Matching Pursuit With Time-Frequency Dictionaries

Stéphane Mallat [1] *and Zhifeng Zhang*

Technical Report 619
October 1992, revised in June 1993
To appear in IEEE Transactions in Signal Processing, in December 1993

Courant Institute of Mathematical Sciences
Computer Science Department
New York University
251 Mercer Street
New York, NY 10012, USA

Abstract

We introduce an algorithm, called matching pursuit, that decomposes any signal into a linear expansion of waveforms that are selected from a redundant dictionary of functions. These waveforms are chosen in order to best match the signal structures. Matching pursuits are general procedures to compute adaptive signal representations. With a dictionary of Gabor functions, a matching pursuit defines an adaptive time-frequency transform. We derive a signal energy distribution in the time-frequency plane, which does not include interference terms, unlike Wigner and Cohen class distributions [1]. A matching pursuit isolates the signal structures that are coherent with respect to a given dictionary. An application to pattern extraction from noisy signals is described. We compare a matching pursuit decomposition with a signal expansion over an optimized wavepacket orthonormal basis, selected with the algorithm of Coifman and Wickerhauser [4].

[1] This work was supported by the AFOSR grant F49620-93-1-0102, ONR grant N00014-91-J-1967 and the Alfred Sloan Foundation

1 Introduction

We can express a wide range of ideas and at the same time easily communicate subtle differences between close concepts, because natural languages have large vocabularies, that include words with close meanings. For information processing, low level signal representations must also provide explicit informations on very different properties, while giving simple cues to differentiate close patterns. The numerical parameters should offer compact characterizations of the elements we are looking for. The wide scope of patterns embedded in complex signals and the precision of their characterization, also motivate decompositions over large and redundant dictionaries of waveforms. Linear expansions in a single basis, whether it is a Fourier, wavelet, or any other basis, are not flexible enough. A Fourier basis provides a poor representation of functions well localized in time, and wavelet bases are not well adapted to represent functions whose Fourier transforms have a narrow high frequency support. In both cases, it is difficult to detect and identify the signal patterns from their expansion coefficients, because the information is diluted across the whole basis. Similar examples can be found for any type of basis. Such decompositions are similar to a text written with a small vocabulary. Although this vocabulary might be sufficient to express all ideas, it requires to use circumvolutions that replace non available words by full sentences.

Flexible decompositions are particularly important for representing signal components whose localizations in time and frequency vary widely. The signal must be expanded into waveforms whose time-frequency properties are adapted to its local structures. Such waveforms are called time-frequency atoms. For example, impulses need to be decomposed over functions well concentrated in time, while spectral lines are better represented by waveforms which have a narrow frequency support. When the signal includes both of these elements, the time-frequency atoms must be adapted accordingly. One must therefore introduce a procedure that chooses the waveforms that are best adapted to decompose the signal structures, among all the time-frequency atoms of a large dictionary. Section 2 briefly reviews the properties of time-frequency atoms and their relations to window Fourier transforms and wavelet transforms.

We introduce an algorithm called matching pursuit, that decomposes any signal into a linear expansion of waveforms that belong to a redundant dictionary of functions. These waveforms are selected in order to best match the signal structures. Although a matching pursuit is non-linear, like an orthogonal expansion, it maintains an energy conservation which guaranties its convergence. It is closely related to projection pursuit strategies, developed by Friedman and Stuetzle [7] for statistical parameter estimation. The general algorithm in the Hilbert space framework is explained in section 3 and the finite dimensional case is further studied in section 4.

The application of matching pursuits to adaptive time-frequency decompositions is described in section 5. The signal is decomposed into waveforms selected among a dictionary of time-frequency atoms, that are the the dilations, translations and modulations of a single window function. We derive a time-frequency energy distribution, by adding the Wigner distribution of the selected time-frequency atoms. Contrarily to the Wigner distribution or Cohen's class distributions, this energy distribution does not include interference terms and thus provides a clear picture in the time-frequency plane. Qian and Chen [14] have developed independently a similar algorithm to expand signals over time-frequency atoms. A fast implementation of the matching pursuit for dictionary of Gabor time-frequency atoms is described in section 6, with numerical examples.

A matching pursuit decomposition provides an interpretation of the signal structures. If a structure does not correlate well with any particular dictionary element, it is sub-decomposed into several elements and its information is diluted. Section 7 formally defines coherent signal structures with respect to a given dictionary, and explains how to detect them. An application to the extraction of patterns from noisy signals is described.

A matching pursuit is a greedy algorithm that chooses at each iteration a waveform that is best adapted to approximate part of the signal. Section 8 compares this locally adaptive method to the algorithm of Coifman and Wickerhauser [4], which selects the basis that is best adapted to the global signal properties, among all bases of a wavepacket family. Numerical results show that the global optimization does not perform well for highly nonstationary signals, as opposed the greedy approach of a matching pursuit. On the other hand, the best basis algorithm is efficient to represent simpler signals that have stationary properties.

Notations

The space $\mathbf{L}^2(\mathbf{R})$ is the Hilbert space of complex valued functions such that

$$\|f\|^2 = \int_{-\infty}^{+\infty} |f(t)|^2 dt < +\infty. \tag{1}$$

The inner product of $(f, g) \in \mathbf{L}^2(\mathbf{R})^{\mathbf{2}}$ is defined by

$$< f, g >= \int_{-\infty}^{+\infty} f(t)\bar{g}(t)dt, \tag{2}$$

where $\bar{g}(t)$ is the complex conjugate of $g(t)$. The Fourier transform of $f(t) \in \mathbf{L}^2(\mathbf{R})$ is written $\hat{f}(\omega)$ and defined by

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-i\omega t}dt. \tag{3}$$

2 Time-Frequency Atomic Decompositions

Decompositions of signals over family of functions that are well localized both in time and frequency have found many applications in signal processing and harmonic analysis. Such functions are called time-frequency atoms. Depending upon the choice of time-frequency atoms, the decomposition might have very different properties. Window Fourier transforms and wavelet transforms are examples of time-frequency signal decomposition that have been studied thoroughly [2] [5] [13] [15]. To extract informations from complex signals, it is often necessary to adapt the time-frequency decomposition to the particular signal structures. This section discusses the adaptivity requirements.

A general family of time-frequency atoms can be generated by scaling, translating and modulating a single window function $g(t) \in \mathbf{L}^2(\mathbf{R})$. We suppose that $g(t)$ is real, continuously differentiable and $O(\frac{1}{t^2+1})$. We also impose that $\|g\| = 1$, that the integral of $g(t)$ is non-zero and that $g(0) \neq 0$. For any scale $s > 0$, frequency modulation ξ and translation u, we denote $\gamma = (s, u, \xi)$ and define

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

The index γ is an element of the set $\mathbf{\Gamma} = \mathbf{R}^+ \times \mathbf{R}^{\mathbf{2}}$. The factor $\frac{1}{\sqrt{s}}$ normalizes to 1 the norm of $g_\gamma(t)$. If $g(t)$ is even, which is generally the case, $g_\gamma(t)$ is centered at the abscissa u. Its energy is mostly concentrated in a neighborhood of u, whose size is proportional to s. Let $\hat{g}(\omega)$ be the Fourier transform of $g(t)$. Equation (4) yields

$$\hat{g}_\gamma(\omega) = \sqrt{s}\hat{g}(s(\omega - \xi))e^{-i(\omega-\xi)u}. \tag{5}$$

Since $|\hat{g}(\omega)|$ is even, $|\hat{g}_\gamma(\omega)|$ is centered at the frequency $\omega = \xi$. Its energy is concentrated in a neighborhood of ξ, whose size is proportional to $1/s$.

The family $\mathcal{D} = (g_\gamma(t))_{\gamma\in\mathbf{\Gamma}}$ is extremely redundant, and its properties have been studied by Torresani [17]. To represent efficiently any function $f(t)$, we must select an appropriate countable subset of atoms $(g_{\gamma_n}(t))_{n\in\mathbf{N}}$, with $\gamma_n = (s_n, u_n, \xi_n)$, so that $f(t)$ can be written

$$f(t) = \sum_{n=-\infty}^{+\infty} a_n g_{\gamma_n}(t). \tag{6}$$

Depending upon the choice of the atoms $g_{\gamma_n}(t)$, the expansion coefficients a_n give explicit information on certain types of properties of $f(t)$. Window Fourier transforms and wavelet transforms correspond to different families of time-frequency atoms, that are frames or bases of $\mathbf{L}^2(\mathbf{R})$.

In a window Fourier transform, all the atoms g_{γ_n} have a constant scale $s_n = s_0$ and are thus mainly localized over an interval whose size is proportional to s_0. If the main signal structures are localized over a time-scale of the order of s_0, the expansion coefficients a_n give important insights on their localization and frequency content. However, a window Fourier transform is not well adapted to describe structures that are much smaller or much larger than s_0. To analyze components of varying sizes, it is necessary to use time-frequency atoms of different scales.

In opposition to the window Fourier transform, the wavelet transform decomposes signals over time-frequency atoms of varying scales, called wavelets. A wavelet family $(g_{\gamma_n}(t))_{n\in\mathbf{N}}$ is built by relating the frequency parameter ξ_n to the scale s_n with $\xi_n = \frac{\xi_0}{s_n}$, where ξ_0 is a constant. The resulting family is composed of dilations and translations of a single function, multiplied by complex phase parameter. The expansion coefficients a_n of functions over wavelet families characterize the scaling behavior of signal structures. This is important for the analysis of fractals and singular behaviors. However, expansion coefficients in a wavelet frame do not provide precise estimates of the frequency content of waveforms whose Fourier transform is well localized, especially at high frequencies. This is due to the restriction on the frequency parameter ξ_n, that remains inversely proportional to the scale s_n.

For signals $f(t)$ that include scaling and highly oscillatory structures, one can not define a priori the appropriate constraints on the scale and modulation parameters of the time-frequency atoms $g_{\gamma_n}(t)$ used in the expansion (6). We need to select adaptively the elements of the dictionary $\mathcal{D} = (g_\gamma(t))_{\gamma\in\mathbf{\Gamma}}$, depending upon the local properties of $f(t)$.

3 Matching Pursuit in Hilbert Spaces

The general issue behind adaptive time-frequency decompositions is to find procedures to expand functions over a set of waveforms, selected appropriately among a large and redundant dictionary. We describe a general algorithm, called matching pursuit, that performs such an adaptive decomposition.

Let $\mathbf{H}$ be a Hilbert space. We define a dictionary as a family $\mathcal{D} = (g_\gamma)_{\gamma\in\mathbf{\Gamma}}$ of vectors in $\mathbf{H}$, such that $\|g_\gamma\| = 1$. Let $\mathbf{V}$ be the closed linear span of the dictionary vectors. Finite linear expansions of vectors in $\mathcal{D}$ are dense in the space $\mathbf{V}$. We say that the dictionary is complete if and only if $\mathbf{V} = \mathbf{H}$. For the dictionary of time-frequency atoms described in section 2, $\mathbf{H} = \mathbf{L^2}(\mathbf{R})$, and each vector g_γ is an atom defined by equation (4). Finite linear expansions of time-frequency atoms are dense in $\mathbf{L}^2(\mathbf{R})$ [17], hence this dictionary is complete.

Let $f \in \mathbf{H}$. We want to compute a linear expansion of f over a set of vectors selected from $\mathcal{D}$, in order to best match its inner structures. This is done by successive approximations of f with orthogonal projections on elements of $\mathcal{D}$. Let $g_{\gamma_0} \in \mathcal{D}$. The vector f can be decomposed into

$$f =< f, g_{\gamma_0} > g_{\gamma_0} + Rf, \tag{7}$$

where Rf is the residual vector after approximating f in the direction of g_{γ_0}. Clearly g_{γ_0} is orthogonal to Rf, hence

$$\|f\|^2 = |< f, g_{\gamma_0} >|^2 + \|Rf\|^2. \tag{8}$$

To minimize $\|Rf\|$, we must choose $g_{\gamma_0} \in \mathcal{D}$ such that $| < f, g_{\gamma_0} > |$ is maximum. In some cases, it is only possible to find a vector g_{γ_0} that is almost the best in the sense that

$$|< f, g_{\gamma_0} >| \geq \alpha \sup_{\gamma\in\mathbf{\Gamma}} |< f, g_\gamma >|, \tag{9}$$

where α is an optimality factor that satisfies $0 < \alpha \leq 1$.

A matching pursuit is an iterative algorithm that sub-decomposes the residue Rf by projecting it on a vector of $\mathcal{D}$ that matches Rf almost at best, as it was done for f. This procedure is repeated each time on the following residue that is obtained. Before giving further details, let us emphasize that the "choice" of a vector g_{γ_0} that satisfies (9) is not random. It is defined by a choice function C, that associates to any sub-set Λ of $\mathbf{\Gamma}$ an index that belongs to Λ. Let us define the set of vector indexes that satisfy (9)

$$\Lambda_0 = \{\beta \in \mathbf{\Gamma} : \ |< f, g_\beta >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< f, g_\gamma >|\}. \tag{10}$$

The choice of a vector g_{γ_0} that satisfies (9) is equivalent to the choice of the index γ_0 within Λ_0, formally defined by $\gamma_0 = C(\Lambda_0)$. The axiom of choice guaranties that there exists at least one choice function, but in practice there are many ways to define it, and it depends upon the numerical implementation.

