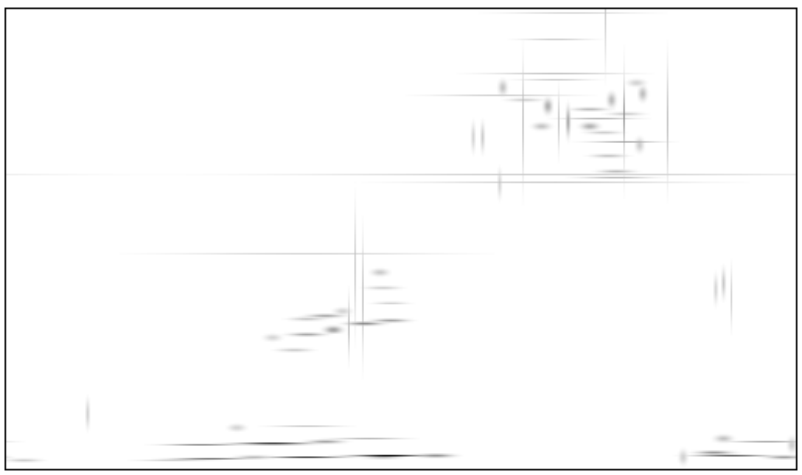

Figure 5(a): time-frequency energy distribution of the $m = 76$ coherent structures of the noisy speech signal shown in Fig.3(a).

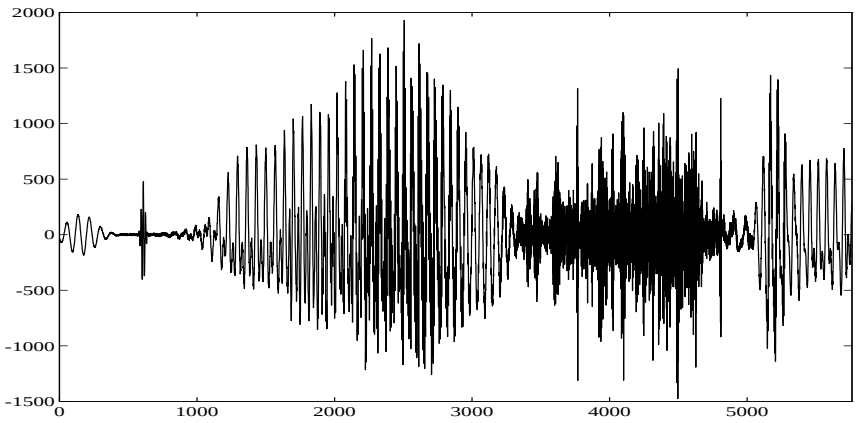

Figure 5(b): time-frequency energy distribution of the $m = 76$ coherent structures of the noisy speech signal shown in Fig. 3. (b): signal reconstructed from the 76 coherent structures shown in (a). The white noise has been removed.

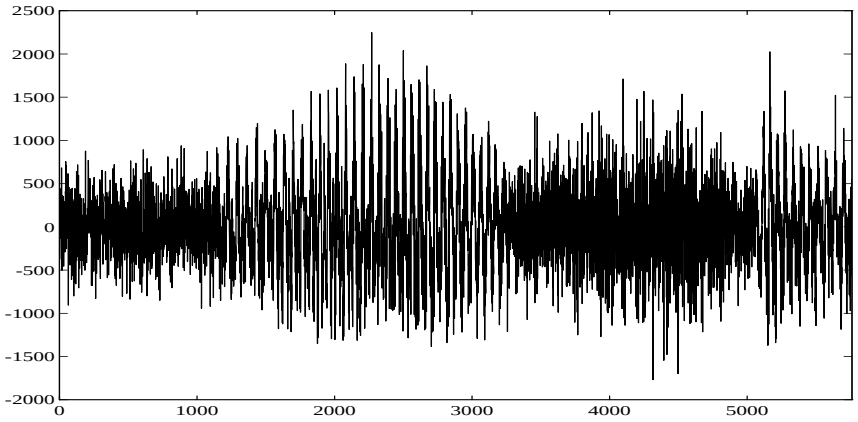

Figure 4(a): signal obtained by adding a Gaussian white noise to the speech recording shown in Fig.2(a). The signal to noise ratio is 1.5db.

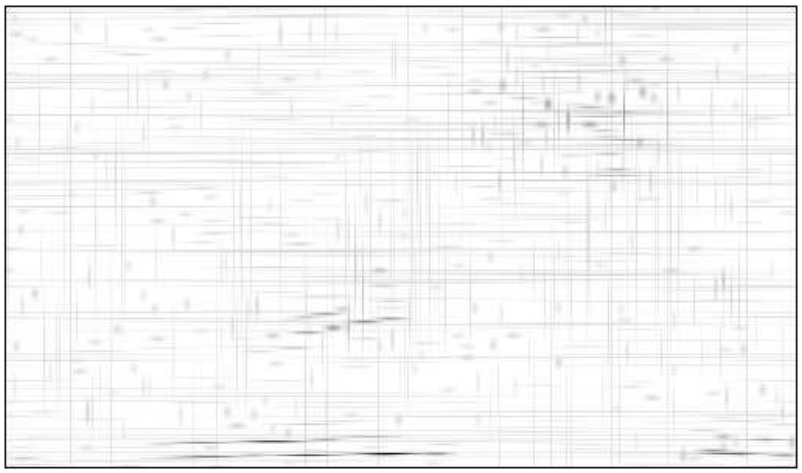

Figure 4(b): time-frequency energy distribution of the noisy speech signal. The energy distribution of the white noise is spread across the whole time-frequency plane.

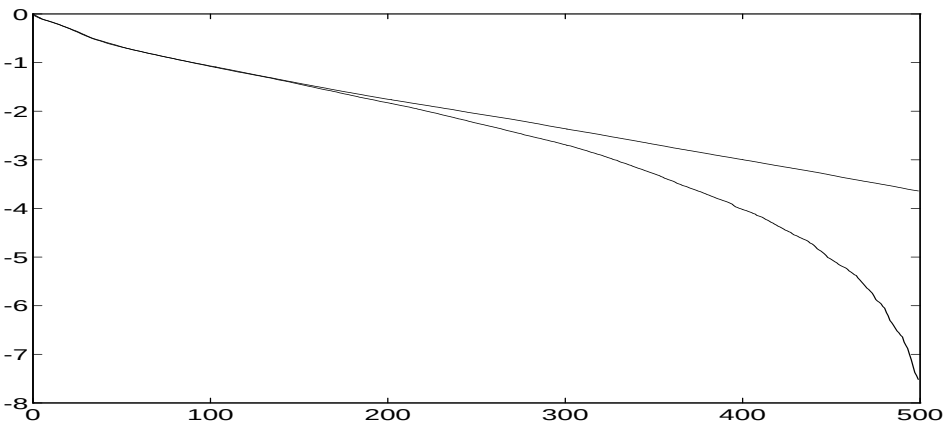

Figure 3(a): The top and bottom curves gives respectively the decay of $\log_{10} \frac{\|R^n f\|}{\|f\|}$ for a non-orthogonal and an orthogonal matching pursuit, applied to the signal in Fig. 1(a).

