MDL-BASED TRANSLATION-INVARIANT DENOISING AND ROBUST TIME-FREQUENCY REPRESENTATIONS

Israel Cohen, Shalom Raz and David Malah

Department of Electrical Engineering, Technion — Israel Institute of Technology
Technion City, Haifa 32000, Israel. cisrael@shoshan.technion.ac.il

ABSTRACT

A translation-invariant denoising method, based on the *Minimum Description Length* (MDL) criterion and the *Shift-Invariant Wavelet Packet Decomposition* (SIWPD) is presented. A collection of signal models is generated using an extended library of orthonormal wavelet-packet bases, and an additive cost function, approximately representing the MDL principle, is derived. We show that the minimum description length of the noisy observed data is achieved by utilizing the SIWPD and thresholding the resulting coefficients. The signal estimator is combined with a *modified Wigner distribution*, yielding robust time-frequency representations, characterized by high resolution and suppressed interference-terms. The proposed method is compared with alternative approaches, and its superiority is demonstrated.

1. INTRODUCTION

The use of wavelet bases for estimating noisy signals has been the object of considerable recent research. Traditional methods often entail noise removal by low-pass filtering, thus blurring sharp signal features. In contrast, wavelet-based methods show good performance for a wide diversity of signals, including those containing jumps, spikes and other nonsmooth features [1, 2]. The widely used transform-based thresholding method consists of three steps: transformation of the noisy data into a time-scale domain, soft or hard thresholding to the resulting coefficients, and transformation back into the original space. This scheme necessitates determination of the "best" basis and threshold value, leading to the best signal estimate. It is constructive to employ the library of wavelet-packet bases as a collection of competing models, and select the best model according to the *Minimum Description Length* (MDL) criterion [2, 3, 4]. However, denoising based on the conventional wavelet-packet decomposition (WPD) [5] may exhibit visual artifacts, attributable to the lack of shift-invariance [6].

One approach to attaining shift-invariance is to average the translation dependence: applying a range of shifts to the noisy data, denoising the shifted versions with the wavelet transform, then unshifting and averaging the denoised data [6]. This procedure, termed *Cycle-Spinning*, generally yields better visual performance on smooth parts of the signal. However, transitory (high-frequency) features may be significantly attenuated [7]. Furthermore, the MDL principle and related information-theoretic arguments cannot be applied. Alternatively, one may optimize the time localization of the signal, so that its features are well-aligned with the basis-functions. In the case of WPD, Pesquet *et al.* [8] suggested to adapt the shift of the signal as follows: (*i*) To each node of the expansion tree assign an information-cost by averaging the Shannon entropy over all translations. (*ii*) Determine the best expansion tree using the conventional WPD algorithm of Coifman and Wickerhauser. (*iii*) Compare the entropy of the 2^{κ} orthonormal representations resulting from 2^{κ} different shift-options, where κ denotes the number of nodes in the best expansion tree, and choose that representation (shift-option) which minimizes the entropy. This procedure is sub-optimal compared with the *Shift-Invariant Wavelet Packet Decomposition* (SIWPD) [9], since the expansion tree is determined by the *averaged* entropy. Additionally, the shift-options in step (*iii*) are examined one by one, whereas the SIWPD not only provides a *recursive* selection method for the optimal shift, but also offers an inherent trade-off between the computational complexity and the information cost.

In this paper, we present a translation-invariant signal estimator, which is based on the SIWPD and the MDL criterion. We show that this estimator, combined with the recently introduced modified Wigner distribution (MWD) [10], yields robust time-frequency representations that are characterized by high resolution, *i.e.*, high concentration and suppressed interference-terms. An *extended* library of wavelet-packet bases [9] is employed for generating a collection of competing models, and the MDL principle is applied for approximating the description length of the observed noisy data. We show that minimum description length is attainable by optimizing the expansion-tree associated with the SIWPD. The optimal signal estimate is subsequently obtained by thresholding the resulting coefficients. This estimator is independent of the alignment of the observed signal with respect to the basis functions. Furthermore, the intrinsic advantages of the SIWPD over the conventional WPD are instrumental in generating a relatively superior estimator.