Let us explain by induction, how the matching pursuit is carried further. Let $R^0 f = f$. We suppose that we have computed the n^{th} order residue $R^n f$, for $n \geq 0$. We choose, with the choice function C, an element $g_{\gamma_n} \in \mathcal{D}$ which closely matches the residue $R^n f$

$$|< R^n f, g_{\gamma_n} >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< R^n f, g_\gamma >|. \tag{11}$$

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

$$R^n f =< R^n f, g_{\gamma_n} > g_{\gamma_n} + R^{n+1} f, \tag{12}$$

which defines the residue at the order n+1. Since $R^{n+1} f$ is orthogonal to g_{γ_n}

$$\|R^n f\|^2 = |< R^n f, g_{\gamma_n} >|^2 + \|R^{n+1} f\|^2. \tag{13}$$

Let us carry this decomposition up to the order m. We decompose f into the concatenated sum

$$f = \sum_{n=0}^{m-1} \left(R^n f - R^{n+1} f \right) + R^m f. \tag{14}$$

Equation (12) yields

$$f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^m f. \tag{15}$$

Similarly, $\|f\|^2$ is decomposed in a concatenated sum

$$\|f\|^2 = \sum_{n=0}^{m-1} \left(\|R^n f\|^2 - \|R^{n+1} f\|^2 \right) + \|R^m f\|^2. \tag{16}$$

Equation (13) yields an energy conservation equation

$$\|f\|^2 = \sum_{n=0}^{m-1} |< R^n f, g_{\gamma_n} >|^2 + \|R^m f\|^2. \tag{17}$$

The original vector f is decomposed into a sum of dictionary elements, that are chosen to best match its residues. Although this decomposition is non-linear, we maintain an energy conservation as if it was a linear orthogonal decomposition. A major issue is to understand the behavior of the residue $R^m f$ when m increases. Let us mention that the algorithm can be modified by selecting several vectors from the dictionary at each iteration and projecting the residue over the space generated by these vectors [12], but we shall not further develop this approach here.

Functional approximations through such iterated orthogonal projections has previously been studied in statistics by Friedman and Stuetzle [7], under the name of projection pursuit regressions. Our algorithm was developed independently in a very different context, but the underlying mathematics are similar, so we adopted the same vocabulary. The statistical problem is to estimate the conditional expectation of a random variable Y with respect to d random variables $X_1, X_2, ..., X_d$. To reduce the dimensionality of the problem, a projection pursuit regression decomposes the conditional expectation as a sum of conditional expectations of successive residues of Y, with respect to one-dimensional random variables that are linear expansions of $X_1, X_2, ..., X_d$. This decomposition is obtained with a strategy similar to the matching pursuit approach. Readers further interested by projections pursuits are referred to a tutorial review written by Huber [10]. The mathematical similarities of the two algorithms allow us to transpose a result of Jones [11] that proves the convergence of projection pursuit algorithms. Let us recall that $\mathbf{V}$ is the closed linear span of vectors in $\mathcal{D}$. We denote by $\mathbf{W}$ the orthogonal complement of $\mathbf{V}$ in $\mathbf{H}$. The orthogonal projectors over $\mathbf{V}$ and $\mathbf{W}$ are respectively written as $\mathbf{P_V}$ and $\mathbf{P_W}$.

Theorem 1 *Let $f \in \mathbf{H}$. The residue $R^m f$ defined by the induction equation (12) satisfies*

$$\lim_{m \to +\infty} \|R^m f - \mathbf{P_W} f\| = 0. \tag{18}$$

Hence

$$\mathbf{P_V} f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{19}$$

and

$$\|\mathbf{P_V} f\|^2 = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2. \tag{20}$$

Theorem 1 proves that the matching pursuit recovers the components of f that belongs to the space spanned by the vectors of $\mathcal{D}$. The proof is in Appendix A. When the dictionary is complete, which means that $\mathbf{V} = \mathbf{H}$, then $\mathbf{P_V} f = f$ and $\mathbf{P_W} f = 0$. Hence

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{21}$$

and

$$\|f\|^2 = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2. \tag{22}$$

The vector f is characterized by the double sequence $(< R^n f, g_{\gamma_n} >, \gamma_n)_{n \in \mathbf{N}}$ that we call structure book. Each γ_n indexes an element selected in the dictionary and $< R^n f, g_{\gamma_n} >$ is the corresponding inner product. The order of elements in a structure book is not important for the reconstruction.

The smallest complete dictionaries are bases. If $\mathcal{D}$ is an orthogonal basis, $< R^n f, g_{\gamma_n} >=< f, g_{\gamma_n} >$. The matching pursuit decomposition is then equivalent to an orthogonal expansion in the basis $\mathcal{D}$. In this case, the indexes γ_n carry no information. Indeed for almost all vectors $f \in \mathbf{H}$, the inner product with elements of the basis are never zero. Hence the sequence $(\gamma_n)_{n \in \mathbf{N}}$ includes exactly once each index of the basis vectors and is thus a permutation of the index set $\mathbf{\Gamma}$ of $\mathcal{D}$. Since the order is unimportant for the reconstruction, the sequence $(\gamma_n)_{n \in \mathbf{N}}$ carries no information. The largest possible dictionary $\mathcal{D}$ is the set of all unit vectors in $\mathbf{H}$. For this dictionary, we can set the optimality factor α to 1 and the matching pursuit converges in one iteration with $g_{\gamma_0} = \frac{f}{\|f\|}$ and $< f, g_{\gamma_0} >= \|f\|$. The index γ_0 characterizes $\frac{f}{\|f\|}$ among all unit vectors of $\mathbf{H}$. If $\mathbf{H}$ has a finite dimension N, the unit sphere is a surface of dimension $N-1$, so γ_0 is characterized by $N-1$ scalars whereas $< f, g_{\gamma_0} >$ is given by 1 scalar. In this case, the index γ_0 carries much more information than $< f, g_{\gamma_0} >$. In general, the balance of information between indexes and inner products depends upon the size of the dictionary.

After m iterations, a matching pursuit decomposes a signal f into

$$f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^m f. \tag{23}$$

If we stop the algorithm at this stage and only record the partial structure book $(< R^n f, g_{\gamma_n} >, \gamma_n)_{0 \leq n < m}$, the summation of equation (23) recovers an approximation of f, with an error equal to $R^n f$. However, this sum is not a linear expansion of the vectors $(g_{\gamma_n})_{0 \leq n < m}$ that approximates f at best. Let $\mathbf{V}_m$ be the space generated by $(g_{\gamma_n})_{0 \leq n < m}$ and $\mathbf{P}_{\mathbf{V}_m}$ be the orthogonal projector on $\mathbf{V}_m$. For any $f \in \mathbf{H}$, $\mathbf{P}_{\mathbf{V}_m} f$ is the closest vector to f that can be written as linear expansion of the m vectors $(g_{\gamma_n})_{0 \leq n < m}$. We derive from (23) that

$$\mathbf{P}_{\mathbf{V}_m} f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + \mathbf{P}_{\mathbf{V}_m} R^m f. \tag{24}$$

If the family of vectors $(g_{\gamma_n})_{0 \leq n < m}$ is not orthogonal, which is generally the case, then $\mathbf{P}_{\mathbf{V}_m} R^m f \neq 0$. The computation of

$$\mathbf{P}_{\mathbf{V}_m} R^m f = \sum_{n=0}^{m-1} x_n g_{\gamma_n}, \tag{25}$$

is called a back-projection. Instead of storing the inner products $< R^n f, g_{\gamma_n} >$ in the structure book, we then store $< R^n f, g_{\gamma_n} > + x_n$ in order to recover $\mathbf{P}_{\mathbf{V}_m} f$ with (24). In this case, the approximation error

$$\mathbf{P}_{\mathbf{W}_m} f = f - \mathbf{P}_{\mathbf{V}_m} f \tag{26}$$

is the orthogonal projection of f on the space $\mathbf{W}_m$, which is orthogonal complement of $\mathbf{V}_m$ in $\mathbf{H}$. One can derive from (23) that

$$\|\mathbf{P}_{\mathbf{W}_m} f\|^2 = \|\mathbf{P}_{\mathbf{W}_m} R^m f\|^2 = \|R^m f\|^2 - \|\mathbf{P}_{\mathbf{V}_m} R^m f\|^2. \tag{27}$$

The reduction of the approximation error thus depends upon $\|\mathbf{P}_{\mathbf{V}_m} R^m f\|$.

The calculation of the coefficients $(x_n)_{0 \leq n < m}$ requires to solve the following linear system. For any g_{γ_k}, $0 \leq k < m$

$$< \mathbf{P}_{\mathbf{V}_m} R^m f, g_{\gamma_k} >=< R^m f, g_{\gamma_k} >= \sum_{n=0}^{m-1} x_n < g_{\gamma_n}, g_{\gamma_k} > . \tag{28}$$

Let us denote $X = (x_n)_{0 \leq n < m}$ and $Y = (< R^m f, g_{\gamma_k} >)_{0 \leq k < m}$. Let $G = (< g_{\gamma_n}, g_{\gamma_k} >)_{0 \leq k < m, 0 \leq n < m}$ be the Gram matrix of the family of selected vectors. The linear system of equations (28) can be written $Y = GX$. The matrix G is non-negative symmetric but might have some zero eigenvalues if the vectors $(g_{\gamma_n})_{0 \leq n < m}$ are linearly dependent. It is often a sparse matrix without any particular structure. Let p be the number of non-zero coefficient of G. The conjugate gradient algorithm, when initialized to $X_0 = 0$, iteratively computes a sequence of vectors X_n that converge to the vector X of minimum norm which satisfies $Y = GX$ [8]. Let κ be the ratio between the largest eigenvalue of G and the smallest non-zero eigenvalue. One can prove [8] that

$$\|X - X_n\| \leq \|X\| \sqrt{\kappa} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n . \tag{29}$$

The main computational burden of each iteration is to apply G to some intermediate residual vector, which requires $O(p)$ operations. The conjugate gradient algorithm thus requires $O(np)$ operations to compute X_n. If n_0 is the rank of G, unless κ^{-1} is comparable to the computational precision, this algorithm guaranties that $X = X_{n_0}$, and clearly $n_0 \leq m$.

A matching pursuit is similar to a shape-gain vector quantizer [16]. The code-book of a shape-gain quantizer is composed of a family of K unit vectors which is equivalent to a dictionary, and of a sequence of scalars to quantize inner product values. The quantization approximates any vector f by projecting it on a vector g_{γ_0}, which correlates best f among the K vectors of the code-book. The inner product $< f, g_{\gamma_0} >$ is quantized by approximating it to the closest scalar stored in the code-book. Vector quantizations algorithms can be extended with a multistage strategy [9]. After quantizing a given vector, the remaining error is quantized once more, and the process continues iteratively. A matching pursuit is similar to a multistage shape-gain vector quantizer. However, a matching pursuit does not quantize the inner products $< R^n f, g_{\gamma_n} >$, as opposed to this vector quantizer. For information processing applications, matching pursuits use very redundant dictionaries of infinite size, whereas vector quantizers are based on finite dictionaries that are best adapted to data compression. Another major difference is that vector quantizations are performed in spaces of low dimension, generally smaller than 16. For example, image quantizers are based on blocks of less than 4 by 4 pixels. On the contrary, a matching pursuits is performed in a signal space $\mathbf{H}$ whose dimension N is equal to the total number of signal samples, which is typically several thousands. The underlined mathematical and algorithmic issues are thus quite different.

4 Matching Pursuit in Finite Spaces

When the signal space H has a finite dimension N, the matching pursuit has specific properties that are studied in this section. The dictionary $\mathcal{D}$ may have an infinite number of elements and we suppose that it is complete. We describe an efficient implementation of matching pursuit algorithms and prove that the norm of the residues decays exponentially.