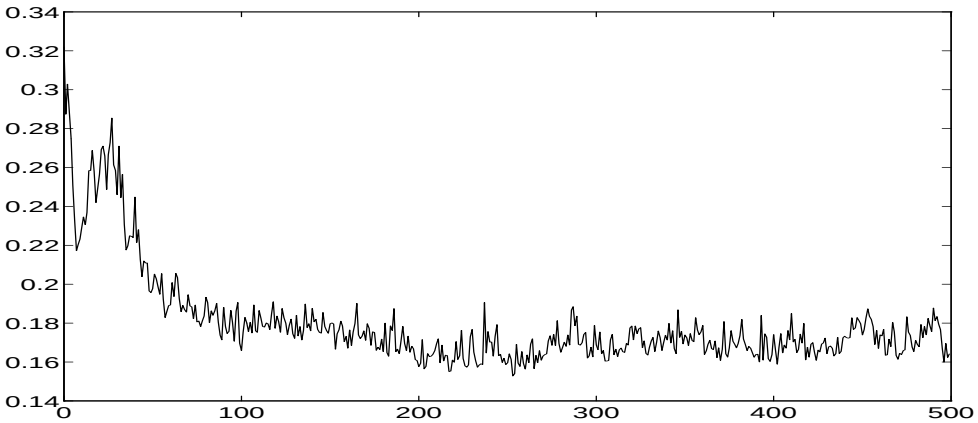

Figure 3(b): $\lambda(R^n f)$ as a function of the number of iterations n, for the signal in Fig. 1(a).

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

such high dimensional spaces. We are currently studying the problem of adapting the dictionary to specific signal properties.

References

[1] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification", *International Journal of Control*, vol. 50, No. 5, pp. 1873-1896, 1989.

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

[3] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Series in Appl. Math., SIAM, 1991.

[4] G. Davis, S. Mallat, and M. Avellaneda, "Chaos in Adaptive Approximations", Technical Report, Computer Science, NYU, April 1994.

[5] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," Journal of the American Statistical Association, Vol. 76, pp. 817-823, 1981.

[6] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Co., New York, 1979.

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

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

[9] 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.

[10] S. Mallat and Z. Zhang "Matching Pursuit with Time-Frequency Dictionaries", *IEEE Trans. on Signal Processing*, Dec. 1993.

[11] Y. C. Pati R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," *Proceedings of the* 27^{th} *Annual Asilomar Conference on Signals, Systems, and Computers*, Nov. 1993.

The ability to find the coherent structures of a signal can be used to remove noise from signals. In this approach, the noise is entirely defined by the choice of the dictionary and corresponds to the invariant measure of the matching pursuit map. For a Gabor dictionary, the dictionary noise is white and stationary, but this is not necessarily true for other dictionaries. The dictionary must be chosen so that its elements correlate as closely as possible the signal inner structures, but so that they avoid high correlations with realizations of the noise to be removed. During the matching pursuit decomposition, we test the correlation ratio (68) and stop when condition (69) is true, i.e. when we have selected all coherent structures. Fig. 4(a) shows a signal obtained by adding a Gaussian white noise to the speech recording given in Fig. 2(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. This signal contains $m = 76$ coherent structures which are displayed in the time-frequency energy distribution of Fig. 5(a). Fig. 5(b) is the signal reconstructed from these coherent time-frequency atoms. The SNR of the reconstructed signal is 6.8 dB. The white noise has been removed and the recovered signal has a good auditory quality because the main time-frequency structures of the original speech signal have been retained.

8 Conclusion

Matching pursuits provide extremely flexible signal representations since the choice of dictionaries is not limited. 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 on the unit sphere of the signal space and are thus well adapted to decomposing signals for 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 for the optimization of code books, such as LBG [7], do not converge to satisfying solutions in

[4].

An important measure of the flatness of the signal f with respect to a dictionary is the correlation ratio

$$\lambda(f) = \sup_{\gamma \in \mathbf{\Gamma}} \frac{| < f, g_\gamma > |}{\|f\|}. \tag{67}$$

The correlation ratio is the fraction of the signal energy which will be removed by one iteration of a matching pursuit when the optimality ratio $\alpha = 1$. Signals which possess structures closely resembling dictionary elements will have large values of $\lambda(f)$. As the matching pursuit proceeds, these structures are removed, and $\lambda(R^k f)$ decreases. When signal residues are as flat as realization of a dictionary noise, the convergence rate of the pursuit is near a minimum. It is often not worth continuing the matching pursuit iterations since the signal includes no structures that strongly correlate to dictionary elements. Let λ_∞ be the expected value of the correlation ratio for the dictionary noise. This constant is computed through numerical experiments over many iterations of the map. We call the coherent structures of a signal f the first m vectors $(g_{\gamma_n})_{0 \leq n < m}$ that correlate with the residue better than the average correlation ratio λ_∞. For $0 \leq n < m$ we have

$$\lambda(R^n f) > \lambda_\infty \tag{68}$$

and for m we have

$$\lambda(R^m f) \leq \lambda_\infty. \tag{69}$$

The decay rate of the energy of the residues depends only upon the correlation ratio. If we set the optimality factor $\alpha = 1$, then (11) implies that

$$\frac{\|R^{n+1} f\|}{\|R^n f\|} = (1 - \lambda^2(R^n f))^{\frac{1}{2}}. \tag{70}$$

Fig. 3(b) displays the correlation ratio of the residues $\lambda^2(R^n f)$ for the signal in Fig. 1(a). After 180 iterations, $\lambda(R^n f)$ fluctuates around the mean λ_∞. It is at that same point that the decay rate of the orthogonal pursuit residues, shown in Fig. 3(a), becomes faster than the decay rate of the non-orthogonal pursuit. This indicates that the orthogonal pursuit converges significantly faster than the non-orthogonal pursuit only when the residues are already close to the realization of a dictionary noise. For information processing applications, it is often not useful to continue the decomposition beyond this point, so for such applications the orthogonal pursuit does not offer much advantage over the non-orthogonal pursuit.

to a 1-dimensional map, which we prove is topologically equivalent to a left-shift map. Thus, renormalized matching pursuit in this case is a chaotic map whose properties are completely understood.