2. PROBLEM FORMULATION

Let $y(t) = f(t) + z(t)$ represent the noisy observed data, where $f(t)$ is the unknown signal to be estimated, and $z(t)$ is a white Gaussian noise with zero mean and a known power spectral density (PSD) σ^2. Let $\{\psi_n(t) : n \in \mathbb{Z}_+\}$ denote a wavelet packet family, generated by an orthonormal scaling function ψ_0 [5]. We assume that $f(t)$ belongs to $\overline{Span}\{\psi_0(t-k) : k \in \mathbb{Z}\}$ and is compactly supported. Accordingly, there exists a finite integer N such that

$$\langle f\,,\ \psi_{\ell,n,m,k}\rangle = 0 \quad \text{if } k < 0 \text{ or } k \geq N2^{\ell}\,, \tag{1}$$

where $\psi_{\ell,n,m,k}(t) \equiv 2^{\ell/2}\psi_n\left(2^{\ell}(t-m)-k\right)$ $(-log_2 N \leq -L \leq \ell \leq 0,\ 0 \leq n,m < 2^{-\ell})$, *i.e.*, $f \in V_0 \equiv$

$\overline{Span}\{\psi_0(t-k) : 0 \le k < N\}$.

To estimate $f(t)$, we use an extended library of wavelet packet bases [9], defined as the collection of all the orthonormal bases for V_0 which are subsets of

$$\left\{ B_{\ell,n,m} \ : \ -L \le \ell \le 0,\ 0 \le n, m < 2^{-\ell} \right\} , \qquad (2)$$

where $B_{\ell,n,m} \equiv \left\{ \psi_{\ell,n,m,k} \ : \ 0 \le k < N2^{\ell} \right\}$. Each basis in this library is given by $\{B_{\ell,n,m} : (\ell,n,m) \in E\}$, where E is a set of indices that represent the terminal nodes of a SIWPD tree [9]. By selecting an appropriate E, the signal f can be represented by a relatively small number $K < N$ of significant expansion coefficients. Thus, we consider a signal estimate of the form

$$\hat{f}(t) = \sum_{n=1}^{K} f_{k_n} \phi_{k_n}(t) \qquad (3)$$

where $\{\phi_k : 1 \le k \le N\} = \{B_{\ell,n,m} : (\ell,n,m) \in E\}$, and $\{k_n\}_{1 \le n \le K}$ are the indices of the significant basis functions.

The problem is to find the best E and $\{k_n\}_{1 \le n \le K}$ (best model) such that the estimate (3) is optimal according to the MDL principle.

3. THE MDL PRINCIPLE

The MDL principle [11] asserts that given a data set and a collection of competing models, the best model is the one that yields the minimal description length of the data. The description length of the data is counted for each model in the collection as the codelength (in bits) of encoding the data using that model, and the codelength needed to specify the model itself. The rationale is that a good model is judged by its ability to "explain" the data, hence the shorter the description length, the better the model.

The encoding, and hence the computation of the codelength, is carried out in three steps: 1) Encoding the observed data assuming E and $\{k_n\}_{1 \le n \le K}$ are given. 2) Encoding $\{k_n\}_{1 \le n \le K}$ assuming that E is given. 3) Encoding E. Accordingly, the total description length of the data is given by

$$\begin{aligned} \mathcal{L}(y) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \quad &\equiv \quad \mathcal{L}\left(y \mid E, \{k_n\}_{1 \le n \le K}\right) \\ &+ \mathcal{L}\left(\{k_n\}_{1 \le n \le K} \mid E\right) + \mathcal{L}(E) . \end{aligned} \qquad (4)$$

Step 1: It was established by Rissanen [11] that the shortest codelength for encoding a data set $\{y_k\}_{1 \le k \le N}$ using the probabilistic model $P(\{y_k\}_{1 \le k \le N} \mid \mu)$, where μ is an unknown parameter vector, is asymptotically given by

$$\mathcal{L}(\{y_k\}_{1 \le k \le N}) = -\log_2 P(\{y_k\}_{1 \le k \le N} \mid \hat{\mu}) + \frac{q}{2} \log_2 N \qquad (5)$$

where $\hat{\mu}$ is the maximum likelihood estimator of μ, and q is the number of free real parameters in the vector μ.