When the dictionary is very redundant, the search for the vectors that match best the signal residues can mostly be limited to a sub-dictionary $\mathcal{D}_\alpha = (g_\gamma)_{\gamma \in \mathbf{\Gamma}_\alpha} \subset \mathcal{D}$. We suppose that $\mathbf{\Gamma}_\alpha$ is a finite index set included in $\mathbf{\Gamma}$ such that for any $f \in \mathbf{H}$

$$\sup_{\gamma \in \mathbf{\Gamma}_\alpha} |< f, g_\gamma >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< f, g_\gamma >|. \tag{30}$$

Depending upon α and the dictionary redundancy, the set $\mathbf{\Gamma}_\alpha$ can be much smaller than $\mathbf{\Gamma}$. The matching pursuit is initialized by computing the inner products $(< f, g_\gamma >)_{\gamma \in \mathbf{\Gamma}_\alpha}$, and continues by induction as follow. Suppose that we have already computed $(< R^n f, g_\gamma >)_{\gamma \in \mathbf{\Gamma}_\alpha}$, for $n \geq 0$. We search in $\mathcal{D}_\alpha$ for an element $g_{\tilde{\gamma}_n}$ such that

$$|< R^n f, g_{\tilde{\gamma}_n} >| = \sup_{\gamma \in \mathbf{\Gamma}_\alpha} |< R^n f, g_\gamma >|. \tag{31}$$

To find a dictionary element that matches f even better than $g_{\tilde{\gamma}_n}$, we then search with a Newton method for an index γ_n in a neighborhood of $\tilde{\gamma}_n$ in $\mathbf{\Gamma}$ where $|< f, g_\gamma >|$ reaches a local maxima. Clearly

$$|< R^n f, g_{\gamma_n} >| \geq |< R^n f, g_{\tilde{\gamma}_n} >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< R^n f, g_\gamma >|. \tag{32}$$

Let us observe that the choice function mentioned in section 3 is defined indirectly by this double search strategy. Once the vector g_{γ_n} is selected, we compute the inner product of the new residue $R^{n+1} f$ with any $g_\gamma \in \mathcal{D}_\alpha$, with an updating formula derived from equation (12)

$$< R^{n+1} f, g_\gamma > = < R^n f, g_\gamma > - < R^n f, g_{\gamma_n} > < g_{\gamma_n}, g_\gamma > . \tag{33}$$

Since we previously stored $< R^n f, g_\gamma >$ and $< R^n f, g_{\gamma_n} >$, this update requires only to compute $< g_{\gamma_n}, g_\gamma >$. Dictionaries are generally built so that this inner product is recovered with a small number of operations. The number of times we sub-decompose the residues of a given signal f depends upon the desired precision ϵ. The number of iterations is the minimum p such that

$$\|R^p f\| = \|f - \sum_{n=0}^{p-1} < R^n f, g_{\gamma_n} > g_{\gamma_n}\| \leq \epsilon\|f\|. \tag{34}$$

The energy conservation (17) proves that this equation is equivalent to

$$\|f\|^2 - \sum_{n=0}^{p-1} |< R^n f, g_{\gamma_n} >|^2 \leq \epsilon^2\|f\|^2. \tag{35}$$

Since we do not compute the residue $R^n f$, at each iteration we test the validity of (35) to stop the decomposition. The number of iterations p depends upon the decay rate of $\|R^n f\|$. It can vary widely depending upon the signals but is much smaller than N in most applications. The energy of the residual error can be decreased with the back-projection algorithm described in section 3. Many types of dictionaries do not contain any subfamily of less than $N+1$ vectors that are linearly dependent. In this case, once the matching pursuit has selected N different vectors, these vectors spans the whole signal space $\mathbf{H}$. Hence, after back-projection there is no more residual error and f is recovered as a linear expansion of the N selected vectors. However, this basis of $\mathbf{H}$ might be badly conditioned which slows down the convergence of the back-projection algorithm.

The decay of $\|R^n f\|$ depends upon the correlation between the residues and the dictionary elements. Let us define the correlation ratio of a function $f \in \mathbf{H}$ with respect to $\mathcal{D}$ as

$$\lambda(f) = \sup_{\gamma \in \mathbf{\Gamma}} \frac{|< f, g_\gamma >|}{\|f\|}. \tag{36}$$

The following lemma guaranties that for any $f \in \mathbf{H}$, $\lambda(f)$ is larger than a strictly positive constant.

Lemma 1 *Let $\mathcal{D}$ be a complete dictionary in a finite dimensional space* $\mathbf{H}$,

$$I(\lambda) = \inf_{f \in \mathbf{H}} \lambda(f) > 0. \tag{37}$$

The proof of this lemma is in Appendix B. The value of $I(\lambda)$ is the cosine of the maximum possible angle between a direction in $\mathbf{H}$ and the closest direction of a dictionary vector. If $\mathcal{D}$ is an orthogonal basis, one can prove that $I(\lambda) = \frac{1}{\sqrt{N}}$. The next lemma guaranties that $\|R^n f\|$ decays exponentially in a finite dimensional space, with a rate proportional to $\alpha^2 I^2(\lambda)$.

Lemma 2 *Let $f \in \mathbf{H}$. For any $m > 0$*

$$\|R^m f\| \leq \|f\|(1 - \alpha^2 I^2(\lambda))^{\frac{m}{2}}. \tag{38}$$

Proof: the matching pursuit chooses a vector g_{γ_n} that satisfies

$$|< R^n f, g_{\gamma_n} >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< R^n f, g_\gamma >| = \alpha\lambda(R^n f)\|R^n f\|. \tag{39}$$

Since $\|R^{n+1} f\|^2 = \|R^n f\|^2 - |< R^n f, g_{\gamma_n} >|^2$,

$$\|R^{n+1} f\| \leq \|R^n f\|(1 - \alpha^2 \lambda^2(R^n f))^{\frac{1}{2}}, \tag{40}$$

and hence, for any $m > 0$

$$\|R^m f\| \leq \|f\| \prod_{n=0}^{m-1} (1 - \alpha^2 \lambda^2(R^n f))^{\frac{1}{2}} \leq \|f\|(1 - \alpha^2 I^2(\lambda))^{\frac{m}{2}}. \tag{41}$$

□

The lower the correlation ratios of a particular signal f and its residues, the slower the decay of their norm. If the signal f is the sum of a few high energy components that belong to the dictionary, the correlation ratios of f and its residues is high so their norm decrease quickly. These high energy components can be viewed as "coherent structures" with respect to the dictionary. If the residues of f have low-correlation ratios, their norm decay slowly and f must be expanded over many dictionary vectors in order to be well approximated. This means that the information of f is diluted across the dictionary. The extraction of coherent signal structures is further studied in section 7.

5 Matching Pursuit With Time-Frequency Dictionaries

For dictionaries of time-frequency atoms, a matching pursuit yields an adaptive time-frequency transform. It decomposes any function $f(t) \in \mathbf{L}^2(\mathbf{R})$ into a sum of complex time-frequency atoms that best match its residues. This section studies the properties of this particular matching pursuit decomposition. We derive a new type of time-frequency energy distribution by summing the Wigner distribution of each time-frequency atom.

Since a time-frequency atom dictionary is complete, Theorem 1 proves that a matching pursuit decomposes any function $f \in \mathbf{L}^2(\mathbf{R})$ into

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{42}$$

where $\gamma_n = (s_n, u_n, \xi_n)$ and

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s_n}} g(\frac{t-u_n}{s_n}) e^{i\xi_n t}. \tag{43}$$

These atoms are chosen to best match the residues of f.

The matching pursuit algorithm depends upon a choice function that selects at each iteration a vector g_{γ_n} among all vectors that satisfy equation (11). Appendix C proves that we can define choice functions for which the matching pursuit is covariant by dilation, translation and modulation. Let us denote $(g_{\gamma_n^0})_{n\in\mathbf{N}}$ and $(g_{\gamma_n^1})_{n\in\mathbf{N}}$, with $\gamma_n^0 = (s_n^0, u_n^0, \xi_n^0)$ and $\gamma_n^1 = (s_n^1, u_n^1, \xi_n^1)$, the family of time-frequency atoms selected to decompose respectively $f^0(t)$ and $f^1(t)$. Appendix C proves that there exists a class of choice functions such that

$$f^1(t) = \frac{d}{\sqrt{a}} f^0(\frac{t-c}{a}) e^{ibt} \tag{44}$$

if and only if for all $n \geq 0$

$$s_n^0 = \frac{s_n^1}{a},\ u_n^0 = \frac{u_n^1 - c}{a},\ \xi_n^0 = a(\xi_n^1 - b), \tag{45}$$

and

$$< R^n f^0, g_{\gamma_n^0} >= d e^{ic(b-\xi_n^1)} < R^n f^1, g_{\gamma_n^1} > . \tag{46}$$

The translation, modulation and dilation of a function appears as simple modifications of the selected atom indexes. The covariance through dilation, translation and modulation is important to perform a signal analysis that takes into account any of these transformations.

From the decomposition of any $f(t)$ within a time-frequency dictionary, we derive a new time-frequency energy distribution, by adding the Wigner distribution of each selected atom. Let us recall that the cross Wigner distribution of two functions $f(t)$ and $h(t)$ is defined by

$$W[f,h](t,\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t+\frac{\tau}{2}) \overline{h}(t-\frac{\tau}{2}) e^{-i\omega\tau} d\tau. \tag{47}$$

The Wigner distribution of $f(t)$ is $Wf(t,\omega) = W[f,f](t,\omega)$. Since the Wigner distribution is quadratic, we derive from the atomic decomposition (42) of $f(t)$ that

$$\begin{aligned} Wf(t,\omega) &= \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g_{\gamma_n}(t,\omega) \qquad (48) \\ &+ \sum_{n=0}^{+\infty} \sum_{m=0, m\neq n}^{+\infty} < R^n f, g_{\gamma_n} > \overline{< R^m f, g_{\gamma_m} >} W[g_{\gamma_n}, g_{\gamma_m}](t,\omega). \end{aligned}$$

The double sum corresponds to the cross terms of the Wigner distribution. It regroups the terms that one usually tries to remove in order to obtain a clear picture of the energy distribution of $f(t)$ in the time-frequency plane. We thus only keep the first sum and define

$$Ef(t,\omega) = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g_{\gamma_n}(t,\omega). \qquad (49)$$

A similar decomposition algorithm over time-frequency atoms was derived independently by Qian and Chen [14], in order to define this energy distribution in the time-frequency plane. From the well known dilation and translation properties of the Wigner distribution and the expression (43) of a time-frequency atom, we derive that for $\gamma = (s, \xi, u)$

$$W g_\gamma(t,\omega) = W g(\frac{t-u}{s}, s(\omega - \xi)), \qquad (50)$$

and hence

$$Ef(t,\omega) = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g(\frac{t-u_n}{s_n}, s_n(\omega - \xi_n)). \qquad (51)$$

The Wigner distribution also satisfies

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W g(t,\omega) dt\ d\omega = \|g\|^2 = 1, \qquad (52)$$

so the energy conservation equation (22) implies

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Ef(t,\omega) dt\ d\omega = \|f\|^2. \qquad (53)$$

We can thus interpret $Ef(t,\omega)$ as an energy density of f in the time-frequency plane (t,ω). Unlike the Wigner and the Cohen class distributions, it does not include cross terms. It also remains positive if $Wg(t,\omega)$ is positive, which the case when $g(t)$ is Gaussian. On the other hand, the energy density $Ef(t,\omega)$ does not satisfy marginal properties, as opposed to certain Cohen class distributions [1]. The importance of these marginal properties for signal processing is however not clear.

When the signal $f(t)$ is real, to get a decomposition with real expansion coefficients, one must use dictionaries of real time-frequency atoms. For any $\gamma = (s, \xi, u)$, with $\xi \neq 0$, and any phase $\phi \in [0, 2\pi[$, we define

$$g_{(\gamma,\phi)} = \frac{K_{(\gamma,\phi)}}{\sqrt{s}} g(\frac{t-u}{s}) cos(\xi t + \phi). \tag{54}$$

The constant $K_{(\gamma,\phi)}$ is adjusted so that $\|g_{(\gamma,\phi)}\| = 1$. The phase ϕ that was hidden in the complex numbers, now appears explicitly as a parameter of the real atoms. The dictionary of real time-frequency atoms is defined by $\mathcal{D} = (g_{(\gamma,\phi)})_{(\gamma,\phi)\in\Gamma\times[0,2\pi[}$, with $\mathbf{\Gamma} = \mathbf{R}^{+} \times \mathbf{R^2}$. The matching pursuit performed with this dictionary decomposes any real signal $f(t)$ into

$$f(t) = \sum_{n=0}^{+\infty} < R^n f, g_{(\gamma_n,\phi_n)} > g_{(\gamma_n,\phi_n)}(t), \tag{55}$$

where the indexes $(\gamma_n, \phi_n) = (s_n, u_n, \xi_n, \phi_n)$ are chosen to best match the residues of f. For any $\gamma = (s, \xi, u)$, real atoms are related to complex atoms by

$$g_{(\gamma,\phi)}(t) = \frac{K_{(\gamma,\phi)}}{2} (e^{i\phi} g_{\gamma}(t) + e^{-i\phi} g_{\gamma^-}(t)), \tag{56}$$

where $\gamma^- = (s, -\xi, u)$. However, one can show that the real matching pursuit decomposition (55) is not equivalent to the complex decomposition (42), because the two vectors $g_\gamma(t)$ and $g_{\gamma^-}(t)$ are not orthogonal.