Numerical experiments with several types of dictionaries, including Gabor dictionaries, provide evidence that the renormalized matching pursuit map possesses an invariant measure and that it is mixing. The mixing property means that if we start with a collection of test signals, then after sufficient iterations of renormalized matching pursuit, the density of the residues in the signal space will be close to the invariant density function. To express this result in another way, consider a random process which yields signals with a probability measure on the signal space given by our invariant measure. Then the mixing property implies that after sufficient iterations of the map, the residues will look like realizations of this process.

The renormalized matching pursuit map continually sets to zero the largest of the dictionary components of the current residue, and redistributes the energy from this removed component over the remainder of the residue. We expect this system to be near equilibrium when all dictionary components of the residue have roughly the same magnitude–when this is so, the operations of setting the largest component to zero and of redistributing the energy have little effect. The residues converge to realizations of a random process for which the distribution of the dictionary components is flat. This invariant process can be interpreted as a generic noise with respect to our dictionary, which we call dictionary noise.

For a Gabor dictionary we observe that after several iterations the residues have statistical properties that are close to realizations of a white stationary process. To better understand this phenomenon, we first note that the Gabor dictionary is invariant under translation and frequency modulation, which means that for any $g_\gamma(t) \in \mathcal{D}$, and $(u, \xi) \in \mathbf{R^2}$, there exists a $\phi \in \mathbf{R}$ such that $e^{i\phi} e^{i\xi t} g_\gamma(t-u) \in \mathcal{D}$. One can prove [4] that for a translation and modulation invariant dictionary, if there exists an invariant measure, then this measure is invariant with respect to operators that translate signals in time or frequency. This invariant measure therefore corresponds to a white stationary process. A detailed analysis of the invariant measure was performed performed for a dictionary composed of a discrete Dirac basis plus a discrete Fourier basis. This dictionary is clearly invariant by translations and frequency modulations. We constructed a stochastic differential equation model of the evolution of the renormalized residues and solved the corresponding Fokker-Planck equation to compute the properties of the invariant measure. We obtained excellent agreement with numerical data

residue

$$\tilde{R}^n f = \frac{R^n f}{\|R^n f\|}. \tag{60}$$

Renormalized matching pursuit is the map defined by

$$M(\tilde{R}^n f) = \tilde{R}^{n+1} f. \tag{61}$$

Since

$$\|R^{n+1} f\|^2 = \|R^n f\|^2 - |< R^n f, g_{\gamma_n} >|^2, \tag{62}$$

we obtain that

$$M(\tilde{R}^n f) = \frac{\tilde{R}^n f - < \tilde{R}^n f, g_{\gamma_n} > g_{\gamma_n}}{\sqrt{1 - | < \tilde{R}^n f, g_{\gamma_n} > |^2}}. \tag{63}$$

At each iteration the renormalized matching pursuit map removes the largest dictionary component of the residue and renormalizes it. This action is much like that of a left-shift operator acting on a base-N decimal number: the shift operator removes the most significant (leftmost) digit of the expansion and then "renormalizes" the expansion by multiplying by N. Let Σ_N be the set of all base N decimals. The left-shift map $L_N : \Sigma_N \to \Sigma_N$ is formally defined by

$$L_N(0.s_1 s_2 s_3 \ldots) = 0.s_2 s_3 s_4 \ldots. \tag{64}$$

where $0.s_1 s_2 \ldots$ is the base-N decimal $\sum_{k=1}^{\infty} \frac{s_k}{N^k}$. The left shift map is known to be a chaotic map, which suggests that renormalized matching pursuits share this property.

Additional evidence that renormalized matching pursuits are chaotic can be found in the fact that the map has "sensitive dependence" on the initial signal f when f is close to a dictionary element. Consider two signals f_1 and f_2 defined by

$$f_1 = (1 - \epsilon)g + \epsilon h_1 \tag{65}$$

and

$$f_2 = (1 - \epsilon)g + \epsilon h_2 \tag{66}$$

where g is the closest dictionary element to f_1 and f_2, $\|h_1\| = \|h_2\| = 1$, and $< h_1, g >=< h_2, g >= 0$. Then $\|f_1 - f_2\| = \epsilon \|h_1 - h_2\|$ can be made arbitrarily small, while $\|\tilde{R} f_1 - \tilde{R} f_2\| = \|h_1 - h_2\|$ is of order 1. The map thus separates points near dictionary elements.

We examined the renormalized matching pursuit map for a dictionary in 3 dimensions [4]. Through symmetry operations this map can be reduced

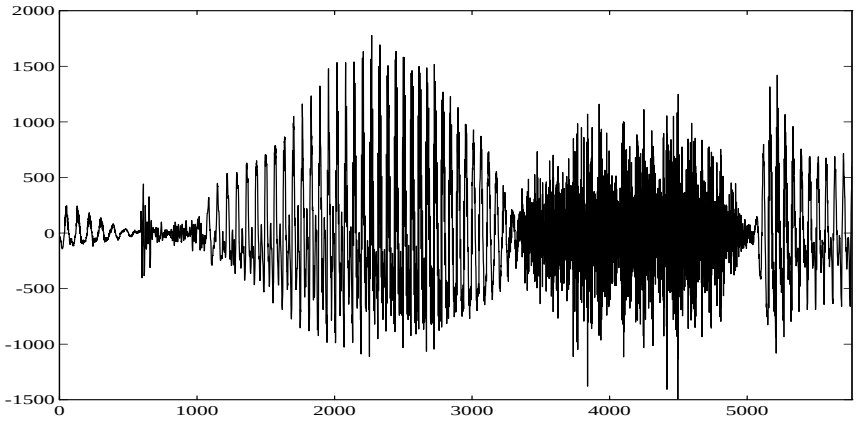

Figure 2(a): speech recording of the word "greasy", sampled at 16 kHz.

