FAST SPARSE SUBBAND DECOMPOSITION USING FIRSP

Mike Davies[†] *and Laurent Daudet*[‡]

[†] DSP & Multimedia Group
Queen Mary University of London
Mile End Road, London E1 4NS, UK
michael.davies@elec.qmul.ac.uk

[‡] Laboratoire d'Acoustique Musicale
Université Paris 6,
11 rue de Lourmel 75015 Paris, France
daudet@lam.jussieu.fr

ABSTRACT

This paper presents a new fast algorithm for generating sparse signal approximations within an overcomplete subband representation. While the current paper concentrates on sparsifying the Modulated Complex Lapped Transform the theory is applicable to representations composed of a general union of orthonormal bases. We illustrate our method on an audio signal and demonstrate the coding gain of such representations.

1. INTRODUCTION

Sparse signal representations have become increasingly popular in a number of fields such as: signal processing [1, 9, 16, 11, 3], Independent Component Analysis (ICA) [19, 12, 4] and Machine Learning [18, 7]. Typically the aim is to exploit the redundancy in an overcomplete dictionary to obtain a compact representation of the signal. Furthermore sparse decompositions are by their very nature *generative*, concentrating on reconstruction equations. This differs markedly from Frame theory [13], which concentrates on the analysis equations.

A number of criteria for sparsity have been proposed and a variety of algorithms for solving the resulting optimization problem have been developed. In [3] we applied an Iterative Re-weighted Least Squares (IRLS) based algorithm to sparsify the Modulated Complex Lapped Transform (MCLT) which is a 2× overcomplete subband decomposition composed of the union of 2 orthonormal bases introduced by Malvar [14]. The algorithm in [3] is equivalent to the regularized FOCUSS algorithm [16] and the sparse regression algorithm of Figueiredo [7].

In this paper we present a new sparsifying algorithm that avoids the expensive matrix inversion that dominates the computational cost of most previous sparsifying methods (e.g. [1, 16, 3]). This is replaced by a sequence of scalar *Shrinkage* operations. The algorithm generalizes the IRLS framework and can be interpreted as a generalized Expectation-Maximization (EM) algorithm. We also note that the framework is applicable to a wide class of overcomplete dictionaries beyond the MCLT.

The rest of this paper is set out as follows. In the next section we discuss the concept of sparsity and identify possible iterative solutions. We then introduce our sparse subband decomposition, based on the MCLT. Using the fact that the MCLT is the union of two orthonormal bases, we construct our new algorithm which we have coined *Fast Iteratively Reweighted SParsifier* (FIRSP). This is followed by a simple audio example where we also examine the coding gain obtainable from the sparse representation. We end with a discussion on generalizations and applications.

2. SPARSE DECOMPOSITIONS AND APPROXIMATIONS

Let $\Phi \in \mathbb{C}^{N\times M}$ define an overcomplete basis ($M > N$). Our aim is to determine an *approximate* overcomplete representation $\Phi s = x + e$ of a signal x such that the coefficients, s, are sparse and approximation error, e, is small. Note that even when the approximation error is constrained to zero, overcompleteness provides us with the flexibility to search for a sparse representation [1].

Unfortunately there are currently a plethora of sparsity measures with little indication of their relative merits. However an interesting class of such measures has been examined in the FOCUSS family of algorithms [16]. These aim to minimise the following cost functions:

$$s = \arg\min_s \frac{1}{2v}||x - \Phi s||_2^2 + \lambda \sum_{k=1}^{M} |s|^p \qquad (1)$$

where $0 < p \leq 1$, v is the variance of e and λ is a scaling parameter for the sparsity measure. This optimization problem also has various probabilistic interpretations [5, 16, 7, 8]. More generally it can be shown, [16] that if $p \leq 1$ then a minimum of the cost function is sparse, by which we mean it has no more than N non-zero coefficients. In practice we are looking for approximations that have $K \ll N$ non-zero coefficients. The value of p can be interpreted as controlling the degree of sparsity of the prior placed on s_n. $p = 1$ is equivalent to a Laplacian prior on s_n. At the other extreme $p \longrightarrow 0$ equation 1 becomes:

$$s = \arg\min_s \frac{1}{2v}||x - \Phi s||_2^2 + \frac{1}{2}\sum_{k=1}^{M} \ln|s| \qquad (2)$$

This cost function corresponds to placing a Jeffrey's prior on the variance in a Hierarchical Gaussian model [7][1]

We also note that the Laplacian prior ($p = 1$) is unique in that this is the only prior model that both guarantees sparseness of the solution while also guaranteeing that the cost function is convex and therefore has a unique minimum [1]. Indeed a reasonable criticism of using $p < 1$ is that the cost function typically has many minima (see [9]).