The time-frequency energy distribution of a real function $f(t)$ is derived from its matching pursuit decomposition, by summing the Wigner distribution of the underlined complex atoms:

$$Ef(t,\omega) = \sum_{n=0}^{+\infty} |< R^n f, g_{(\gamma_n,\phi_n)} >|^2 \frac{1}{2} (Wg_{\gamma_n}(t,\omega) + Wg_{\gamma_n^-}(t,\omega)). \tag{57}$$

By inserting equation (50) in this expression, we obtain

$$Ef(t,\omega) = \frac{1}{2} \sum_{n=0}^{+\infty} |< R^n f, g_{(\gamma_n,\phi_n)} >|^2 \Big(Wg(\tfrac{t-u_n}{s_n}, s_n(\omega - \xi_n)) + Wg(\tfrac{t-u_n}{s_n}, s_n(\omega + \xi_n))\Big). \tag{58}$$

This distribution also satisfies the energy density property (53).

In signal processing applications of time-frequency matching pursuits, we process directly the discrete parameters $(\gamma_n, \phi_n) = (s_n, \xi_n, u_n, \phi_n)$ and $< R^n f, g_{\gamma_n} >$ of the selected atoms, rather than the energy density $Ef(t,\omega)$. Indeed, these parameters carry all the necessary information and are much easier to manipulate than the two-dimensional map $Ef(t,\omega)$. This energy distribution is rather used for the visualization of the structure book information. If $g(t)$ is the Gaussian window

$$g(t) = 2^{1/4} e^{-\pi t^2}, \tag{59}$$

then

$$Wg(t,\omega) = 2e^{-2\pi(t^2 + (\frac{\omega}{2\pi})^2)}. \tag{60}$$

The time-frequency atoms $g_\gamma(t)$ are then called Gabor functions. The time-frequency energy distribution $Ef(t,\omega)$ is a sum of Gaussian blobs whose locations and variances along the time and frequency axes depend upon the parameters (s_n, u_n, ξ_n).

As explained in section 4, to implement efficiently a matching pursuit, we must avoid computing the inner products of the signal residues with all the dictionary vectors. The following theorem guaranties that if we discretize appropriately the Gabor dictionary, one can obtain a sub-dictionary that satisfies the property (30).

Theorem 2 *Let Δu and $\Delta\xi$ be respectively a time and a frequency discretization interval that satisfy*

$$\Delta u = \frac{\Delta\xi}{2\pi} < 1. \tag{61}$$

Let $a > 1$ be an elementary dilation factor. Let $\mathbf{\Gamma}_\alpha$ be the discrete subset of $\mathbf{\Gamma} = \mathbf{R}^+ \times \mathbf{R^2}$, of all indexes $\gamma = (a^j, pa^j\Delta u, ka^{-j}\Delta\xi)$, for $(j,p,k) \in \mathbf{Z^3}$. There exists a constant $\alpha > 0$ such that for all $f \in \mathbf{L}^2(\mathbf{R})$

$$\sup_{\gamma\in\mathbf{\Gamma}_\alpha} |< f, g_\gamma >| \geq \alpha \sup_{\gamma\in\mathbf{\Gamma}} |< f, g_\gamma >|. \tag{62}$$

Appendix D gives a proof of this theorem. The fast numerical implementation of a matching pursuit in a Gabor dictionary is based on this theorem.

6 Discrete Matching Pursuit in Gabor Dictionaries

We explain the discrete implementation of a matching pursuit for a dictionary of Gabor time-frequency atoms. Numerical examples are shown at the end of this section. We suppose that our signal is real and has N samples. The space $\mathbf{H}$ is the set of infinite discrete signals of period N. Due to the limitations of the sampling rate and the signal size, the scale s can only vary between 1 and N. The window function $g(t)$ is the normalized Gaussian given by equation (59). To obtain a discrete and periodic signal, at any scale s, the window function is uniformly sampled and periodized over N points

$$g_s(n) = \frac{K_s}{\sqrt{s}} \sum_{p=-\infty}^{+\infty} g(\frac{n-pN}{s}). \tag{63}$$

The constant K_s normalizes the discrete norm of g_s. For any integer $0 \leq p < N$ and $0 \leq k < N$, we denote $\gamma = (s, p, \frac{2\pi k}{N})$ and define the discrete Gabor atom

$$g_\gamma(n) = g_s(n-p)e^{i\frac{2\pi k}{N}n}. \tag{64}$$

The discrete complex Gabor dictionary is the set of all such atoms for $s \in]1, N[$ and p, k integers between 0 and N. To this dictionary of atoms, we add the canonical basis of discrete Diracs and the discrete Fourier basis of complex exponentials. For $\gamma = (1, p, 0)$, $g_\gamma(n)$ is a discrete Dirac centered at p. For $\gamma = (N, 0, k)$, $g_\gamma(n) = \frac{1}{\sqrt{N}} e^{i\frac{2\pi k}{N}n}$.

Similarly to equation (56), for any $\gamma = (s, p, \frac{2\pi k}{N})$ and $\phi \in [0, 2\pi[$, real discrete time-frequency atoms are given by

$$g_{(\gamma,\phi)}(n) = K_{(\gamma,\phi)} g_s(n-p) cos(\frac{2\pi k}{N}n + \phi), \tag{65}$$

with $K_{(\gamma,\phi)}$ such that $\|g_{(\gamma,\phi)}\| = 1$. Appendix E describes an efficient implementation of a matching pursuit with this real discrete Gabor dictionary and gives informations to obtain a copy of a matching pursuit software. The implementation follows the general algorithm described in section 4. We compute the inner products of the signal residues with the complex Gabor atoms (64) and recover the phase from the complex coefficients. As suggested by theorem 2 and the implementation algorithm of section 4, we only compute the inner product of the signal residues with a subset $\mathcal{D}_\alpha = (g_\gamma)_{\gamma \in \mathbf{\Gamma}_\alpha}$ of the complex Gabor dictionary. The index set $\mathbf{\Gamma}_\alpha$ is composed of all

$\gamma = (a^j, pa^j \Delta u, ka^{-j} \Delta\xi)$, with $a = 2$, $\Delta u = \frac{1}{2}$, $\Delta\xi = \pi$, $0 < j < \log_2 N$, $0 \le p < N2^{-j+1}$ and $0 \le k < 2^{j+1}$. We also add the discrete Dirac and Fourier bases to $\mathcal{D}_\alpha$. The number of vectors in $\mathcal{D}_\alpha$ is $O(N \log_2 N)$. The implementation of the matching pursuit iterations is further described in Appendix E. The search over $\mathcal{D}_\alpha$ finds the approximate scale, time and frequency localization of the main signal structures. These values are then refined with a Newton search strategy to recover the time-frequency parameters that best match the signal components. Each iteration requires $O(N \log N)$ operations and as much CPU time as a fast Fourier transform subroutine applied to a signal of N samples.

Fig. 1(a) is a signal f of 512 samples that is built by adding chirps, truncated sinusoidal waves and waveforms of different time-frequency localizations. No Gabor function have been used to construct this signal. Fig. 1(b) shows the time-frequency energy distribution $Ef(t,\omega)$. Since $Ef(t,\omega) = Ef(t,-\omega)$, we only display its values for $\omega \ge 0$. Each Gabor time-frequency atom selected by the matching pursuit is an elongated Gaussian blob in the time-frequency plane. We clearly see appearing two chirps that cross each others, with a localized time-frequency waveform at the top of their crossing point. We can also detect closely spaced Diracs, and truncated sinusoidal waves having close frequencies. Several isolated localized time-frequency components also appear in this energy distribution. The curve (a) in Fig. 2 gives the decay of $\log_{10} \frac{\|R^n f\|}{\|f\|}$ as a function of the number of iterations n. For $n \le 130$, $\|R^n f\|$ has a relatively faster decay. These iterations correspond to the coherent signal structures, as shown in section 7. For $n \ge 130$, the decay rate is almost constant. This confirms the exponential decay proved by lemma 2. For any $n \ge 0$, the back-projection algorithm described in section 3 recovers a better approximation of f from the n atoms selected from the dictionary. The reconstruction error is then the orthogonal projection of f on the space $\mathbf{W}_n$ that is orthogonal to the n vectors selected by the matching pursuit. The back-projection requires much less computation than the matching pursuit. The curve (b) in Fig. 2 gives the decay of $\log_{10} \frac{\|\mathbf{P}_{\mathbf{W}_n} f\|}{\|f\|}$. For $n \le 130$, $\|R^n f\| \approx \|\mathbf{P}_{\mathbf{W}_n} f\|$. It means that the matching pursuit computes a close approximation of the orthogonal projection of f on the n vectors selected from the dictionary. For $n = 300$, $\|R^n f\| = 1.5 \|\mathbf{P}_{\mathbf{W}_n} f\|$. For $n = N = 512$, $\log_{10} \frac{\|\mathbf{P}_{\mathbf{W}_n} f\|}{\|f\|}$ drops to $-\infty$ because $\mathbf{P}_{\mathbf{W}_n} f = 0$. This indicates that the N vectors selected by the matching pursuit are linearly independent and thus define a basis of the signal space $\mathbf{H}$. The relative gain of the back-projection is important when the number of iterations is of the order of the dimension of the signal space. For almost all signals f, the decays of $\|R^n f\|$ and $\|\mathbf{P}_{\mathbf{W}_m} f\|$ have the same qualitative behavior as in Fig. 2.

Fig. 3(a) is the graph of a speech recording corresponding to the word "greasy", sampled at 16 kHz. From the time-frequency energy displayed in Fig. 3(b), we can see the low-frequency component of the "g" and the quick burst transition to the "ea". The "ea" has many harmonics that are lined up but we can also see localized high-frequency impulses that correspond to the pitch. The "s" component has a time-frequency energy spread over a high-frequency interval. Most of the signal energy is characterized by few time-frequency atoms. For $n = 250$ atoms, $\frac{\|R^n f\|}{\|f\|} = .169$, although the signal has 5782 samples, and the sound recovered from these atoms is of excellent quality.

Fig. 4(a) shows a signal obtained by adding a Gaussian white noise to the speech recording given in Fig. 3(a), with a signal to noise ratio of 1.5 db. Fig. 4(b) is the time-frequency energy distribution of this noisy signal. The white noise generates time-frequency atoms spread across the whole time-frequency plane, but we can still distinguish the time-frequency structures of the original signal because their energy is better concentrated in this plane.

7 Noise and Coherent Structures

Generally, the notion noise versus signal information is ill-defined. Even though a signal component might carry a lot of information, it is often considered as noise if we can not make sense out of it. In a crowd of people speaking a language we do not understand, surrounding conversations are generally perceived as a noise background. However, our attention will be attracted by any remote conversation in a language we know. In this case, what is important is not the information content but whether this information is in a coherent format with respect to our system of interpretation. A matching pursuit decomposition in a given dictionary defines a system of interpretation for signals. We study the notion of coherence and describe an algorithm that isolates signal structures that are coherent with respect to a given dictionary.

Coherent signal components have a strong correlation with some dictionary vectors. The more coherent a signal, the larger the correlation ratios

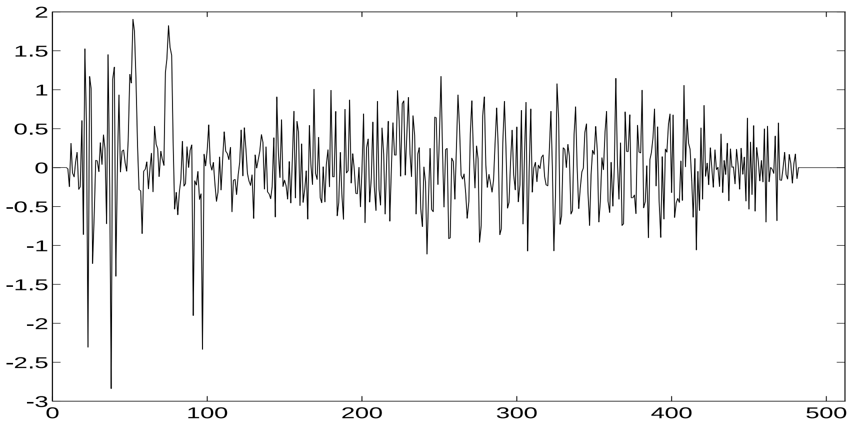

Figure 1(a): Signal of 512 samples built by adding chirps, truncated sinusoidal waves and waveforms of different time-frequency localizations.

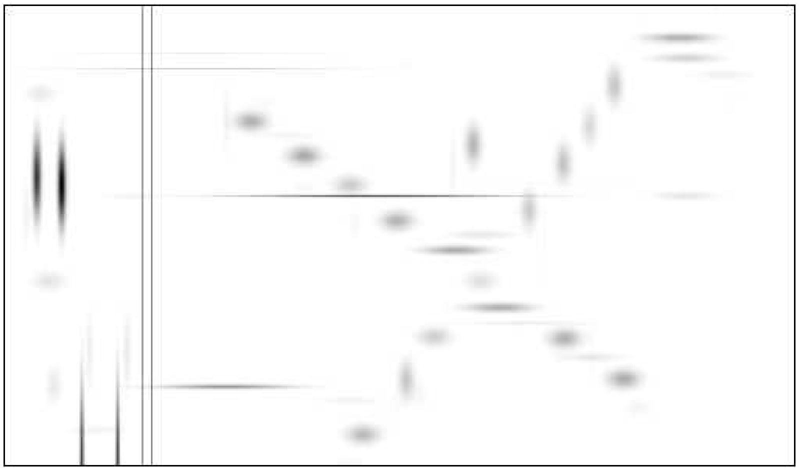

Figure 1(b): Time-frequency energy distribution $Ef(t, \omega)$ of the signal shown in (a). The horizontal axis is time. The vertical axis is frequency. The highest frequencies are on the top. The darkness of this time-frequency image increases with the value $Ef(t, \omega)$.