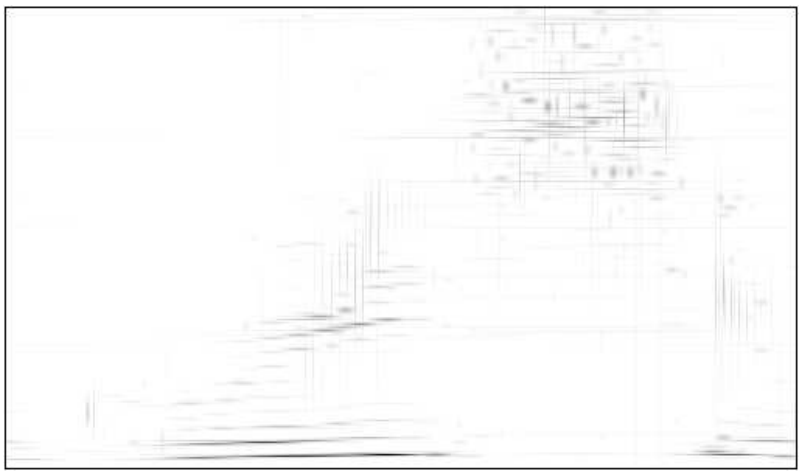

Figure 2(b): time-frequency energy distribution of the speech recording shown in (a). We see the low-frequency component of the "g", the quick burst transition to the "ea" and the harmonics of the "ea". The "s" has energy spread over high frequencies.

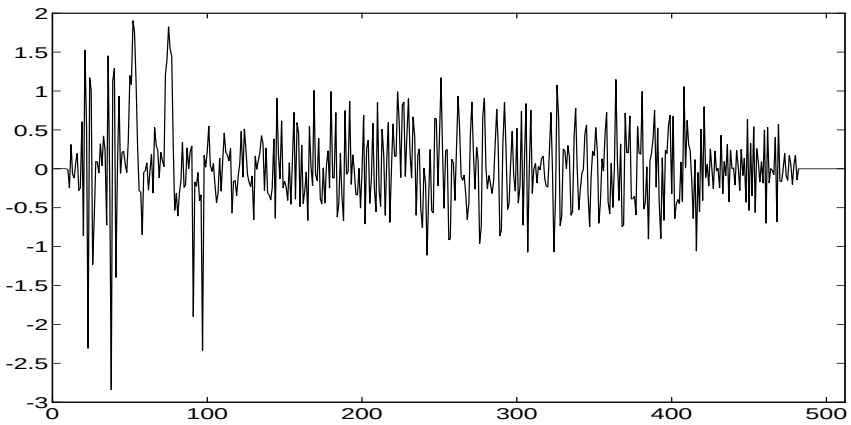

Figure 1(a): Signal of 512 samples built by adding chirps, truncated sinusoidal waves and waveforms of different time-frequency localizations.

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 $Ef(t,\omega)$.

a high-frequency interval. Most of the signal energy is characterized by a 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.

Let us now compare the decay rate of the residues for non-orthogonal matching pursuits versus orthogonal matching pursuits. The top curve in Fig. 3(a) gives the decay of $\log_{10} \frac{\|R^n f\|}{\|f\|}$ as a function of the number of iterations n, for a non-orthogonal matching pursuit. For $n \geq 130$, the decay rate is almost constant. This confirms the exponential decay proved by Theorem 2. The bottom curve in Fig. 3(a) gives the decay of $\log_{10} \frac{\|R^n f\|}{\|f\|}$ as a function of the number of iterations n, for an orthogonal matching pursuit. For the first 180 iterations we cannot distinguish the two curves, which means that the decay of the residues for the two algorithms is nearly identical. This indicates that the atoms selected by the non-orthogonal pursuit are nearly orthogonal. After this point, we see that the residues of the orthogonal pursuit decay much faster and for $n = N = 512$, the number of samples of the signal, the residue is zero. The next section analyzes more precisely this divergence of the convergence rates of the orthogonal and non-orthogonal pursuits. The point of separation of the rates corresponds to a stage at which the residues have no more "coherent" structures. At this stage, the residue $R^n f$ has statistical properties that are very close to a the realization of a stationary white noise. For many information processing applications, we stop the decomposition when these coherent structures have disappeared, so for such applications the orthogonal pursuit does not offer much advantage.

7 Chaos in Matching Pursuit and Noise Removal

A non-orthogonal matching pursuit can require an infinite number of iterations to converge, even in finite dimensions. We know already that the norms of these residues converge to zero; we now examine in more detail the asymptotic behavior of the residues. Numerical experiments suggest that matching pursuits are chaotic maps, and we can prove this for a particular dictionary. Experiments also show that these matching pursuits possess invariant measures, and we use this property to develop a noise removal algorithm.

To study the asymptotic properties of the residues, we first renormalize them to prevent their convergence to zero. We define the renormalized

The Wigner distribution also satisfies

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Wg(t,\omega)dt\ d\omega = \|g\|^2 = 1, \tag{56}$$

so the energy conservation equation (18) implies

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Ef(t,\omega)dt\ d\omega = \|f\|^2. \tag{57}$$

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.

If $g(t)$ is the Gaussian window

$$g(t) = 2^{1/4}e^{-\pi t^2}, \tag{58}$$

then

$$Wg(t,\omega) = 2e^{-2\pi(t^2+(\frac{\omega}{2\pi})^2)}, \tag{59}$$

so $Ef(t,\omega)$ remains positive. 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).

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 functions 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 $\omega \geq 0$. 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 other, with a localized time-frequency waveform on 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.

Fig. 2(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. 2(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

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

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s_n}} g(\frac{t-u_n}{s_n}) e^{i\xi_n t}. \tag{50}$$

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

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

$$W[f,h](t,\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t+\frac{\tau}{2}) \overline{h}(t-\frac{\tau}{2}) e^{-i\omega\tau} d\tau. \tag{51}$$

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 (49) of $f(t)$ that

$$\begin{aligned} Wf(t,\omega) &= \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g_{\gamma_n}(t,\omega) \qquad (52) \\ &+ \sum_{n=0}^{+\infty} \sum_{m=0, m\neq n}^{+\infty} < R^n f, g_{\gamma_n} > \overline{< R^m f, g_{\gamma_m} >} W[g_{\gamma_n}, g_{\gamma_m}](t,\omega). \end{aligned}$$

The double sum corresponds to the cross terms of the Wigner distribution. It contains 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 therefore keep only the first sum and define

$$Ef(t,\omega) = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g_{\gamma_n}(t,\omega). \tag{53}$$

A similar decomposition algorithm over time-frequency atoms was derived independently by Qian and Chen [12] 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 (50) of a time-frequency atom, we derive that for $\gamma = (s, \xi, u)$

$$W g_\gamma(t,\omega) = W g(\frac{t-u}{s}, s(\omega-\xi)), \tag{54}$$

and hence

$$Ef(t,\omega) = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g(\frac{t-u_n}{s_n}, s_n(\omega-\xi_n)). \tag{55}$$

