal wave at frequency ξ (Figure 2-a) is first decomposed into a large atom at frequency ξ (middle horizontal line on Figure 2-a-MP) that covers the time support of both bumps. Then, in order to remove the energy created between the two bumps by this first atom, MP chooses two atoms of the same size as the first one, with frequencies $\xi + \Delta\xi$ (upper line) and $\xi - \Delta\xi$ (lower line).

Moreover, we observed that MP does not keep a good localization of attack patterns (Figure 2-b-MP), which leads to a little, but still audible, pre-echo at re-synthesis stage. This is due to the atom selection criterion that allows the creation of energy where there was none previously.

Aiming at avoiding this problem, Donoho and Chen [CD95] introduced the Basis Pursuit, which makes a full optimization, by minimizing $\sum_{\gamma \in \Gamma} |\alpha_\gamma|$ over all possible decompositions

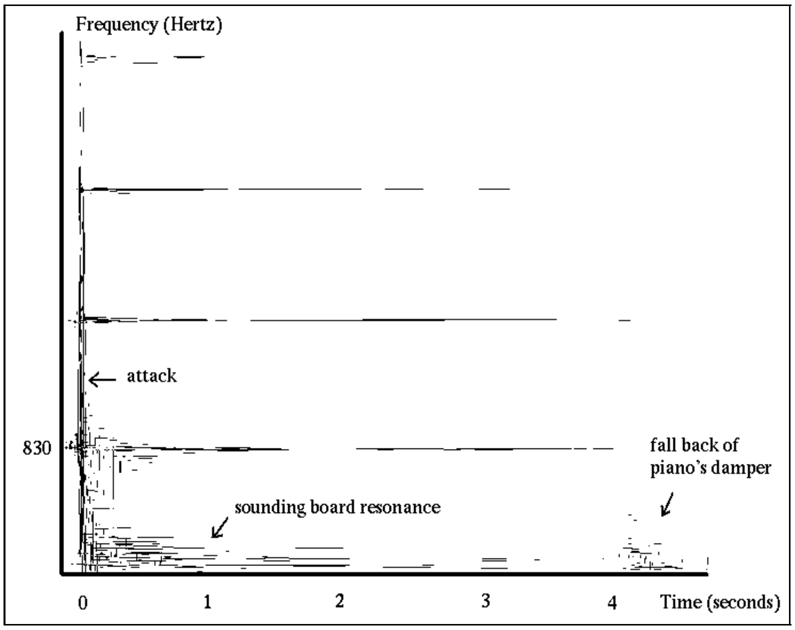

Figure 1: Time-Frequency distribution of a piano note obtained with HRMP

$f = \sum_{\gamma\in\Gamma} \alpha_\gamma g_\gamma$. However this leads to large scale linear-programming problems and therefore is very expensive in terms of calculation time.

The new algorithm, that we called High Resolution Matching Pursuit (HRMP), is an enhanced version of Matching Pursuit (MP), extending to time-frequency dictionaries the pursuit over non-modulated spline dictionaries introduced by Jaggi et. al. [JCMW95]. It uses a different correlation function, that allows the pursuit to emphasize local fit over global fit at each step. The fast algorithm structure of MP is however kept.

For each time-frequency atom g_γ a set I_γ of *sub-atom* indexes is introduced. I_γ corresponds to smaller atoms $g_{\gamma_i}, \gamma_i \in I_\gamma$ with a time support included in the support of g_γ, and modulated at the same frequency.

Let suppose that the atom g_γ is chosen in a pursuit. $Rf = f - C(f, g_\gamma) g_\gamma$ becomes the residue of this pursuit on the signal f. For all $\gamma_i \in I_\gamma$, $< Rf, g_{\gamma_i} >$ represents the amount of "energy" of Rf located on the time-frequency support of g_{γ_i}. This amount must be smaller than the signal "energy" $< f, g_{\gamma_i} >$ at the same location. Moreover the corresponding decrease $< C(f, g_\gamma)g_\gamma, g_{\gamma_i} >$ of signal energy cannot be greater than the initial signal energy itself. This is formalized in Equations (6) and (7) :

$$|< Rf, g_{\gamma_i} >| \leq |< f, g_{\gamma_i} >|, \tag{6}$$

$$|< C(f, g_\gamma)g_\gamma, g_{\gamma_i} >| \leq |< f, g_{\gamma_i} >|. \tag{7}$$

From these relations, we derive the new correlation function $C(f, g_\gamma)$, which maximizes the amount of signal energy that the pursuit can grasp, when choosing the atom g_γ :

$$C(f, g_\gamma) = \varepsilon \min_{\gamma_i \in I_\gamma} \frac{|< f, g_{\gamma_i} >|}{|< g_\gamma, g_{\gamma_i} >|} \tag{8}$$

where ε is evaluated as follows:

- if $< f, g_{\gamma_i} >$ have the same sign, for all $\gamma_i \in I_\gamma$, then ε is this common sign.
- else $\varepsilon = 0$.