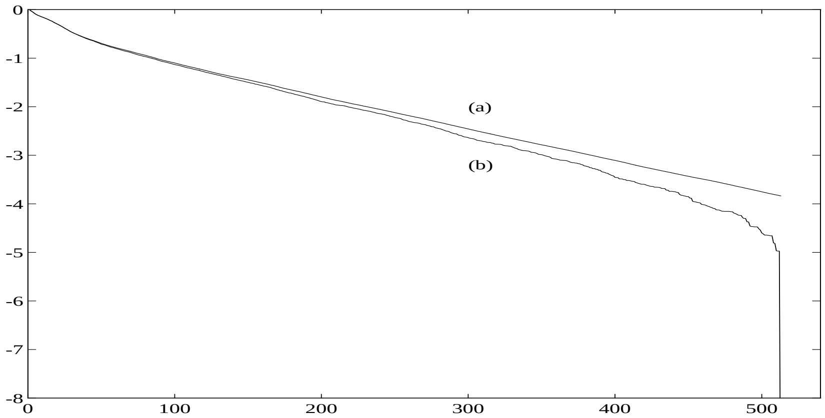

Figure 2(a): The curve (a) gives the decay of $log_{10}\frac{\|R^n f\|}{\|f\|}$, as function of the number of iterations n, for the signal f of Fig.1(a). The curve (b) gives the decay of $log_{10}\frac{\|\mathbf{P}_{\mathbf{W_n}}\mathbf{f}\|}{\|\mathbf{f}\|}$.

of the signal residues

$$\lambda(R^n f) = \sup_{\gamma \in \mathbf{\Gamma}} \frac{|< R^n f, g_\gamma >|}{\|R^n f\|}. \tag{66}$$

The matching pursuit selects vectors g_{γ_n} that almost best correlate the signal residues. Let us denote

$$\tilde{\lambda}(R^n f) = \frac{|< R^n f, g_{\gamma_n} >|}{\|R^n f\|}. \tag{67}$$

Equation (11) implies that

$$\tilde{\lambda}(R^n f) \leq \lambda(R^n f) \leq \frac{1}{\alpha}\tilde{\lambda}(R^n f). \tag{68}$$

For any $h \in \mathbf{H}$, the choice procedure implies that $\tilde{\lambda}(h) = \tilde{\lambda}(\frac{h}{\|h\|})$. Hence, $\tilde{\lambda}(h)$ only depends upon the position of $\frac{h}{\|h\|}$ on the unit sphere of the signal space $\mathbf{H}$. Let W be a discrete Gaussian white noise. For any $n \geq 0$, the average value of $\tilde{\lambda}(R^n f)$ measured with a uniform probability distribution over the unit sphere, is equal to the expected value $E(\tilde{\lambda}(R^n W))$. Indeed, after normalization, the realizations of a discrete Gaussian white noise have

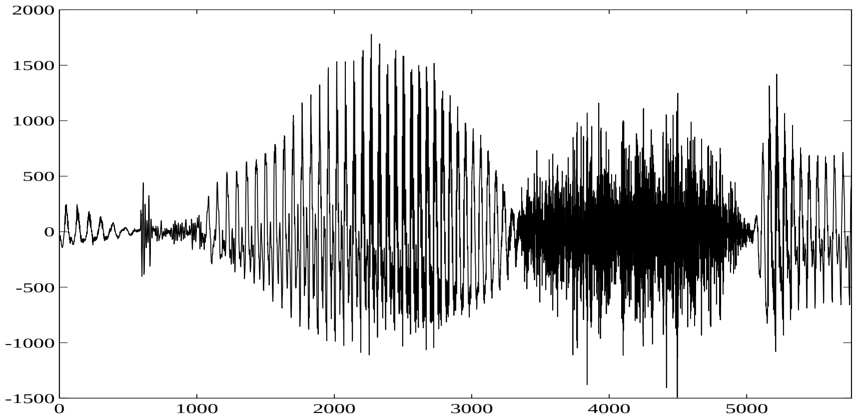

Figure 3(a): speech recording of the word "greasy", sampled at 16 kHz.

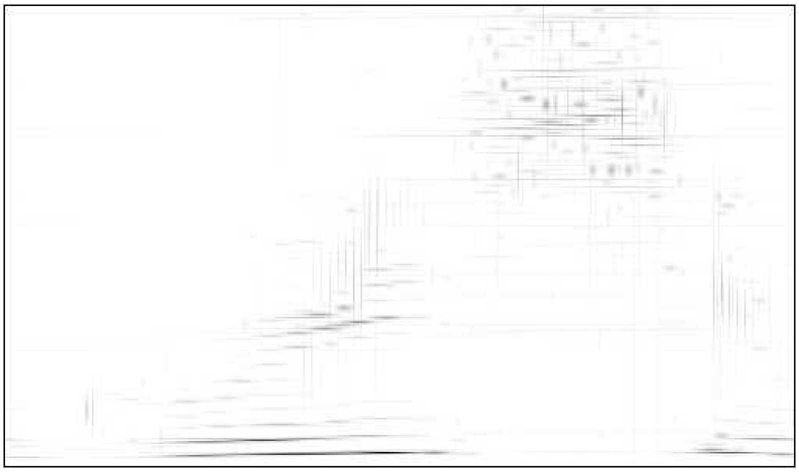

Figure 3(b): time-frequency energy distribution of the speech recording shown in (a). We see the low-frequency component of the "g", the quick burst transition to the "ea" and the harmonics of the "ea". The "s" has energy spread over high frequencies.

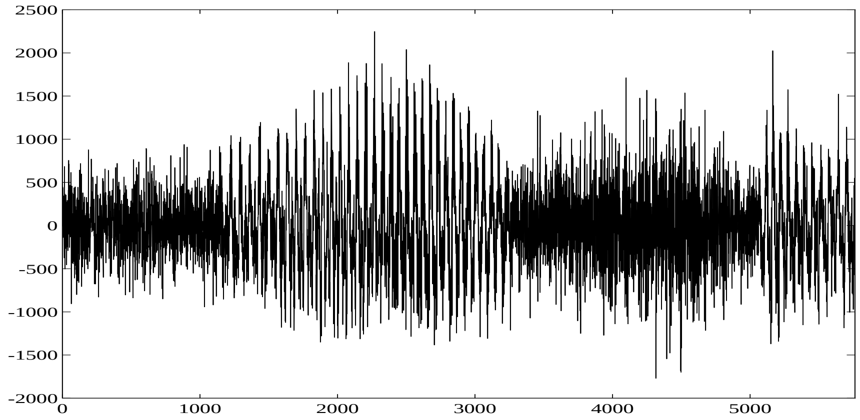

Figure 4(a): signal obtained by adding a Gaussian white noise to the speech recording shown in Fig.3(a). The signal to noise ratio is 1.5db.

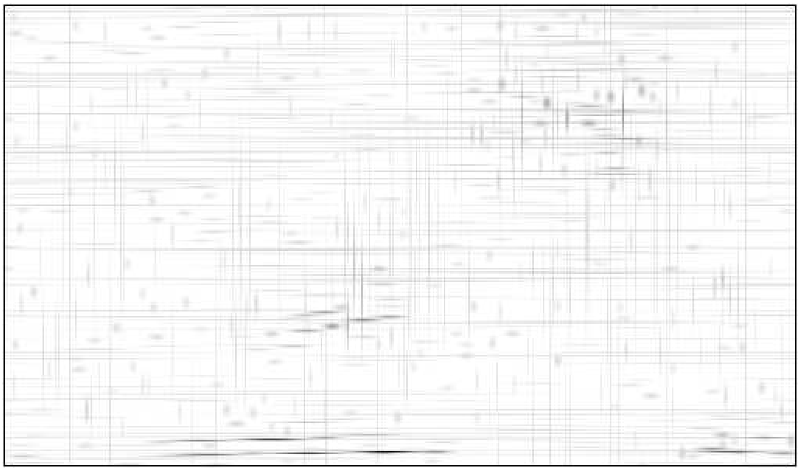

Figure 4(b): time-frequency energy distribution of the noisy speech signal. The energy distribution of the white noise is spread across the whole time-frequency plane.

a uniform probability distribution over the unit sphere of $\mathbf{H}$. We define the coherent structures of a signal f as the first m vectors $(g_{\gamma_n})_{0\leq n<m}$ that have a higher than average correlation with $R^n f$. In other words, f has m coherent structures if and only if for $0 \leq n < m$

$$\tilde{\lambda}(R^n f) > E(\tilde{\lambda}(R^n W)), \tag{69}$$

and

$$\tilde{\lambda}(R^m f) \leq E(\tilde{\lambda}(R^m W)). \tag{70}$$

Equation (13) proves that $\tilde{\lambda}(R^n f)$ is related to the decay of $\|R^n f\|$ by

$$\tilde{\lambda}(R^n f) = \sqrt{1 - \frac{\|R^{n+1} f\|^2}{\|R^n f\|^2}}. \tag{71}$$

One can verify that for a Gabor dictionary, the signal shown in Fig. 1(a) has $m = 130$ coherent structures that correspond to the iterations where the $\|R^n f\|$ has a relatively faster decay in Fig. 2.

For all the dictionaries that we studied numerically, we have observed that when n increases, $E(\tilde{\lambda}(R^n W))$ converges quickly to a constant $E(\tilde{\lambda}(R^\infty W))$. In fact, the process $R^n W$ seems to converge to a process $R^\infty W$ that we call dictionary noise, whose properties are now being studied. The realizations of a dictionary noise have an energy that is uniformly spread across the whole dictionary. For a Gabor dictionary this process is a stationary white noise, that is not Gaussian. The curve (c) in Fig. 5 gives the value of $E(\tilde{\lambda}(R^n W))$ as a function of n, for a discrete Gaussian white noise of 5762 samples, decomposed in a Gabor dictionary. The limit is $E(\tilde{\lambda}(R^\infty W)) = 0.0506$.

The curve (a) in Fig. 5 gives the value of $\tilde{\lambda}(R^n f)$ as a function of n for the speech recording f shown in Fig. 3(a). The number of coherent structures is the abscissa of the first intersection between curves (a) and (c), which is located at $n = 698$. We have observed numerically that after removing the coherent structures from a signal f, the residue $R^m f$ behaves like a realization of the dictionary noise $R^\infty W$. This property remains to be studied more precisely. The curve (b) in Fig. 5 gives the value of $\tilde{\lambda}(R^n f)$ for the noisy speech signal in Fig. 4(a). The noise has destroyed the low-energy coherent structures and only 76 coherent structures remains at an SNR of 1.5db. Fig. 6(a) is the time-frequency energy distribution of these $m = 76$ coherent structures. Fig 6(b) is the signal reconstructed from these time-frequency atoms. The SNR of the reconstructed signal is $6.8db$. The white noise has been removed and this signal has a good auditory quality because the main time-frequency structures of the original speech signal have been retained.

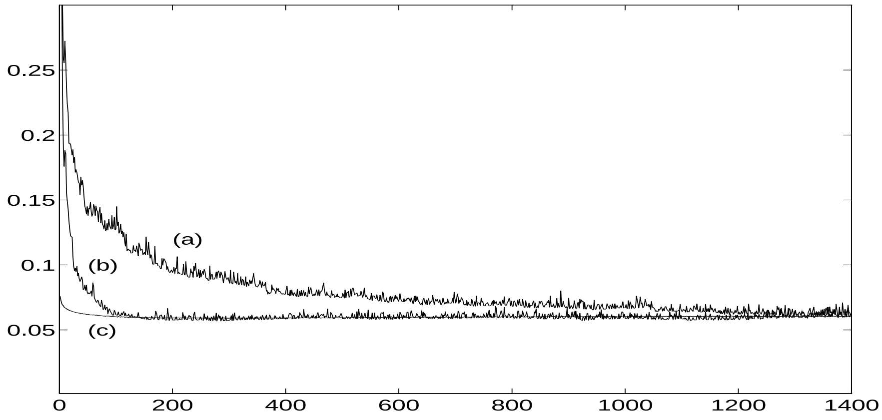

Figure 5(a): the curve (c) is a plot of $E(\tilde{\lambda}(R^nW))$ as a function of n, for a discrete Gaussian white noise of 5786 samples. The curves (a) and (b) give the values of $\tilde{\lambda}(R^nf)$ for the speech signal in Fig.3(a) and the noisy speech signal in Fig.4(a).