If the inner product of any two elements in $\mathcal{D}$ is calculated in $O(I)$ operations, the $(b_{k,n})_{0 \leq k \leq n}$ are obtained in $O(nI + n^2)$ operations. We then compute the inner product of the new residue $R^{n+1}f$ with any $g_\gamma \in \mathcal{D}$, from the orthogonal updating formula (31)

$$< R^{n+1}f, g_\gamma > = < R^n f, g_\gamma > - < R^n f, g_{\gamma_n} > < u_n, g_\gamma > . \qquad (47)$$

Since

$$< u_n, g_\gamma > = \sum_{k=0}^{n} b_{k,n} < g_{\gamma_k}, g_\gamma >, \qquad (48)$$

computing $(< R^{n+1}f, g_\gamma >)_{\gamma \in \mathbf{\Gamma}}$ requires $O(nIZ)$ operations. The total number of operations to compute P orthogonal matching pursuit iterations is therefore $O(P^3 + P^2 IZ)$ operations. For a dictionary of discrete Gabor signals with signals of size N, since $I = 1$, $Z = N$, and $P \leq N$, the number of operations is $O(NP^2)$. It also requires $O(N \log N + P^2)$ memory to store the inner products $(< R^n f, g_\gamma >)_{\gamma \in \mathbf{\Gamma}}$ and the expansion coefficients $(b_{k,p})_{0 \leq k,p \leq n}$.

For P iterations, the non-orthogonal matching pursuit algorithm is P times faster than the orthogonal one. When P is large, which is the case in many signal processing applications, the orthogonal pursuit algorithm is much slower than the non-orthogonal one, and requires much more memory. When P remains small, the orthogonal pursuit is more advantageous because it converges faster.

6 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 the sum of complex time-frequency atoms that best match its residues. This section describes the properties of this particular matching pursuit decomposition. We derive a new type of time-frequency energy distribution by summing the Wigner distributions of the time-frequency atoms.

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

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \qquad (49)$$

so that the first step is to find g_{γ_n} such that

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

If all the inner products are stored in an open hash table, finding this supremum requires on average $O(1)$ operations. Otherwise, one needs to search across the whole set of inner products.

For a non-orthogonal matching pursuit, once the vector g_{γ_n} is selected, we compute the inner product of the new residue $R^{n+1} f$ with all $g_\gamma \in \mathcal{D}$, with an updating formula derived from equation (10)

$$< R^{n+1} f, g_\gamma > = < R^n f, g_\gamma > - < R^n f, g_{\gamma_n} > < g_{\gamma_n}, g_\gamma > . \qquad (43)$$

Since we previously stored $< R^n f, g_\gamma >$ and $< R^n f, g_{\gamma_n} >$, this update requires only the computation of $< g_{\gamma_n}, g_\gamma >$. Dictionaries are generally built so that few of these inner products are non-zero, and non-zero inner products are computed with a small number of operations. Let us suppose that such inner product computations are done in $O(I)$ operations and that there are $O(Z)$ non-zero inner products for any g_{γ_n}. Computing $\{< R^{n+1} f, g_\gamma >\}_{\gamma \in \mathbf{\Gamma}}$ thus requires $O(IZ)$ operations. Hence, the total numerical complexity of computing P matching pursuit iterations is $O(PIZ)$. An efficient implementation of a discrete Gabor dictionary [10] has $I = 1$ and $Z = N$, so that P iterations require $O(NP)$ operations. Since the dictionary has $O(N \log N)$ vectors, the total memory needed for the algorithm is $O(N \log N)$.

For an orthogonal matching pursuit algorithm, once the vector g_{γ_n} is selected, we must compute the orthogonal vector u_n with the Gram-Schmidt formula (28). We suppose that for $p < n$, we have already computed the expansion coefficient of each u_p in $(g_{\gamma_k})_{0 \leq k \leq p}$,

$$u_p = \sum_{k=0}^{p} b_{k,p} g_{\gamma_k}. \qquad (44)$$

From

$$u_n = g_{\gamma_n} - \sum_{p=0}^{n-1} \frac{< g_{\gamma_n}, u_p >}{||u_p||^2} u_p, \qquad (45)$$

we can compute the expansion

$$u_n = \sum_{k=0}^{n} b_{k,n} g_{\gamma_k}. \qquad (46)$$

The coefficients $b_{p,n}$ can be calculated while computing the orthogonal matching pursuit, as explained in section 5. Inserting expression (39) into (36) yields

$$f = \sum_{0 \leq n < M} \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2} \sum_{p=0}^{n} b_{p,n} g_{\gamma_p}. \tag{40}$$

One could naively try to rearrange the terms of this double summation to obtain

$$f = \sum_{0 \leq p < M} g_{\gamma_p} \sum_{p \leq n < M} b_{p,n} \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2}. \tag{41}$$

However, when $M = +\infty$ the infinite sum over n that defines each coefficient β_p may not converge. Such a situation arises when the family $(g_{\gamma_n})_{0 \leq n < M}$ is not a Riesz basis of the closed space $\mathbf{V}_M$ that it generates. For such a case, we cannot obtain an expansion of the form (38) from the orthogonal matching pursuit. If the signal space $\mathbf{H}$ has a finite dimension N, then M is finite, so we can always invert the two sums of (40) to obtain (41). The basis $(g_{\gamma_n})_{0 \leq n < M}$ may, however, be very badly conditioned in which case we can have numerical instabilities:

$$\sum_{0 \leq n < M} |\beta_n|^2 >> ||f||^2. \tag{42}$$

The residues of orthogonal matching pursuits in general decrease faster than the residues of non-orthogonal matching pursuits. However, this orthogonal procedure can yield unstable expansions and requires much more numerical computation because of the Gram-Schmidt orthogonalization. The implementation and computational complexity of these two algorithms is compared in the next section.

5 Numerical Implementations of Matching Pursuits

We describe fast implementations of non-orthogonal and orthogonal matching pursuits in finite dimensional spaces, and compare their performance. Software implementing matching pursuits for time-frequency dictionaries is available through anonymous ftp at the address cs.nyu.edu . Instructions are in the file README of the directory /pub/wave/software.