There has also been some interesting research showing that under certain circumstances the minimum l_1 and l_0 solutions are actually equivalent [6]. Unfortunately this only holds when there is no approximation error (as is easily seen from the difference in the hard and soft thresholding solutions that emerge when the dictionary is orthonormal). While

[1] although this is the cost function for $p \longrightarrow 0$ it does not correspond to an l_0 cost function. However, in practice, it appears to be a good approximation

the algorithms presented below can equally be applied with $p = 1$ we have so far found that the benefits of a guaranteed single minimum do not compensate the mildness of the sparsity model. For this reason we concentrate on the strongest sparsity model, using equation (2).

3. SPARSE IRLS SOLUTIONS

We next review the IRLS solution that we used in [3] since this provides the starting point for our new algorithm. Let $W \in \mathbb{R}^{M\times M}$ be a non-negative diagonal weighting matrix. A Weighted Least Squares estimate for s can be obtained through matrix inversion as follows:

$$s = \left(vW + \Phi^H\Phi\right)^{-1}\Phi^H x \tag{3}$$

This solves the following problem:

$$s = \arg\min_s \frac{1}{2v}||\Phi s - x||_2^2 + \frac{1}{2}s^H W s \tag{4}$$

We can now use this to solve equation (1) by iteratively adapting the weighting matrix as a function of the previous estimate for s. To minimize equation (2), at the ith iteration, we choose the weighting matrix to be:

$$W^{(i)}(n,n) = |s_n^{(i-1)}|^{-2} \tag{5}$$

This approach can also be viewed as an EM algorithm when applied to hierarchical Gaussian models [5] and is the approach used by Figueiredo [7] for sparse regression. The Weighted Least Squares solution forms the maximization step, while the re-weighting is equivalent to the Expectation step. (the EM framework also provides a simple means of optimizing v - see [3]).

Finally since the IRLS can be formulated as an EM algorithm we know that it will exhibit the usual monotonic convergence property of EM [5].

4. SPARSE MCLT SUBBAND DECOMPOSITIONS

We now consider the specific case where Φ is the Modulated Complex Lapped Transform (MCLT) introduced by Malvar, [14], which is a simple 2× overcomplete subband decomposition. The MCLT is very similar to the Short Time Fourier Transform [13] and is also the union of 2 orthonormal transforms since the real component of the coefficients is the Modified Cosine Transform (MDCT), while the imaginary part is the Modified Discrete Sine Transform (MDST).

Sparsifying the MCLT was initially examined by the authors in [3] using the fact that the operator $\Phi^H\Phi$ is block tridiagonal. Here we will instead make use of the special orthonormal structure within the MCLT.

A further important aspect of the MCLT is that its redundancy makes it approximately shift invariant, a similar concept to shiftability proposed by Simoncelli [17]. By using a 2× overcomplete complex filterbank (e.g. [10, 14]) each complex subband is approximately alias free and hence approximately shiftable [17].

In generating sparse subband decompositions we should try to preserve this property. This can be done by imposing priors on the coefficients that are *phase-invariant*. That is, in our sparsity model we impose sparsity on $|s_n|$ rather than on the real and imaginary components of s_n individually (see[3] for more details). This means that we have N weights to calculate rather than $2N$ which is important when we measure the coding cost of the decomposition in section 6.1.

5. FAST ITERATIVELY RE-WEIGHTED SPARSIFIER

When adopting the IRLS approach to sparsification each update requires the inversion of the $M \times M$ matrix $(W + \Phi^H\Phi)$. Similar matrix inversion is necessary in Linear Programming solutions [1]. Efficiency depends crucially on how easily we can perform this step. For example we can try to exploit fast transform properties to help this calculation. Here we present a new algorithm called the *Fast Iteratively Reweighted SParsifier* (FIRSP), which also exploits the transform properties, however, we also crucially do not attempt to fully solve the Weighted Least Squares problem. Instead of using EM to solve equation (2) let us consider generalizations of EM that might take on a simpler algorithmic form. Specifically we know from EM theory that the maximization step can be replaced by any operation that guarantees to increase the likelihood (decrease the cost function). One such generalization is the Expectation Conditional Maximization (ECM) algorithm [15]. This replaces the Maximization step by a sequence of Conditional Maximization steps. The nature of EM means that there is a great deal of flexibility in the ordering of the various CM steps and the corresponding E step as discussed below.