8 Wavepacket Dictionary

A wavepacket dictionary is a family of orthonormal bases composed of vectors that are well localized both in time and frequency. It is computed with a quadrature mirror filter bank algorithm [15]. Through our numerical experiments with wavepacket dictionaries, we intend to compare matching pursuit decompositions with the best basis algorithm of Coifman and Wikerhauser [4], that selects an "optimal" orthonormal basis within the wavepacket dictionary. This highlights the respective advantages of procedures that globally adapt the signal representation versus the greedy strategy of a matching pursuit, that locally optimizes the decomposition.

For signals of N samples, each vector g_γ of a wavepacket dictionary is indexed by $\gamma = (j, p, k)$, with $0 \leq j \leq \log_2(N)$, $0 \leq p \leq 2^{-j}N$, $0 \leq p \leq 2^j$. Such a vector has a similar time-frequency localization properties as a discrete window function, dilated by 2^j, centered at $2^j(p + \frac{1}{2})$, and modulated by sinusoidal wave of frequency $2\pi 2^{-j}(k + \frac{1}{2})$. The wavepacket dictionary $\mathcal{D} = (g_\gamma)_{\gamma\in\mathbf{\Gamma}}$ includes $(N+1)\log_2(N)$ vectors. For any discrete signal $f(n)$ of N samples, the inner products $(< f, g_\gamma >)_{\gamma\in\mathbf{\Gamma}}$ are computed with a filter bank algorithm based on quadrature mirror filters, that requires

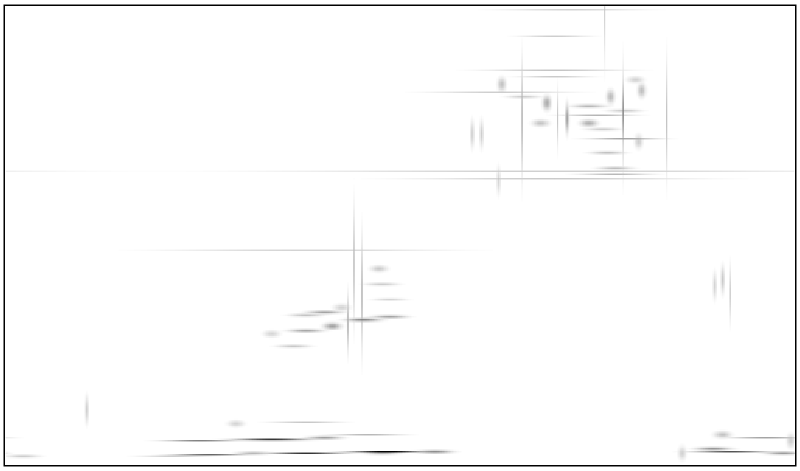

Figure 6(a): time-frequency energy distribution of the $m = 76$ coherent structures of the noisy speech signal shown in Fig.4(a).

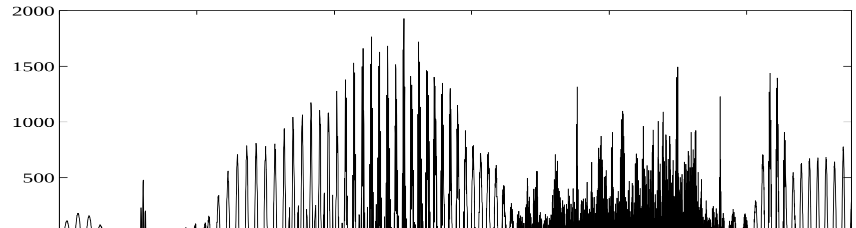

Figure 6(b): Fig. 6(a): time-frequency energy distribution of the $m = 76$ coherent structures of the noisy speech signal shown in Fig. 4. (b): signal reconstructed from the 76 coherent structures shown in (a). The white noise has been removed.

$O(N \log_2(N))$ operations [4]. The implementation of the matching pursuit decomposition follows the general outline of the algorithm described in section 4. In this case, we set the optimality factor α to 1 and search over the whole dictionary $\mathcal{D}$ because it is not to large. To compute the inner product updating formula (33), we calculate the inner product of wavepacket vectors from the coefficients of the quadrature mirror filters [4]. Each matching pursuit iteration requires $O(N \log_2(N))$ operations.

Fig. 7(a) shows the structure book $(< R^n f, g_{\gamma_n} >, \gamma_n)_{n \in \mathbf{N}}$ of the signal in Fig. 1(a), with the display conventions of Coifman and Wickerhauser[4]. The wavepacket dictionary is built with the Daubechies 6 quadrature mirror filters [5]. The horizontal and vertical axes of Fig. 10 are respectively the time and frequency axes. Each vector g_{γ_n}, for $\gamma_n = (j_n, k_n, p_n)$, is represented by a rectangle which is centered at the time $2^{j_n}(p_n + \frac{1}{2})$ and at the frequency $2\pi 2^{-j_n}(k_n + \frac{1}{2})$. This rectangle has a width of 2^{j_n} along time and $2^{-j_n}\pi$ along frequencies. It gives an approximate idea of the localization in time and frequency of the atom g_{γ_n}, but in reality g_{γ_n} is much more spread in time and frequency than the zone indicated by its rectangle. Wavepacket functions are not as well localized in time and frequency as Gabor functions. When the scale 2^j increases, these atoms have a complicated time-frequency localization studied by Coifman, Meyer and Wickerhauser [3]. The time-frequency image obtained with this wavepacket dictionary is similar to the energy distribution in Fig. 1(b), obtained with Gabor dictionary. Some signal features do not appear as clearly because wavepackets are not as well localized in time and frequencies as Gabor functions. Moreover, wavepacket functions do not include a phase parameter and thus can not match signal components as well Gabor functions. We must also mention that the Gabor dictionary includes Gabor functions translated in time and frequency over a much finner grid than wavepackets, so that the different time-frequency signal features can be located more precisely. Although the Gabor dictionary is much larger than the wavepacket dictionary, the matching pursuit does not require much more calculations because we limit most of the computations to a sub-dictionary $\mathcal{D}_\alpha$ that is approximately of the same size as the wavepacket dictionary.

By combining the vectors of a wavepacket dictionary, Coifman and Wickerhauser [4] proved that we can build 2^N different orthonormal bases. They have introduced an algorithm that finds the orthonormal basis $(g_{\gamma_n})_{1 \leq n \leq N}$

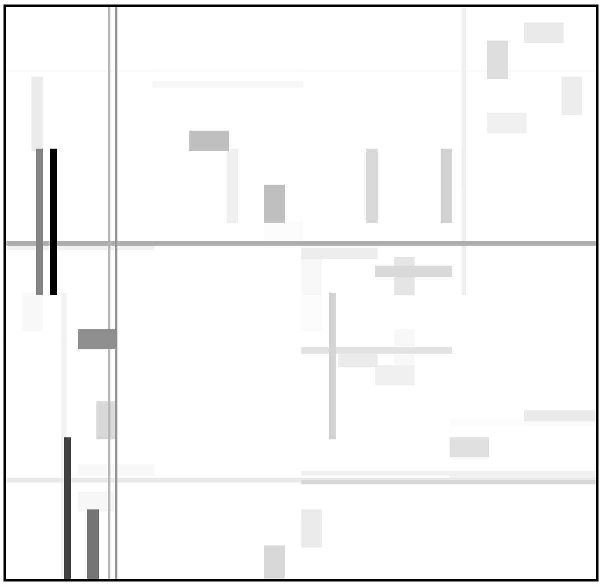

Figure 7(a): time-frequency display of the wavepacket structure book of the signal shown in Fig.1(a). Each rectangle roughly represents the location and time-frequency spread of a selected wavepacket function.

Figure 7(b): time-frequency display of the signal in Fig.1(a) decomposed in the best wavepacket orthonormal basis.

of $\mathcal{D}$ which minimizes the entropy

$$\sum_{n=1}^{N} |< f, g_{\gamma_n} >|^2 \log_2(|< f, g_{\gamma_n} >|^2). \tag{72}$$

The choice of this "optimal" orthonormal basis is thus obtained through a global minimization over all the signal components. Fig. 7(b) displays the structure book $(< f, g_{\gamma_n} >, \gamma_n)_{n \in \mathbf{N}}$ that is obtained by decomposing the signal of Fig. 1(a) in the optimal wavepacket orthonormal basis. One can hardly distinguish many of the signal components, including the two chirps. The entropy optimization creates a competition between the signal components that are in the same frequency range, but have different time-frequency signatures. Since the signal is not stationary, the global entropy minimization is driven by the transients of highest energy. It leads to a choice of orthonormal basis that is well adapted to represent the corresponding transients, but not to represent other signal structures that have different time-frequency behaviors. For highly non-stationary signals, the entropy minimization produces mismatch between the "best" orthonormal basis and many local signal components. On the contrary, a matching pursuit is a greedy algorithm that locally optimizes the choice of the wavepacket function, for each signal residue. It can thus adapt itself to varying structures. On the other hand, this greedy strategy requires more computations than the best basis decomposition algorithm, whose total complexity is $O(N \log N)$. The best basis algorithm is thus better suited to represent simpler signals that have stationary properties. The global optimization is then valid locally, and yields good results.

9 Conclusion

Matching pursuits provide extremely flexible signal representations since the choice of dictionaries is not limited. We showed that time-frequency dictionaries yield adaptive decompositions where signal structures are represented by atoms that match their time-frequency signature. The properties of the signal components are explicitly given by the scale, frequency, time and phase indexes of the selected atoms. This representation is therefore well adapted to information processing.

Compact signal coding is another important domain of application of matching pursuits. For a given class of signals, if we can adapt the dictionary to minimize the storage for a given approximation precision, we are guaranteed to obtain better results than decompositions on orthonormal bases. Indeed, an orthonormal decomposition is a particular case of matching pursuit where the dictionary is the orthonormal basis. For dictionaries that are not orthonormal bases, we must code the inner products of the structure book but also the indexes of the selected vectors. This requires to quantize the inner product values and use a dictionary of finite size. The matching pursuit decomposition is then equivalent to a multistage shape-gain vector quantization in a very high dimensional space.

For information processing or compact signal coding, it is important to have strategies to adapt the dictionary to the class of signal that is decomposed. Time-frequency dictionaries include vectors that are spread between the Fourier and Dirac bases. They are regularly distributed of the unit sphere of the signal space and are thus well adapted to decompose signals over which we have little prior information. When enough prior information is available, one can adapt the dictionary to the probability distribution of the signal class within the signal space $\mathbf{H}$. Learning a dictionary is equivalent to finding the important inner structures of the signals that are decomposed. Classical algorithms such as LBG to optimize code-books [9] do not converge to satisfying solutions in such high dimensional vector spaces. Finding strategies to optimize dictionaries in high dimensions is an open problem that shares similar features with learning problems in neural networks.

Acknowledgements We thank Francois Bergeaud, Wen Liang Hwang and Mike Orszag who helped us to develop the software. We are also grateful to Dave Donoho and Iain Johnstone for showing us the relations between this work and projection pursuit regressions.

Appendix A: *Proof of Theorem 1*

This appendix is a translation in the matching pursuit context of Jones's proof [11] for the convergence of projection pursuit regressions.