At stage n of the pursuit, we suppose that the inner products $(< R^n f, g_\gamma >)_{\gamma \in \mathbf{\Gamma}}$ have already been computed. We choose the optimality factor $\alpha = 1$,

An orthogonal pursuit guarantees that the selected vectors $(g_{\gamma_n})_{0 \leq n \leq m}$ are linearly independent, and computes the best possible approximation of f from these vectors. Since $R^0 f = f$, we derive from equations (31) and (32) that for any $m > 0$

$$f = \sum_{0 \leq n < m} \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2} u_n + R^m f, \tag{33}$$

and

$$\|f\|^2 = \sum_{0 \leq n < m} \frac{| < R^n f, g_{\gamma_n} > |^2}{||u_n||^2} + \|R^m f\|^2. \tag{34}$$

The derivations are similar to those for equations (13) and (14). The next theorem is similar to Theorem 2 and guarantees the convergence of the orthogonal pursuit [4].

Theorem 3 *Let $f \in \mathbf{H}$. Let N be the dimension of $\mathbf{H}$ (N may be infinite). The orthogonal matching pursuit converges in $M \leq N$ iterations (M may be infinite if N is infinite). The residue $\mathbf{R^n f}$ defined inductively by equation (31) satisfies*

$$\lim_{n \to M} \|R^n f\| = 0, \tag{35}$$

$$f = \sum_{0 \leq n < M} \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2} u_n, \tag{36}$$

and

$$\|f\|^2 = \sum_{0 \leq n < M} \frac{| < R^n f, g_{\gamma_n} > |^2}{||u_n||^2}. \tag{37}$$

If $\mathbf{H}$ is of finite dimension, the orthogonal pursuit converges within a finite number of iterations.

Our primary objective is not to expand f over $(u_n)_{0 \leq n < M}$ but rather over $(g_{\gamma_n})_{0 \leq n < M}$. We want coefficients $(\beta_n)_{0 \leq n < M}$ such that

$$f = \sum_{0 \leq n < M} \beta_n g_{\gamma_n}. \tag{38}$$

Since $u_n \in \mathbf{V}_n$ and $(g_{\gamma_p})_{0 \leq p \leq n}$ is a basis of $\mathbf{V}_n$, we can decompose u_n into

$$u_n = \sum_{p=0}^{n} b_{p,n} g_{\gamma_p}. \tag{39}$$

Let us explain by induction how to compute the orthogonal residue $R^{n+1}f$ from $R^n f$. We suppose that we have already selected n vectors $(g_{\gamma_p})_{0\leq p<n}$ that are linearly independent and that we computed the corresponding Gram-Schmidt orthogonal basis $(u_p)_{0\leq p<n}$. Both $(g_{\gamma_p})_{0\leq p<n}$ and $(u_p)_{0\leq p<n}$ are bases of the space $\mathbf{V}_n$ and

$$R^n f = f - \mathbf{P}_{\mathbf{V}_n} f. \tag{26}$$

We choose a vector $g_{\gamma_n} \in \mathcal{D}$ which satisfies

$$|< R^n f, g_{\gamma_n} >| \geq \alpha \sup_{\gamma\in\mathbf{\Gamma}} |< R^n f, g_\gamma >|. \tag{27}$$

If $< R^n f, g_{\gamma_n} > \neq 0$, then the vector g_{γ_n} cannot belong to the space $\mathbf{V}_n$ since $R^n f$ is orthogonal to $\mathbf{V}_n$. Hence the vectors $(g_{\gamma_p})_{0\leq p\leq n}$ are linearly independent . The next vector u_n of the Gram-Schmidt basis is obtained by subtracting from g_{γ_n} its projection on the space $\mathbf{V}_n$

$$u_n = g_{\gamma_n} - \sum_{p=0}^{n-1} \frac{< g_{\gamma_n}, u_p >}{||u_p||^2} u_p. \tag{28}$$

The family $(u_p)_{0\leq p\leq n}$ is an orthogonal basis of $\mathbf{V}_{n+1}$. The residue $R^{n+1}f$ is defined by

$$R^{n+1} f = f - \mathbf{P}_{\mathbf{V}_{n+1}} f = f - \sum_{p=0}^{n} \frac{< f, u_p >}{||u_p||^2} u_p. \tag{29}$$

This can also be rewritten

$$R^{n+1} f = R^n f - \frac{< R^n f, u_n >}{||u_n||^2} u_n. \tag{30}$$

Since $R^n f$ is orthogonal to the vectors $(g_{\gamma_p})_{0\leq p<n}$, equation (28) implies that $< R^n f, u_n >=< R^n f, g_{\gamma_n} >$ and thus

$$R^{n+1} f = R^n f - \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2} u_n. \tag{31}$$

This equation is similar to the residue updating equation (10) of a matching pursuit, but instead of subtracting a vector in the direction of g_{γ_n}, we subtract a component in a direction orthogonal to all vectors previously selected. Since $R^{n+1}f$ and u_n are orthogonal,

$$||R^{n+1} f||^2 = ||R^n f||^2 - \frac{|< R^n f, g_{\gamma_n} >|^2}{||u_n||^2}. \tag{32}$$

is called a back-projection. Instead of storing the inner products $< R^n f, g_{\gamma_n} >$ in the structure book, we store $< R^n f, g_{\gamma_n} > + x_n$ in order to recover $\mathbf{P}_{\mathbf{V}_m} f$ with (20). In this case, the approximation error

$$\mathbf{P}_{\mathbf{W}_m} f = f - \mathbf{P}_{\mathbf{V}_m} f \tag{22}$$

is the orthogonal projection of f on the space $\mathbf{W}_m$, the orthogonal complement of $\mathbf{V}_m$ in $\mathbf{H}$. The calculation of the coefficients $(x_n)_{0 \leq n < m}$ requires that we solve the following linear system. For any g_{γ_k}, $0 \leq k < m$,

$$< \mathbf{P}_{\mathbf{V}_m} R^m f, g_{\gamma_k} > = < R^m f, g_{\gamma_k} > = \sum_{n=0}^{m-1} x_n < g_{\gamma_n}, g_{\gamma_k} > . \tag{23}$$

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 (23) can be written $Y = GX$. A solution of this system is computed efficiently with a conjugate gradient algorithm [10]. If $\mathbf{H}$ is of finite dimension N, there are many classes of dictionaries for which any collection of N distinct dictionary vectors is a basis of $\mathbf{H}$. This is the case for the Gabor dictionary used for time-frequency decompositions. Hence, after selecting N different vectors with a matching pursuit, the back-projection reduces to 0 the remaining residue.