We can now formally describe the iteration of our new algorithm. Let our overcomplete basis be divided into two orthonormal bases: $\Phi = (\Phi_c, \Phi_s)$ with associated real coefficients c and s. For the MCLT these correspond to the inverse MDCT and inverse MDST transforms. We can now solve a Weighted Conditional Least Squares problem where we freeze the values of s and optimize for c:

$$c = \left(vW + \Phi_c^T\Phi_c\right)^{-1}\Phi_c^T(x - \Phi_s s) \tag{6}$$

Since Φ_c is orthonormal $\Phi_c^T\Phi_c = I$ and therefore the matrix inversion reduces to a diagonal *shrinkage* operator [13]:

$$c_n = \frac{1}{(vW_{(n,n)} + 1)}\left(\Phi_c^T(x - \Phi_s s)\right)_n \tag{7}$$

An identical expression can be calculated for the Conditional Maximization of s.

The iteration is finally completed by calculating the new weights:

$$W_{(n,n)} = (|s_n|^2 + |c_n|^2)^{-1} \tag{8}$$

W takes this particular form since we want the sparsity model to act on the complex pairs $c_n + js_n$, i.e. be phase-invariant. Note that in practice it is more convenient to work with W^{-1}.

In the implementation below the re-weighting step is performed after each CM. One full iteration is therefore considered to be composed of two CM steps and two re-weighting steps. Examining this iteration we see that the computational cost is dominated by the need to map from one transform domain to another. The cost of the shrinkage and weight calculations are trivial by comparison. Overall one iteration takes approximately 4× the computation for a single MDCT, which itself is implemented through fast FFT-based algorithms.

Despite the fact that we are no longer solving the full Weighted Least Squares, the convergence of the algorithm does not appear to have been drastically altered (see below). Similar observations for the ECM algorithm have been made in other applications [15].

5.1 Comment

There is considerable flexibility in the order in which we apply the various CM and E steps. For example we could transform into, say, the MDCT domain and repeatedly apply a CM-step followed by the E-step until convergence, then we could transform into the MDST domain and repeat the procedure, etc. A possible motivation for this is that the asymptotic mapping due to repeated CM followed by re-weighting can be calculated analytically and takes the form of a single diagonal nonlinear shrinkage operator. In this particular application we have not found this to give us significant improvements, however it might prove beneficial for unions of different orthonormal bases.

6. AUDIO EXAMPLE

To illustrate the performance of our method we apply the iterative shrinkage algorithm to a short extract (scaled between -1 and $+1$) from a guitar solo. The audio was sampled at $44.1kHz$ and we used an MCLT with a frame size of 1024. Figure 1 shows the MCLT "spectrogram" for the audio signal. The signal was then sparsified using a fixed $v = 10^{-5}$. 10 iterates of the FIRSP algorithm were computed. In contrast to the initial redundant basis only 6% of the complex coefficients remained non-zero and the resulting approximation had a SNR of: 38dB. The generative MCLT "spectrogram" for the sparse coefficients is shown in figure 2.

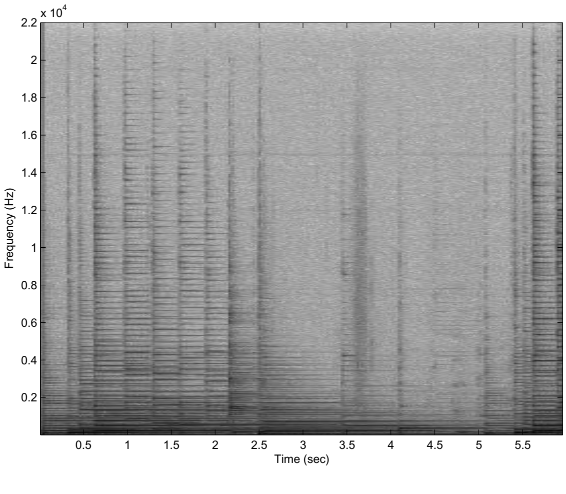

Figure 1: *MCLT "spectrogram" of the guitar data.*

The evolution of the algorithm is best seen by plotting the size of the coefficients sorted in order of magnitude for each iterate. This is shown in figure 3 (solid lines) along with the magnitude of the MDCT coefficients (dashed). From this it can be seen that most of the coefficients shrink to zero after a few iterates. Indeed the number of iterations required is very similar to the full IRLS scheme investigated in [3], while the computational cost of each iterate is a small fraction of that for the IRLS scheme.