Here, $\mu \equiv \{f_{k_n}\}_{1 \le n \le K}$ and $\hat{f}_{k_n} = y_{k_n} \equiv \langle y, \phi_{k_n} \rangle$ $(1 \le n \le K)$. Consequently,

$$\mathcal{L}_1 = \frac{1}{2\sigma^2 \ln 2} \sum_{n=K+1}^{N} y_{k_n}^2 + \frac{N}{2} \log_2(2\pi\sigma^2) + \frac{K}{2} \log_2 N . \qquad (6)$$

Step 2: The indices $\{k_n\}_{1 \le n \le K}$ can be specified by a binary string of length N containing exactly K 1s. Since there are $\binom{N}{K}$ such possible strings, and K $(1 \le K \le N)$ requires $\log_2 N$ bits (the probability density function for K is unknown), we have

$$\mathcal{L}_2 = \log_2 \binom{N}{K} + \log_2 N = \log_2 \frac{N \cdot N!}{K!(N-K)!} . \qquad (7)$$

By applying Stirling's formula to the factorials and ignoring constant terms, we have $\mathcal{L}_2 \approx K \log_2 N$ for $N \gg K$. The optimal $\{k_n\}_{1 \le n \le K}$ are obtained by minimizing $\mathcal{L}_1 + \mathcal{L}_2$. Clearly,

$$\sum_{n=1}^{N} \min\left(y_n^2 ,\ 3\sigma^2 \ln N\right) \le \sum_{n=K+1}^{N} y_{k_n}^2 + \sum_{n=1}^{K} (3\sigma^2 \ln N) \qquad (8)$$

and equality holds for

$$\{k_n\}_{1 \le n \le K} = \left\{ n : |y_n| > \sigma\sqrt{3 \ln N},\ 1 \le n \le N \right\} . \qquad (9)$$

Hence,

$$\mathcal{L}_1 + \mathcal{L}_2 = \frac{1}{2\sigma^2 \ln 2} \sum_{n=1}^{N} \min\left(y_n^2 ,\ 3\sigma^2 \ln N\right) . \qquad (10)$$

Step 3: A SIWPD tree can be specified by a 3-ary string, which contains 2s for terminal nodes, and either 0s or 1s for internal nodes, depending on their expansion mode [12]. There are $|E|$ terminal nodes and $|E|-1$ internal nodes, where $|E|$ is the cardinality of E. Since the tree always ends with a terminal node, the last 2 in the string can be discarded, and thus we need to encode a sequence containing $|E|-1$ 2s and $|E|-1$ symbols from $\{0,1\}$, The description length of such sequence is

$$\mathcal{L}(E) = \log_2 \binom{2|E|-2}{|E|-1} + (|E|-1) + \log_2 |E| , \qquad (11)$$

where the first term is required to specify the locations of 2s in the sequence, the second term to discriminate between 0s and 1s, and the third term to encode the number of terminal terms. Applying Stirling's formula to the factorials and ignoring constant terms, it follows that $\mathcal{L}(E) \approx 3|E|$ for $|E| \gg 1$. Hence, the total codelength is given by

$$\mathcal{L}(y) = 3\,|E| + \frac{1}{2\sigma^2 \ln 2} \sum_{n=1}^{N} \min\left(y_n^2 ,\ 3\sigma^2 \ln N\right) . \qquad (12)$$

4. SIGNAL ESTIMATION

Let $\mathcal{L}(By)$ denote the description length of y represented on a basis $B = \{B_{\ell,n,m} : (\ell,n,m) \in E\}$, and let $B_{\ell,n,m}y = \left\{ C_{\ell,n,m,k}(y) = \langle y ,\ \psi_{\ell,n,m,k} \rangle \ : \ 1 \le k \le 2^{\ell} N \right\}$ be the expansion coefficients of the observed data. Then, by Eq. (12)

$$\mathcal{L}(By) = \sum_{(\ell,n,m) \in E} \mathcal{L}(B_{\ell,n,m}y) \qquad (13)$$

where the description length associated with a tree-node (ℓ,n,m) is given by

$$\mathcal{L}(B_{\ell,n,m}y) = 3 + \frac{1}{2\sigma^2 \ln 2} \sum_{k=1}^{2^{\ell}N} \min\left\{ C_{\ell,n,m,k}^2(y) ,\ 3\sigma^2 \ln N \right\} .$$

The optimal basis for y is that B for which $\mathcal{L}(By)$ is minimal. Since the codelength in Eq. (13) is an additive