Lemma 3 *Let $h_n = < R^n f, g_{\gamma_n} > g_{\gamma_n}$. For any $n \geq 0$ and $m \geq 0$,*

$$|< h_m, R^n f >| \leq \frac{1}{\alpha} \|h_m\| \ \|h_n\|. \tag{73}$$

Proof: Since $h_m = < R^m f, g_{\gamma_m} > g_{\gamma_m}$,

$$|< h_m, R^n f >| = |< R^m f, g_{\gamma_m} >< g_{\gamma_m}, R^n f >| = \|h_m\| \ |< g_{\gamma_m}, R^n f >|. \tag{74}$$

Equation (11) implies

$$|< h_m, R^n f >| \leq \|h_m\| \ |< R^n f, g_{\gamma_n} >| \frac{1}{\alpha} = \frac{1}{\alpha} \|h_m\| \ \|h_n\|. \tag{75}$$

□

Lemma 4 *If $(s_n)_{n \in \mathbf{Z}}$ is a positive sequence such that $\sum_{n=0}^{+\infty} s_n^2 < +\infty$, then*

$$\lim_{n \to +\infty} \inf s_n \sum_{k=0}^{n} s_k = 0. \tag{76}$$

Proof: For any $\epsilon > 0$, we choose n such that $\sum_{i=n}^{+\infty} s_i^2 < \epsilon/2$. Since $\lim_{k \to +\infty} s_k = 0$, we can choose k large enough such that $s_k \sum_{i=0}^{n} s_i \leq \epsilon/2$. Let s_j be the minimum element for indexes between $n+1$ and k,

$$s_j \sum_{k=0}^{j} s_k = s_j \sum_{k=0}^{n} s_k + s_j \sum_{k=n+1}^{j} s_k \leq \epsilon/2 + \sum_{k=n+1}^{j} s_k^2 \leq \epsilon. \tag{77}$$

□

To prove Theorem 1, we prove that the sequence $(R^n f)_{n \in \mathbf{N}}$ is a Cauchy sequence. Let $N \geq 0$ and $M \geq 0$

$$\|R^N f - R^M f\|^2 = \|R^M f - R^M f + \sum_{n=N}^{M-1} h_n\|^2 \tag{78}$$

$$= \|R^N f\|^2 + \|R^M f\|^2 - 2\|R^M f\|^2 - 2 \sum_{n=N}^{M-1} Real(< R^M f, h_n >). \tag{79}$$

Lemma 3 implies that

$$\|R^N f - R^M f\|^2 \leq \|R^N f\|^2 - \|R^M f\|^2 + \frac{2}{\alpha}\|h_M\| \sum_{n=N}^{M-1} \|h_n\|. \tag{80}$$

The energy conservation equation (13) proves that the sequence $(\|R^n f\|)_{n\in\mathbf{N}}$ is monotonically decreasing and thus converges to some value R_∞. Let $\epsilon > 0$, there exist $K > 0$ such that for all $m > K$, $\|R^m f\|^2 \leq R_\infty^2 + \epsilon^2$. Let $p > 0$. We want to estimate $\|R^m f - R^{m+p} f\|$, for $m > K$. Equation (17) proves that $\sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 = \sum_{n=0}^{+\infty} \|h_n\|^2 \leq \|f\|^2 < +\infty$, hence lemma 4 implies that there exist $q > m + p$ such that

$$\|h_q\| \sum_{n=0}^{q} \|h_n\| \leq \epsilon^2. \tag{81}$$

We can decompose

$$\|R^m f - R^{m+p} f\| \leq \|R^m f - R^q f\| + \|R^{m+p} f - R^q f\|. \tag{82}$$

Equation (80) for $N = m$ and $M = q$ implies

$$\|R^m f - R^q f\|^2 \leq \epsilon^2 + \frac{2}{\alpha}\epsilon^2. \tag{83}$$

Similarly

$$\|R^{m+p} f - R^q f\|^2 \leq \epsilon^2 + \frac{2}{\alpha}\epsilon^2. \tag{84}$$

Hence

$$\|R^m f - R^{m+p} f\| \leq 2\epsilon\sqrt{1 + \frac{2}{\alpha}}, \tag{85}$$

which proves that $(R^n f)_{n\in\mathbf{N}}$ is a Cauchy sequence. Let

$$R^\infty f = \lim_{n\to+\infty} R^n f.$$

We know that $\lim_{n\to+\infty} |< R^n f, g_{\gamma_n} >| = 0$. Since

$$|< R^n f, g_{\gamma_n} >| \geq \alpha \sup_{\gamma\in\mathbf{\Gamma}} |< R^n f, g_\gamma >|,$$

for any $\gamma \in \mathbf{\Gamma}$, $\lim_{n\to+\infty} |< R^n f, g_\gamma >| = 0$, and thus $|< R^\infty f, g_\gamma >| = 0$. This implies that $R^\infty f \in \mathbf{W}$. Since

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^\infty f, \tag{86}$$

and $\sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n} \in \mathbf{V}$, we derive that

$$\mathbf{P_V} f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{87}$$

and

$$\mathbf{P_W} f = R^\infty f. \tag{88}$$

□

Appendix B: *Proof of Lemma 1*

If this lemma was wrong, one could construct a sequence of unit vectors $(f_n)_{n\in\mathbf{N}}$ and a sequence $(\lambda_n)_{n\in\mathbf{N}}$ of decreasing real numbers converging to zero such that for all $n > 0$

$$\sup_{\gamma\in\mathbf{\Gamma}} |< f_n, g_\gamma >| \leq \lambda_n. \tag{89}$$

Since the unit sphere of the finite dimensional space $\mathbf{H}$ is compact, there exists a sub-sequence $(f_{n_p})_{p\in\mathbf{N}}$ that converges to a unit vector $f \in \mathbf{H}$. Hence

$$\sup_{\gamma\in\mathbf{\Gamma}} |< f, g_\gamma >| = \lim_{p\to+\infty} \sup_{\gamma\in\mathbf{\Gamma}} |< f_{n_p}, g_\gamma >| \leq \lim_{p\to+\infty} \lambda_{n_p} = 0. \tag{90}$$

Since f has a norm 1, the inner product which each element of $\mathcal{D}$ can not be zero since $\mathcal{D}$ is complete and thus includes at least a basis of $\mathbf{H}$. This contradicts our assumption, which finishes the proof. □

Appendix C: *Dilation, Translation and Modulation Covariance*

We say that a subset Λ of $\mathbf{\Gamma}$ is admissible and associated to $f \in \mathbf{L}^2(\mathbf{R})$ if

$$\Lambda = \{\beta \in \mathbf{\Gamma} \ : \ |< f, g_\beta >| \geq \alpha \sup_{\gamma\in\mathbf{\Gamma}} |< f, g_\gamma >|\}. \tag{91}$$

Let Λ be an admissible set and $(a,b,c) \in \mathbf{R}^+ \times \mathbf{R^2}$. Let

$$\Lambda_{(a,b,c)} = \{\beta = (s,u,\xi) \in \mathbf{\Gamma} \ : \ (\frac{s}{a}, \frac{u-c}{a}, a(\xi - b)) \in \Lambda\}. \tag{92}$$

A choice function C is said to be covariant if and only if for any admissible set Λ, $C(\Lambda) = (s_0, u_0, \xi_0)$ implies that

$$C(\Lambda_{(a,b,c)}) = (\frac{s_0}{a}, \frac{u_0 - c}{a}, a(\xi_0 - b)). \tag{93}$$

If we restrict our signal space to functions that are bounded and absolutely integrable, the matching pursuit residues are also bounded and absolutely integrable. Let us give an example of covariant choice function. For any admissible set Λ, associated to a bounded and absolutely integrable function, $C(\Lambda) = (s_1, u_1, \xi_1)$, such that $s_1 = \sup\{s : \exists (u, \xi) \in \mathbf{R^2}, (\mathbf{s}, \mathbf{u}, \xi) \in \mathbf{\Lambda}\}$, $u_1 = \sup\{u : \exists \xi \in \mathbf{R}, (\mathbf{s_1}, \mathbf{u}, \xi) \in \mathbf{\Lambda}\}$, and $\xi_1 = \sup\{\xi : (s_1, u_1, \xi) \in \Lambda\}$. The following lemma proves that the index (s_1, u_1, ξ_1) is well defined and belongs to Λ

Lemma 5 *For any admissible set Λ, associated to a bounded and absolutely integrable function, $(s_1, u_1, \xi_1) \in \Lambda$.*

Proof: let Λ be an admissible index set associated to f. Since $g(t)$ is bounded and $f(t)$ is absolutely integrable, one can prove that

$$\lim_{s \to +\infty} \sup_{(u,\xi) \in \mathbf{R^2}} |< f, g_\gamma >| = 0.$$

We can thus derive that there exist a finite s_1 that is the supremum of all s such that $(s, u, \xi) \in \Lambda$. Since Λ is closed, there exists $(s_1, u, \xi) \in \Lambda$. Since

$$\lim_{|t| \to \infty} |g(t)| = 0$$

and $f(t)$ is absolutely integrable, we can prove that for $\gamma = (s_1, u, \xi)$,

$$\lim_{u \to +\infty} \sup_{\xi \in \mathbf{R}} |< f, g_\gamma >| = 0.$$

We can then derive that there exists u_1 that is the supremum of all u such that $(s_1, u, \xi) \in \Lambda$. Since Λ is closed, there exists ξ such that $(s_1, u_1, \xi) \in \Lambda$. Since

$$\lim_{|\omega| \to \infty} |\hat{g}(\omega)| = 0$$

and $\hat{f}(\omega)$ is absolutely integrable, we can prove that for $\gamma = (s_1, u_1, \xi)$,

$$\lim_{\xi \to +\infty} |< f, g_\gamma >| = 0.$$

We can finally derive that ξ_1 that is the supremum of all ξ such that $(s_1, u_1, \xi) \in \Lambda$. Since Λ is closed, $(s_1, u_1, \xi_1) \in \Lambda$. This finishes the proof of the lemma. $\square$

Let us prove the covariance of a matching pursuit based on covariant choice functions. Let us define

$$f^1(t) = \frac{d}{\sqrt{a}} f^0(\frac{t-c}{a}) e^{ibt}. \tag{94}$$

Let $\gamma^1 = (s, u, \xi)$ and $\gamma^0 = (\frac{s}{a}, \frac{u-c}{a}, a(\xi - b))$. With a change of variable, we prove that

$$< f^1, g_{\gamma^1} >= d e^{ic(b-\xi)} < f^0, g_{\gamma^0} > . \tag{95}$$

Hence $\sup_{\gamma \in \mathbf{\Gamma}} |< f^1, g_\gamma >| = d \sup_{\gamma \in \mathbf{\Gamma}} |< f^0, g_\gamma >|$. Let us define

$$\Lambda = \{\beta \in \mathbf{\Gamma} \; : \; |< f^1, g_\beta >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< f^1, g_\gamma >|\}. \tag{96}$$

Equation (95) proves that the set $\Lambda_{(a,b,c)}$ defined in (92), also satisfies

$$\Lambda_{(a,b,c)} = \{\beta \in \mathbf{\Gamma} \; : \; |< f^0, g_\beta >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< f^0, g_\gamma >|\}. \tag{97}$$

The covariance of the choice function implies that if $C(\Lambda) = \gamma_0^1 = (s_0, u_0, \xi_0)$, then $C(\Lambda_{(a,b,c)}) = \gamma_0^0 = (\frac{s_0}{a}, \frac{u_0-c}{a}, a(\xi_0 - b))$. We can thus derive that

$$Rf^1(t) = \frac{d}{\sqrt{a}} Rf^0(\frac{t-c}{a}) e^{ibt}. \tag{98}$$

Similarly, we can prove by induction that for any $n \geq 0$

$$R^n f^1(t) = \frac{d}{\sqrt{a}} R^n f^0(\frac{t-c}{a}) e^{ibt}, \tag{99}$$

and if $\gamma_n^1 = (s_n, u_n, \xi_n)$ then $\gamma_n^0 = (\frac{s_n}{a}, \frac{u_n-c}{a}, a(\xi_n - b))$, and

$$< R^n f^1, g_{\gamma_n^1} >= d e^{ic(b-\xi_n)} < R^n f^0, g_{\gamma_n^0} > . \tag{100}$$

Conversely, if the residues of $f^0(t)$ and $f^1(t)$ satisfies these equalities, equation (42) proves that equation (94) is satisfied. Hence, a matching pursuit based on covariant choice functions is covariant by dilation, translation and modulation.

Appendix D: *Proof of Theorem 2*

We denote $g_{j,p,k}(t) = g_\gamma(t)$ for $\gamma = (a^j, pa^j\Delta u, ka^{-j}\Delta\xi) \in \mathbf{\Gamma}_\alpha$. Since $\Delta u = \frac{\Delta\xi}{2\pi}$ and $\Delta u\Delta\xi < 2\pi$, Daubechies [6] proved that for the Gaussian window $g(t)$ specified by equation (59), $(g_{0,p,k}(t))_{(p,k)\in\mathbf{Z^2}}$ is a frame of $\mathbf{L}^2(\mathbf{R})$. The dual frame is given by $(\tilde{g}_{0,p,k}(t))_{(p,k)\in\mathbf{Z^2}}$, where $\tilde{g}(t) \in \mathbf{L}^2(\mathbf{R})$ and

$$\tilde{g}_{j,p,k}(t) = \frac{1}{\sqrt{a^j}}\tilde{g}(\frac{t - pa^j\Delta u}{a^j})e^{ika^{-j}\Delta\xi t}. \tag{101}$$

The dual window $\tilde{g}(t)$ has an exponential decay and its Fourier transform $\hat{\tilde{g}}(\omega)$ also has an exponential decay [5]. For any $f \in \mathbf{L}^2(\mathbf{R})$,

$$f(t) = \sum_{p=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} < f, g_{0,p,k} > \tilde{g}_{0,p,k}(t). \tag{102}$$

With a change of variable, one can derive that for any $j \in \mathbf{Z}$

$$f(t) = \sum_{p=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} < f, g_{j,p,k} > \tilde{g}_{j,p,k}(t). \tag{103}$$

Let $\gamma_0 = (s, u, \xi) \in \mathbf{\Gamma}$ and $j \in \mathbf{Z}$ be such that $a^{j-1/2} < s \leq a^{j+1/2}$

$$< f, g_{\gamma_0} >= \sum_{p=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} < f, g_{j,p,k} >< \tilde{g}_{j,p,k}, g_{\gamma_0} > . \tag{104}$$

Since $g_{j,p,k} = g_\gamma$ with $\gamma \in \mathbf{\Gamma}_\alpha$

$$|< f, g_{\gamma_0} >| \leq \sup_{\gamma\in\mathbf{\Gamma}_\alpha} |< f, g_\gamma >| \sum_{p=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} |< \tilde{g}_{j,p,k}, g_{\gamma_0} >|. \tag{105}$$