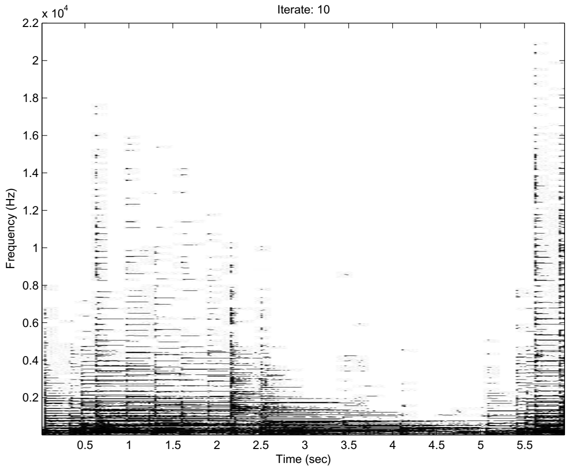

Figure 2: *sparse MCLT representation of the guitar data.*

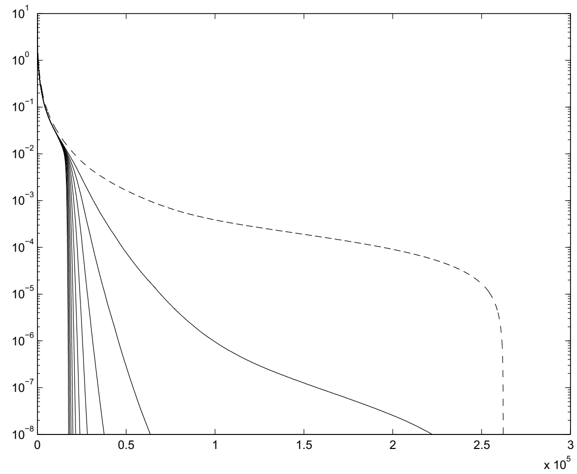

Figure 3: *A sorted plot of MCLT coefficient amplitude for the FIRSP algorithm (solid lines - iterations increasing from right to left) and MDCT coefficient amplitude (dashed).*

6.1 Coding cost for overcomplete representations

For a sparsely coded signal, increasing sparsity typically results in more efficient coding since the dominant coding cost is due to the significance map, [13]. For this reason a simple measure of the coding rate, R, (in bits per sample) for a quantization of K bits per significant coefficient is as follows:

$$R = \mathscr{H}(p_s) + 2p_sK \tag{9}$$

where p_s is the probability that a coefficient will be significant. The first term, $\mathscr{H}(p_s)$ measures the cost of the significance map, while the second term measures the cost of encoding the significant bits (with no additional compression).

Note that in our complex transform, the number of coefficients is the same as the number of samples. Thus, while we need 2× the number of bits in order to encode the significant complex coefficients we do not need to account for twice the entropy of the significance map (as we would have to if we had treated the cosine and sine coefficients independently).

We applied this measure of coding cost to our sparse subband architecture for the audio sample examined above.

Again the frame size was 1024 and 30 iterations of the FIRSP operator were applied. Figure 4 shows the rate-distortion plots for the sparse coefficients with different levels of approximation (i.e. different values of v) and differing levels of quantization ($2 \leq K \leq 14$)). From the plot it is clear that the best values of v and K are linked. The figure also shows the coding cost for the quantization of the MDCT coefficients, using the same formula. When the noise level and quantization resolution are well matched the sparse subband representation can be seen to provide a coding gain over the basic MDCT transform.