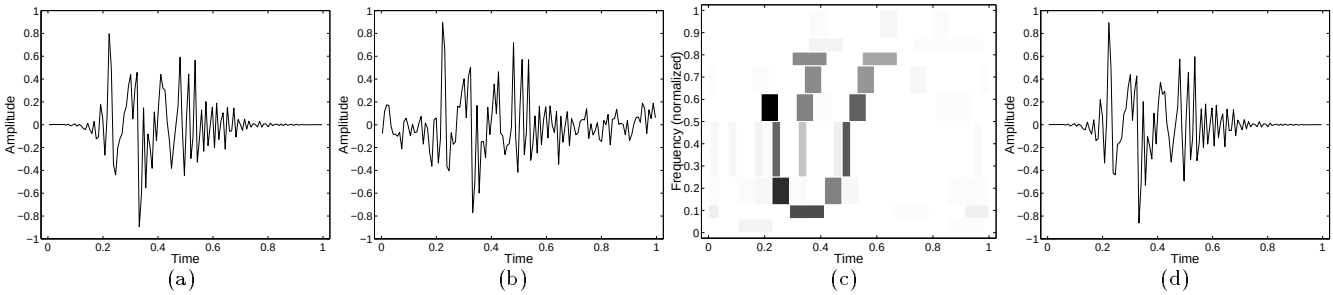

Figure 1. MDL-based translation-invariant denoising: (a) Synthetic signal; (b) Noisy measurement with SNR= 7dB; (c) SIWPD of the noisy measurement using the MDL criterion; (d) Signal estimate, SNR= 19dB.

cost function, which directly results from the expressions and approximations derived in the previous section, applying the SIWPD to y yields the optimal basis. The signal estimate is subsequently obtained by hard-thresholding the coefficients by $\sigma\sqrt{3\ln N}$. Specifically,

$$\hat{f}(t) = \sum_{k\in\Lambda} y_k \hat{\phi}_k(t) \tag{14}$$

where $\left\{\hat{\phi}_k\right\}_{1\le k\le N}$ is the optimal basis, $y_k = \left\langle y, \hat{\phi}_k \right\rangle$, and $\Lambda \stackrel{\triangle}{=} \{k_n\}_{1\le n\le K}$ is obtained by (9).

The proposed estimator, combined with the modified Wigner distribution (MWD) [10], yields robust time-frequency representations. Denote by W_ϕ the auto WD of ϕ, and by W_{ϕ_1,ϕ_2} the cross WD of ϕ_1 and ϕ_2. Then, from [10] and Eq. (14), the MWD estimate of y is given by

$$\hat{T}_y = \sum_{k\in\Lambda} |y_k|^2 W_{\hat{\phi}_k} + 2 \sum_{\{k,k'\}\in\Gamma} \mathrm{Re}\{y_k y_{k'}^* W_{\hat{\phi}_k,\hat{\phi}_{k'}}\}\,. \tag{15}$$

The set $\Gamma = \left\{\{k,k'\} : k,k' \in \Lambda,\, 0 < d(\hat{\phi}_k, \hat{\phi}_{k'}) \le D\right\}$ restricts the cross terms to *neighboring* pairs of basis-functions, *i.e.*, basis-functions whose time-frequency distance is smaller than a certain threshold D. This threshold is adjusted to balance the cross-term interference, the useful properties of the distribution, and the computational complexity [13]. The distance measure in the time-frequency plane is defined by

$$d(\hat{\phi}_k, \hat{\phi}_{k'}) = \left[\frac{(\bar{t}_k - \bar{t}_{k'})^2}{\Delta t_k \Delta t_{k'}} + \frac{(\bar{\omega}_k - \bar{\omega}_{k'})^2}{\Delta\omega_k \Delta\omega_{k'}}\right]^{1/2} \tag{16}$$

where $(\bar{t}_k, \bar{\omega}_k)$ is the position of the cell associated with $\hat{\phi}_k$; Δt_k and $\Delta\omega_k$ are, respectively, the widths (uncertainties) in time and frequency. Similar notations apply to $\hat{\phi}_{k'}$.

The proposed estimate is robust to noise and possesses the useful properties of the modified Wigner distribution, namely high energy concentration, well delineated components, low interference-terms, *etc.* [12].

5. AN EXAMPLE

A synthetic signal, created by a linear superposition of a few wavelet packets (12−tap coiflet filters), and a noisy version with signal-to-noise ratio SNR= 7dB are depicted in Figs. 1(a)-(b). The optimal SIWPD of the noisy signal using the MDL criterion is shown in Fig. 1(c). The expansion coefficients are thresholded by $\sigma\sqrt{3\ln N}$ and transformed back into the signal domain. Compared to the noisy measurement, the signal estimate (Fig. 1(d)) is enhanced to SNR= 19dB. Table 1 compares the SNRs obtained by alternative methods. Their deficient performance results from the restricted compression capability of the WPD. The proposed method uses the SIWPD, which optimizes the representation by incorporating translations of wavelet-packets for signal components that are not aligned with the basis elements.