Let us now prove that there exists a finite constant K such that for all γ_0

$$S_{\gamma_0} = \sum_{p=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} |< \tilde{g}_{j,p,k}, g_{\gamma_0} >| \leq K. \tag{106}$$

For this purpose, we shall prove that there exists $D_1 > 0$, $C_1 > 0$ and $D_2 > 0$, $C_2 > 0$ such that

$$|< \tilde{g}_{j,p,k}, g_{\gamma_0} >| \leq D_1 \exp(-C_1|p\Delta u - a^{-j}u|), \tag{107}$$

and

$$|<\tilde{g}_{j,p,k}, g_{\gamma_0}>| \leq D_2 \exp(-C_2 |k\Delta\xi - a^j \xi|). \tag{108}$$

Clearly

$$|<\tilde{g}_{j,p,k}, g_{\gamma_0}>| \leq \int_{-\infty}^{+\infty} |\tilde{g}_{j,p,k}(t)||g_{\gamma_0}(t)| \; dt. \tag{109}$$

With a change of variable, we derive that

$$|<\tilde{g}_{j,p,k}, g_{\gamma_0}>| \leq \int_{-\infty}^{+\infty} \sqrt{\frac{a^j}{s}} |g(\frac{a^j t}{s})||\tilde{g}(t + a^{-j}u - p\Delta u)| \; dt. \tag{110}$$

Since $\frac{1}{\sqrt{a}} \leq \frac{a^j}{s} \leq \sqrt{a}$, and since both $g(t)$ and $\tilde{g}(t)$ have an exponential decay, one can derive that there exists two constant $C_1 > 0$ and $D_1 > 0$ that satisfy equation (107). To prove equation (108), we observe that

$$|<\tilde{g}_{j,p,k}, g_{\gamma_0}>| \leq \int_{-\infty}^{+\infty} |\hat{\tilde{g}}_{j,p,k}(\omega)||\hat{g}_{\gamma_0}(\omega)| \; d\omega. \tag{111}$$

From the expression of $\tilde{g}_{j,p,k}(\omega)$ in equation (5), with a change of variable we derive that

$$|<\tilde{g}_{j,p,k}, g_{\gamma_0}>| \leq \int_{-\infty}^{+\infty} \sqrt{\frac{s}{a^j}} |\hat{g}(\frac{s\omega}{a^j})||\hat{\tilde{g}}(\omega + a^j \xi - k\Delta\xi)| \; d\omega. \tag{112}$$

Since both $\hat{g}(\omega)$ and $\hat{\tilde{g}}(\omega)$ have an exponential decay and $\frac{1}{\sqrt{a}} \leq \frac{s}{a^j} \leq \sqrt{a}$, we can also derive that there exists two constant C_2 and D_2 that satisfy (108). From the upper bounds (107) and (108) we can show that the sum S_{γ_0} of equation (106) is bounded by a finite constant K that is independent of γ_0. From equation (105), we derive that any constant $\alpha \leq \frac{1}{K}$ satisfies the condition (62) of theorem 2.

Appendix E: *Matching Pursuit Implementation with Gabor Dictionaries*

This appendix describes the numerical implementation of a matching pursuit for a Gabor dictionary. Instructions to obtain a free copy of the software implementing this transform are available through anonymous ftp at the address cs.nyu.edu, in the file README of the directory /pub/wave.

For any $\gamma = (s, p, \frac{2\pi k}{N})$ and $\phi \in [0, 2\pi[$, real discrete time-frequency atoms are related to complex atoms by

$$g_{(\gamma,\phi)} = \frac{K_{(\gamma,\phi)}}{2}(e^{i\phi} g_\gamma + e^{-i\phi} g_{\gamma^-}). \tag{113}$$

One can derive that the normalization constant is

$$K_{(\gamma,\phi)} = \frac{\sqrt{2}}{\sqrt{1 + Real(e^{i2\phi} < g_\gamma, g_{\gamma^-} >)}}, \tag{114}$$

where $Real(z)$ is the real part of the complex number z. For any residue $R^n f$,

$$|< R^n f, g_{(\gamma,\phi)} >| = K_{(\gamma,\phi)} |Real(e^{-i\phi} < R^n f, g_\gamma >)|. \tag{115}$$

By chosing ϕ equal to the complex phase ϕ_γ of $< R^n f, g_\gamma >$, we obtain

$$|Real(e^{-i\phi_\gamma} < R^n f, g_\gamma >)| = |< R^n f, g_\gamma >|. \tag{116}$$

We search for an index $\tilde{\gamma}_n$ that maximizes $|< R^n f, g_\gamma >|$ for γ in the subset $\mathbf{\Gamma}_\alpha$ of $\mathbf{\Gamma}$. With a Newton algorithm, we then look in the neighborhood of $\tilde{\gamma}_n$ in $\mathbf{\Gamma}$ for an index $\gamma_n = (s_n, p_n, \frac{2\pi k_n}{N}) \in \mathbf{\Gamma}$, where $|< R^n f, g_\gamma >|$ reaches a local maxima. One can verify that there exists $\alpha > 0$ such that

$$|< R^n f, g_{(\gamma_n,\phi_{\gamma_n})} >| \geq \alpha \sup_{(\gamma,\phi)\in\Gamma_\alpha\times[0,2\pi[} |< R^n f, g_{(\gamma,\phi)} >|. \tag{117}$$

Since

$$R^{n+1} f = R^n f - < R^n f, g_{(\gamma_n,\phi_{\gamma_n})} > g_{(\gamma_n,\phi_{\gamma_n})}, \tag{118}$$

for the next iteration we must compute for any $\gamma \in \mathbf{\Gamma}_\alpha$

$$< R^{n+1} f, g_\gamma > = < R^n f, g_\gamma > - < R^n f, g_{(\gamma_n,\phi_{\gamma_n})} >< g_{(\gamma_n,\phi_{\gamma_n})}, g_\gamma > . \tag{119}$$

We therefore estimate

$$< g_{(\gamma_n,\phi_{\gamma_n})}, g_\gamma > = \frac{K_{(\gamma_n,\phi_{\gamma_n})}}{2} (e^{i\phi_{\gamma_n}} < g_{\gamma_n}, g_\gamma > + e^{-i\phi_{\gamma_n}} < g_{\gamma_n^-}, g_\gamma >), \tag{120}$$

To compute fast this inner product, we use an analytical formula that gives the inner product of two discrete complex Gabor signals. This formula is derived from the following lemma.

Lemma 6 *Let $f(t)$ and $h(t)$ be two continuously differentiable functions such that $f(t) = O(\frac{1}{1+t^2})$ and $h(t) = O(\frac{1}{1+t^2})$. Let f_d and h_d be the discrete signals of period N defined by*

$$f_d(n) = \sum_{m=-\infty}^{+\infty} f(n + mN), \tag{121}$$

$$h_d(n) = \sum_{m=-\infty}^{+\infty} h(n+mN). \tag{122}$$

Then

$$< f_d, h_d >= \sum_{n=1}^{N} f_d(n)\bar{h}_d(n) = \sum_{m=-\infty}^{+\infty} \sum_{q=-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(t)\bar{h}(t+mN)e^{i2q\pi t}dt. \tag{123}$$

Proof:

$$\begin{aligned} < f_d, h_d >= \sum_{n=1}^{N} f_d(n)\bar{h}_d(n) &= \sum_{n=1}^{N} \sum_{q=-\infty}^{+\infty} f(n+qN) \sum_{m=-\infty}^{+\infty} \bar{h}(n+mN) \\ &= \sum_{p=-\infty}^{+\infty} f(p) \sum_{m=-\infty}^{+\infty} \bar{h}(p+mN). \end{aligned} \tag{124}$$

Hence,

$$< f_d, h_d > \; = \; \sum_{m=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(t)\bar{h}(t+mN)\delta(t-p)dt. \tag{125}$$

Let us recall the Poisson formula:

$$\sum_{p=-\infty}^{+\infty} \delta(t-p) = \sum_{q=-\infty}^{+\infty} e^{i2\pi q t}. \tag{126}$$

Inserting this in (125) yields equation (123). This finishes the proof of the lemma. □

For $\gamma_1 = (s_1, p_1, \frac{2\pi k_1}{N})$ and $\gamma_2 = (s_2, p_2, \frac{2\pi k_2}{N})$ and $g(t)$ given by (59), one can derive from equation (123) of lemma 6 that the inner product of two discrete Gabor signals is

$$\begin{aligned} < g_{\gamma_1}, g_{\gamma_2} > \; &= \; K_{s_1} K_{s_2} \sqrt{\frac{2 s_1 s_2}{s_1^2 + s_2^2}} \exp(-i p_2 \frac{2\pi(k_2 - k_1)}{N}) \times \\ \sum_{m=-\infty}^{+\infty} \sum_{q=-\infty}^{+\infty} & \Bigg(\exp(-\pi \frac{(p_2 - p_1 + mN)^2}{s_1^2 + s_2^2}) \; \exp(-\pi \frac{(k_2 - k_1 + qN)^2}{N^2(s_1^{-2} + s_2^{-2})}) \times \\ & \exp(i \frac{s_2^2}{s_1^2 + s_2^2} \frac{2\pi}{N}(k_2 - k_1 + qN)(p_2 - p_1 + mN)) \Bigg). \end{aligned} \tag{127}$$

When g_{γ_1} or g_{γ_2} is a discrete Dirac or a discrete complex exponential, different formula must be used. If we limit the computation to a precision ϵ, for any Gabor atom g_{γ_1}, there are $O(N\sqrt{|\log\epsilon|})$ other vectors g_{γ_2} such that $< g_{\gamma_1}, g_{\gamma_2} >$ is not negligible. One can show that equation (127) requires $O(N|\log\epsilon|^{3/2})$ operations to compute the inner product of any atom g_{γ_1} with all other discrete atoms $(g_\gamma)_{\gamma\in\mathbf{\Gamma}_\alpha}$. The total numerical complexity for one matching pursuit iteration is $O(N\log N)$. By tabulating the Gaussian and complex exponential functions, each iteration requires approximatively as much CPU time as a Fast Fourier Transform on a signal of N samples. In the experiments shown in this paper, we restricted the scale s_n of the selected atoms to powers of 2, to minimize the memory required by the tabulation. However, the choice of s_n may have no such restriction, if we do not use any tabulation.

References

[1] C. K. Chui, "Wavelets: a tutorial in theory and applications", Academic Press, Inc., New York, 1992.

[2] L. Cohen "Time-frequency distributions: a review" Proceedings of the IEEE, vol. 77, No. 7, p. 941-979, July 1989.

[3] R. Coifman, Y. Meyer and V. Wickerhauser, "Size properties of wavelet-packets", *Wavelets and Their Applications*, Mary Beth Ruskai, Ed., Jones and Bartlett Publishers, Boston, p. 453-470, 1992.

[4] R. Coifman and V. Wickerhauser, "Entropy-based algorithms for best basis selection", *IEEE Trans. on Information Theory.* vol. 38, no. 2, March 1992. March, 1992.

[5] I. Daubechies, "Ten lectures on wavelets", CBMS-NSF Series in Appl. Math., SIAM, 1991.

[6] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis", *IEEE Trans. on Information Theory*, vol. 36, p. 961-1005, Sept. 1990.

[7] J. H. Friedman and W. Stuetzle, "Projection pursuit regression", *Journal of the American statistical Association*, vol. 76, p. 817-823, 1981.

[8] G. Golub and C. Van Loan "Matrix Computations", The Johns Hopkins University Press, Baltimore, Maryland.

[9] R. Gray, "Vector quantization", *IEEE Acoustic Speech and Signal Processing Magazine*, April 1984.

[10] P. J. Huber, "Projection Pursuit", *The Annals of Statistics*, vol. 13, No. 2, p. 435-475, 1985.

[11] L. K. Jones, "On a conjecture of Huber concerning the convergence of projection pursuit regression", *The Annals of Statistics*, vol. 15, No. 2, p. 880-882, 1987.

[12] S. Mallat and Z. Zhang, "Local time-frequency multilayer orthogonal transforms", *Proc. of the Workshop on the Role of Wavelets in Signal Processing Applications,* Dayton, Ohio, March 1992.

[13] Y. Meyer, "Ondelettes et Operateurs", Hermann, 1990.

[14] S. Qian and D. Chen, "Signal Representation via Adaptive Normalized Gaussian Functions," *IEEE Trans. on Signal Processing*, vol. 36, no. 1, Jan. 1994.

[15] O. Rioul and M. Vetterli, "Wavelets and signal processing", *IEEE Signal Processing Magazine*, Oct., 1991.

[16] M. Sabin and R. Gray, "Product code vector quantizer for waveform and voice coding", *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. 32. no. 3, June 1984.

[17] B. Torresani, "Wavelets associated with representations of the affine Weyl-Heisenberg group", *Journal of Mathematical Physics*, vol. 32, No. 5, p. 1273-1279, May 1991.