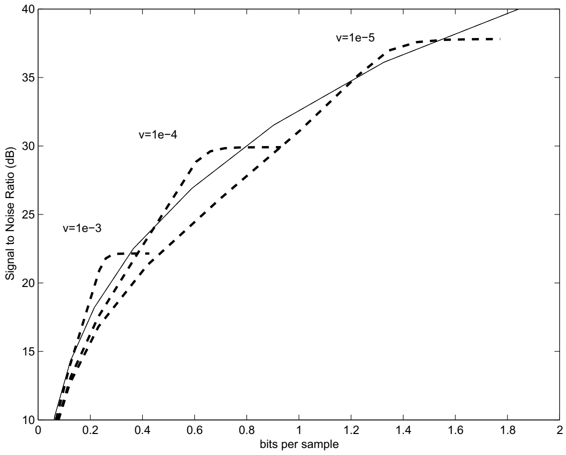

Figure 4: *A plot of Signal-to-Noise Ratio against the coding rate for:* $v = 10^{-3}, 10^{-4}$ *and* 10^{-5} *and different levels of quantization (dashed); and for MDCT coefficients (solid).*

7. DISCUSSION

In this paper we have presented a scheme for generating sparse subband decompositions based on the MCLT. To our knowledge this is the first sparsifying algorithm that can be practically applied to large datasets, requiring as few as 10 iterates, each roughly equivalent to 4 MDCT calculations. This is orders of magnitude fast than previous Basis/Matching Pursuits, [1, 13, 3]. As such, we believe that this algorithm is destined to have a large impact on a wide range of coding and signal processing problems. For example we believe that it could prove a strong competitor to Matching Pursuit as the basis for new scalable and low bit rate coding schemes. Furthermore sparse representations also have excellent statistical properties, providing better performance for de-noising and source separation ([13, 19, 4]).

While we have concentrated on the MCLT for our overcomplete dictionary here, the theory is applicable to general unions of orthonormal bases such as: combinations of the MDCT and wavelet bases [2]; or Kingsbury's Dual Tree Complex Wavelet Transform [10]. There is also potential to extend the framework in other directions by exploiting the hierarchical nature of our model and introducing further structure e.g. persistence or harmonic dependencies.

REFERENCES

[1] S. Chen and D.L. Donoho, "Atomic decomposition by basis pursuit," SIAM J. Sci. Computation, Vol. 20(1), pp 33-61, 1999.

[2] Daudet L. and Torrésani, B., Hybrid representations for audiophonic signal encoding, Signal Processing, vol. 82, pp.1597-1617, 2002.

[3] M. E. Davies and L. Daudet, "Sparsifying Subband Decompositions" *Proc. of IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoustics*, Oct. 2003.

[4] M.E. Davies and N. Mitianoudis, "A Simple Mixture Model for Sparse Overcomplete ICA," *IEE Proceedings-VISP*, special issue on nonlinear and non-Gaussian signal processing, in press, 2004.

[5] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B*, Vol. 39(1), pp1-38, 1977.

[6] D. L. Donoho and X. Huo, "Uncertainty Principles and Ideal Atomic Decomposition" *IEEE Trans. IT*, Vol. 47(7), 2001.

[7] M. Figueiredo, "Adaptive sparseness using Jeffreys prior," NIPS, 2001.

[8] M. Girolami, "A Variational Method for Learning Sparse and Overcomplete Representations," *Neural Computation*, 13(11), pp. 2517-2532, 2002.

[9] I. F. Gorodnitsky and B.D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a reweighted minimum norm algorithm," *IEEE Trans. SP*, vol. 45(3), 1997.

[10] N. Kingsbury, "Complex Wavelets for Shift Invariant Analysis and Filtering of Signals," *Appl. and Comp. Harmonic Analysis*, vol. 10, pp. 234-253, 2001.

[11] K. Kreutz-Delgado and B.D. Rao, "Measures and algorithms for best basis selection," *Proc. ICASSP 98*, pp. 1881-1884, 1998.

[12] M.S. Lewicki and T.J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337-365 2000.

[13] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[14] H. Malvar, "A modulated complex lapped transform and its applications to audio processing," ICASSP'99, 1999.

[15] G.J. McLachlin and T. Krishnan, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, 1997.

[16] B.D. Rao, K. Engan, S.F. Cotter, J. Palmer and K. Kreutz-Delgado, "Subset Selection in Noise Based on Diversity Measure Minimization," *IEEE Trans. SP*, vol. 51(3), pp.760-770, 2003.

[17] E.P. Simoncelli, W.T. Freeman, E.H. Adelson and D.J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. IT*, vol. 38(2), pp 587-607, 1992.

[18] M.E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. of Machine Learning*, vol. 1 pp. 211-244, 2001.

[19] M. Zibulevsky and B. A. Pearlmutter, "Blind separation of sources with sparse representations in a given signal dictionary," *Neural Computation*, vol. 13(4) pp. 863-882, 2001.