In MP, the inner-product, used as a correlation function between a time-frequency atom and an audio signal, disregards whether the signal contains energy on the whole time-frequency support of the chosen atom. On the contrary, the new correlation function avoids creating energy at time locations where there was none. It can thus distinguish close time features as shown in Figure 2-a-HRMP. Moreover it can avoid pre-echo effects, *i.e.* creation of energy just before the beginning of the sound. Indeed, as shown in Figure 2-b, MP introduces a pre-echo effect by choosing atoms that overlap the attack time-location, whereas HRMP does not choose any such atom.

Because of the new correlation function, the atoms chosen for the decomposition have a smaller time support than with a usual Matching Pursuit decomposition, hence, because of Heisenberg inequalities, they also have a larger frequency support. HRMP frequency-resolution is thus decreased but it performs a higher time-resolution decomposition than MP.

However for such audio applications as attack pattern recognition or precise tracking of partials,

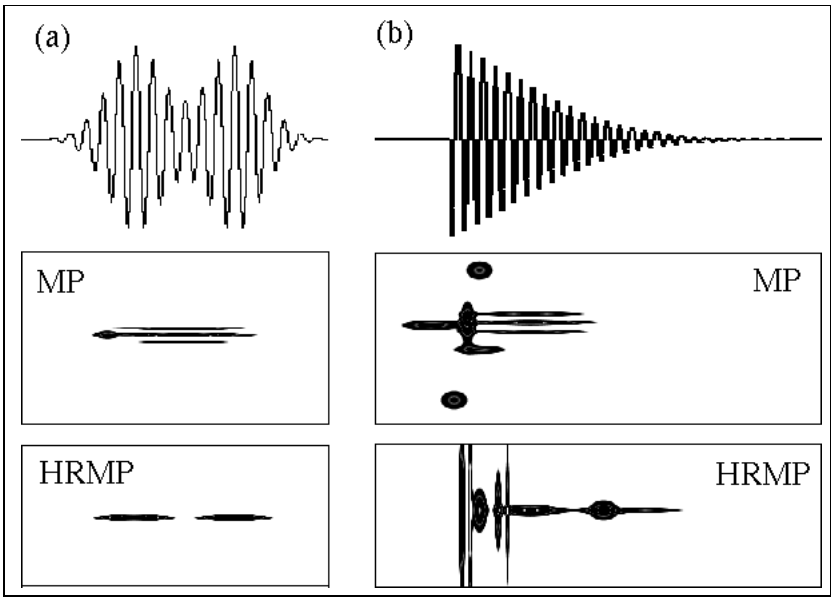

Figure 2: Time-Frequency distributions of signals (top) obtained with MP (middle) and HRMP (bottom): (a) two close bumps, with a four atom decomposition (b) an attack pattern, with a ten atom decomposition

the most important is to keep a good localization of the attacks, because the ear is very sensitive to transients : hearing the attack of a musical instrument is often almost sufficient to identify it.

6 Summary

HRMP provides a time-frequency representation adapted to the specificities of sound signals. Moreover, the signal representation HRMP provides is easily related to perceptual entities (transients, partials, clicks, ...). Thus, HRMP allows more precise or selective sound processing. We have presented, for example, its ability to process separately the sustained and transients parts of a piano sound. We are also considering the ability of the method to extract easily the parameters of formant-waveform synthesizers (central frequency, amplitude, bandwidth and especially excitation duration).

References

[BCG94] J. Berger, R. Coifman, and M.J. Goldberg. A method of denoising and reconstructing audio signals. In *Proc. Int. Computer Music Conf. (ICMC'94)*, pages 344–347, 1994.

[CD95] S. Chen and D.L. Donoho. Atomic decomposition by basis pursuit. Technical report, Statistics Department, Stanford University, 1995.

[CW92] R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, 38(2):713–718, March 1992.

[GBM+96] R. Gribonval, E. Bacry, S. Mallat, Ph. Depalle, and X. Rodet. Analysis of sound signals with high resolution matching pursuit. In *Proc. IEEE Conf. Time-Freq. and Time-Scale Anal. (TFTS'96)*, June 1996.

[JCMW95] S. Jaggi, W.C. Carl, S. Mallat, and A.S. Willsky. High resolution pursuit for feature extraction. Technical report, MIT, November 1995.

[KM88] R. Kronland-Martinet. The wavelet transform for analysis, synthesis, and processing of speech and music sounds. *Comp. Music. Journal*, 12(4), 1988.

[Moo78] J.A. Moorer. The use of the phase vocoder in computer music applications. *Journal of the AES*, (26):42–45, 1978.

[MZ93] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, December 1993.

[RS78] L.R. Rabiner and R.W. Schafer. *Digital Coding of Speech Signals*. Prentice Hall, 1978.