Instead of recovering the orthogonal projection $\mathbf{P}_{\mathbf{V}_m} f$ at the end of the matching pursuit, one can modify the pursuit algorithm by computing this orthogonal projection when selecting each new vector of the dictionary. It is more efficient to orthogonalize the family of selected vectors with a Gram-Schmidt algorithm than to perform a back projection. This type of algorithm was first introduced for control applications [1] and also also studied independently from this work by Pati et al. [11]. It has the advantage of providing better approximations than the matching pursuit algorithm, but it requires much more computation and can introduce numerical instabilities into the expansions. We describe by induction this orthogonal pursuit.

For $n = 0$, we set $R^0 f = f$. Like in a matching pursuit, we define a supremum factor α, with $0 < \alpha \leq 1$, and choose $g_{\gamma_0} \in \mathcal{D}$ which satisfies

$$| < f, g_{\gamma_0} > | \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} | < f, g_\gamma > |. \tag{24}$$

The space $\mathbf{V}_1$ is generated by the single vector g_{γ_0}. The first vector u_0 of the Gram-Schmidt basis is g_{γ_0}. The next residue is defined by

$$Rf = f - \mathbf{P}_{\mathbf{V}_1} f = f - < f, g_{\gamma_0} > g_{\gamma_0}. \tag{25}$$

Theorem 2 *Let $f \in \mathbf{H}$. The residue $R^m f$ defined by the induction equation (10) satisfies*

$$\lim_{m \to +\infty} \|R^m f\| = 0. \tag{16}$$

Hence

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{17}$$

and

$$\|f\|^2 = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2. \tag{18}$$

When $\mathbf{H}$ is of finite dimension, $||R^m f||$ decays exponentially to zero.

This theorem proves that any vector f is characterized by the double sequence $(< R^n f, g_{\gamma_n} >, \gamma_n)_{n \in \mathbf{N}}$, called a structure book, which specifies the expansion coefficients and the index of each chosen vector within the dictionary.

4 Back-projection and Orthogonal Pursuit

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

$$f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^m f. \tag{19}$$

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 (19) recovers an approximation of f with error $R^m f$. However, this sum is not the linear expansion of the vectors $(g_{\gamma_n})_{0 \leq n < m}$ which best approximates f. 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 onto $\mathbf{V}_m$. 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 (19) that

$$\mathbf{P}_{\mathbf{V}_m} f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + \mathbf{P}_{\mathbf{V}_m} R^m f. \tag{20}$$

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

$$\mathbf{P}_{\mathbf{V}_m} R^m f = \sum_{n=0}^{m-1} x_n g_{\gamma_n}, \tag{21}$$

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 $R^n f$

$$| < R^n f, g_{\gamma_n} > | \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} | < R^n f, g_\gamma > |. \tag{9}$$

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

$$R^n f = < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^{n+1} f, \tag{10}$$

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

$$\|R^n f\|^2 = | < R^n f, g_{\gamma_n} > |^2 + \|R^{n+1} f\|^2. \tag{11}$$

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

$$f = \sum_{n=0}^{m-1} \left(R^n f - R^{n+1} f \right) + R^m f. \tag{12}$$

Equation (10) yields

$$f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^m f. \tag{13}$$

Similarly, we write $\|f\|^2$ as a telescoping sum

$$\|f\|^2 = \sum_{n=0}^{m-1} \left(\|R^n f\|^2 - \|R^{n+1} f\|^2 \right) + \|R^m f\|^2 \tag{14}$$

which we combine with (11) to obtain an energy conservation equation

$$\|f\|^2 = \sum_{n=0}^{m-1} | < R^n f, g_{\gamma_n} > |^2 + \|R^m f\|^2. \tag{15}$$

Thus, the original vector f is decomposed into a sum of dictionary elements which are chosen to best match its residues. Although this decomposition is non-linear, we maintain an energy conservation as though it were a linear, orthogonal decomposition. An important issue is to understand the behavior of the residue $R^m f$ when m increases. By transposing a result proved by Jones [9] for projection pursuit algorithms [5], one can prove [10] that the matching pursuit algorithm converges, even in infinite dimensional spaces.

This instability can be avoided by imposing the constraint that the approximation must satisfy

$$\sum_{0 \leq n < M} |\beta_n|^2 \leq K ||f||^2, \tag{5}$$

for some fixed $K \geq 1$. One can prove that there always exist such ϵ-approximations.

When ϵ is modified, the dictionary vectors that appear in the optimal approximation can change completely, which prevents us from computing optimal approximations using progressive refinement. This instability of the optimal approximations, together with the computational intractibility of computing them, leads us to use greedy sub-optimal algorithms that progressively refine the functional approximation by choosing appropriate dictionary vectors.

3 Matching Pursuit

Let $f \in \mathbf{H}$. We want to compute a linear expansion of f over a set of vectors selected from $\mathcal{D}$ which best matches the inner structures of f. A matching pursuit is a greedy algorithm which successively approximates f with orthogonal projections onto elements of $\mathcal{D}$. Let $g_{\gamma_0} \in \mathcal{D}$. The vector f can be decomposed into

$$f = < f, g_{\gamma_0} > g_{\gamma_0} + Rf, \tag{6}$$

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

$$\|f\|^2 = |< f, g_{\gamma_0} >|^2 + \|Rf\|^2. \tag{7}$$

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

$$|< f, g_{\gamma_0} >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< f, g_\gamma >|, \tag{8}$$

where $\alpha \in (0, 1]$ is an optimality factor.

We sub-decompose the residue Rf by projecting it onto the vector of $\mathcal{D}$ that best matches Rf, as was done for f. This projection of Rf generates a second residue, $R^2 f$, which we again decompose to obtain a third residue, and so on.

best adapted to expand a given $f(t)$. A first issue is to define a notion of "optimal" approximation within a given dictionary.

Definition 1 *Let $\epsilon > 0$. An ϵ-approximation of $f \in \mathbf{H}$ is a linear expansion of dictionary vectors*

$$\tilde{f} = \sum_{0 \leq n < M} \beta_n g_{\gamma_n},$$

for which

$$\|\tilde{f} - f\| < \epsilon. \tag{3}$$

$\tilde{f}$ is an optimal ϵ-approximation if M is the minimum integer for which (3) is satisfied.

To determine the computational complexity of obtaining these optimal solutions, we consider a space $\mathbf{H}$ of finite dimension and dictionaries $\mathcal{D}$ of size $O(N^k)$ for some $k > 0$. We encode all quantities with $O(N^p)$ bits for some positive p. We say that an algorithm solves the optimal ϵ-approximation problem if, for any given $f \in \mathbf{H}$, any $\mathcal{D}$ of size $O(N^k)$, and any $\epsilon > 0$, we can find an optimal ϵ-approximation for f. The following result shows that this problem is computationally intractable.

Theorem 1 *The optimal ϵ-approximation problem is NP-hard.*

In large dimensional spaces, it is therefore not feasible to compute optimal ϵ-approximations. Theorem 1 is proved by showing that any instance of the Exact Cover by 3-Sets problem [6] can be transformed in polynomial time into an optimal ϵ-approximation problem. Thus, an algorithm which solves the ϵ-approximation problem can solve the NP-complete Exact Cover by 3-Sets problem. The intractability of the approximation problem is due to the coupling between terms in the function expansions when the dictionary elements are not orthogonal. We map the overlapping sets in the Exact Cover by 3-Sets problem to a set of coupled, non-orthogonal dictionary vectors. When the dictionary vectors are orthogonal, this coupling-induced complexity vanishes. We can solve the problem in $O(N \log N)$ by sorting the inner products $\{| < f, g_\gamma > |^2\}_{\gamma \in \mathbf{\Gamma}}$ and finding the minimum M for which the sum of the M largest terms satisfies $\|f\|^2 - \sum_{i=1}^{M} | < f, g_{\gamma_i} > |^2 < \epsilon$.

In addition to the computational complexity, an important issue is that this optimization criteria can lead to numerically unstable expansions. We can construct examples where the $\mathbf{l}^2$ norm of the expansion coefficients is arbitrarily larger than $||f||^2$, i.e.

$$\sum_{0 \leq n < M} |\beta_n|^2 >> ||f||^2. \tag{4}$$

From this asymptotic behavior, we derive an algorithm that selects coherent signal structures from noisy signals. We describe an application to noise removal in speech recordings.

2 Optimal Adaptive Approximations in Dictionaries

We expand functions from a Hilbert space $\mathbf{H}$ into linear combinations of vectors from a large collection $\mathcal{D} = (g_\gamma(t))_{\gamma \in \mathbf{\Gamma}}$, with $||g_\gamma|| = 1$, called a dictionary. The dictionary is constructed so that linear combinations of dictionary vectors are dense in $\mathbf{H}$. The smallest possible dictionary is a basis of $\mathbf{H}$, but in practice a dictionary is a very redundant set. The redundancy gives an increased degree of freedom in constructing function expansions, and this freedom is used in order to obtain improved convergence properties of the expansions.

For signal processing applications, we study the properties of dictionaries composed of waveforms that are well concentrated both in time and frequency. Our signal space is $\mathbf{L}^2(\mathbf{R})$ and we construct such a dictionary by scaling, translating and modulating a single window function $g(t) \in \mathbf{L}^2(\mathbf{R})$. We suppose that $g(t)$ is real and centered at 0. We also impose that $\|g\| = 1$, that the integral of $g(t)$ is non-zero and $g(0) \neq 0$. For any scale $s > 0$, frequency modulation ξ and translation u, we denote $\gamma = (s, u, \xi)$ and define

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

The index γ is an element of the set $\mathbf{\Gamma} = \mathbf{R}^+ \times \mathbf{R^2}$. The factor $\frac{1}{\sqrt{s}}$ normalizes to 1 the norm of $g_\gamma(t)$. 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. Let $\hat{g}(\omega)$ be the Fourier transform of $g(t)$. Equation (1) yields

$$\hat{g}_\gamma(\omega) = \sqrt{s}\hat{g}(s(\omega - \xi)) e^{-i(\omega-\xi)u}. \tag{2}$$

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 dictionary of time-frequency atoms $\mathcal{D} = (g_\gamma(t))_{\gamma \in \mathbf{\Gamma}}$ is a very redundant set of functions in $\mathbf{L}^2(\mathbf{R})$ that includes window Fourier frames and wavelet frames [3]. Instead of expanding the signal on such a frame that is chosen a priori, we want to choose within $\mathcal{D}$ the time-frequency atoms that are

1 Introduction

In this paper we focus on the problem of approximating functions using linear combinations of a small number waveforms. To obtain a compact expansion of a function which contains complex structures, we must adapt our expansion to the various components of the function. Examples of such a linear expansions include triangular mesh approximations to surfaces, used in scientific computing applications. These mesh expansions can be adapted to obtain low approximation error with a small number of triangles by varying the size and shape of the triangles according to the approximated surface's local properties. Adaptive linear expansions can be used to extract information from signals. We obtain an adaptive time-frequency decomposition of a signal by expanding the signal into a sum of waveforms whose localizations in time and frequency match those of the different signal structures. Such adaptive time-frequency representations are important in signal processing applications such as speech analysis.

The waveforms which we use for our expansions are drawn from a large and redundant collection, called a dictionary. In section 2 we examine the computational complexity of optimally approximating a function with a linear combination of vectors from a dictionary. We prove that in a finite dimensional space, computing the optimal solution is an NP-complete problem, which motivates the use of sub-optimal greedy algorithms. We introduce the matching pursuit algorithm, a greedy algorithm which computes function expansions by iteratively selecting dictionary vectors which best correlate to signal structures. An orthogonal version of the matching pursuit algorithm is also described and compared with the non-orthogonal algorithm. An application of matching pursuits to finding adaptive time-frequency decompositions is explained in section 6. A signal is decomposed into waveforms selected from a dictionary of time-frequency atoms, a collection of dilations, translations, and modulations of a single window function. We construct a time-frequency energy distribution by summing the Wigner distributions of the selected time-frequency atoms. Unlike the Wigner distribution or Cohen's class distributions, this energy distribution does not include interference terms and thus provides a clear picture of the time-frequency plane.

Matching pursuits have chaotic properties which are analyzed for particular dictionaries. As the number of iterations of a matching pursuit increases, the approximation error converges to the realization of a noise process whose energy is uniformly spread across all the dictionary vectors.

Adaptive Time-Frequency Approximations with Matching Pursuits

Geoffrey Davis, [1] *Stéphane Mallat,* [2] *and Zhifeng Zhang*

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

Abstract

Computing the optimal expansion of a signal in a redundant dictionary of waveforms is an NP-complete problem. We introduce a greedy algorithm called a matching pursuit which computes a sub-optimal expansion. The dictionary waveforms which best match a signal's structures are chosen iteratively. An orthogonalized version of the matching pursuit is also developed. Matching pursuits are general procedures for computing 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 contain interference terms, unlike the Wigner and Cohen class distributions. A matching pursuit is a chaotic map whose asymptotic properties are studied. We describe an algorithm which isolates the coherent structures of a signal and show an application to pattern extraction from noisy signals.

[1] Supported by an ONR/ASEE graduate fellowship

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