Denoising Method	SNR (dB)
Saito [2]	1.1
Basis-Pursuit [14]	4.3
Donoho-Johnstone [1]	6.4
Matching-Pursuit	7.5
The proposed method	19.1

Table 1. Signal-to-noise ratios for the signal estimates of the synthetic signal using the library of wavelet packets (12-tap coiflet filters) and various denoising methods.

Fig. 2 shows the WD, smoothed pseudo Wigner distribution and MWD for the synthetic signal. The results for the noisy version are depicted in Fig. 3. Expectedly, the WD of the multicomponent signal is corrupted by interference terms, while the smoothed pseudo Wigner distributions are robust in the noisy environment. However, we can readily observe that the energy concentration of the signal components is poor. In contrast, the proposed estimate of the MWD retains its robustness and insensitivity to noise while providing the desired time-frequency resolution.

6. RELATION TO OTHER WORK

In [2], several libraries of wavelet-packet bases are considered. The "best basis" in each is selected using WPD and the Shannon entropy criterion. Subsequently, the MDL principle is applied for determining the optimum among the "best bases". The proposed method translates the MDL criterion into an additive information cost function, thus rendering the best-basis search manageable. Furthermore, it uses the SIWPD, which yields sparser representations and better estimates than the WPD.

In [3], WPD is applied to the noisy data using an information cost

$$\mathcal{M}(\{y_n\}) = \sum_n \min\left(y_n^2\,,\, 2\sigma^2 \log_2 N\right)\,. \tag{17}$$

The signal estimate is reconstructed from coefficients whose magnitudes are larger than $\sigma\sqrt{2\log_2 N}$. This method ignores the description length of the expansion tree (see

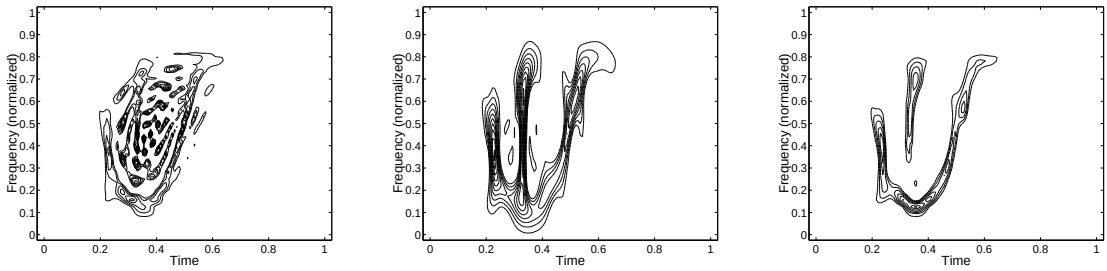

Figure 2. Contour plots for the test signal: (a) Wigner distribution; (b) Smoothed pseudo Wigner distribution; (c) The modified Wigner distribution [10].

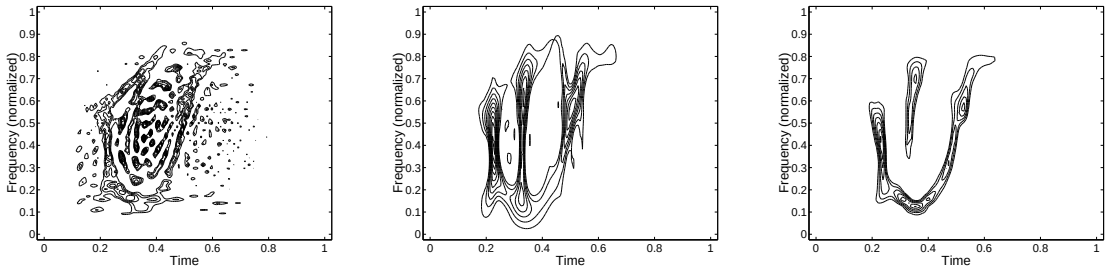

Figure 3. Contour plots for noisy test signal: (a) Wigner distribution; (b) Smoothed pseudo Wigner distribution; (c) The proposed time-frequency distribution estimate.

Eq. (12)), and expectedly the estimator is sensitive to signal translations.

In [1], the criterion for the optimal basis selection is minimum mean-squared error, rather than MDL. The best-basis minimizes

$$\mathcal{M}(\{y_n\}) = \sum_n \min\left(y_n^2\,,\,\zeta^2\right)\,, \tag{18}$$

where $\zeta = \nu\sigma(1+\sqrt{2\ln M_N})$, M_N is the number of distinct basis-functions contained in the library (for WPD, $M_N = N\log_2 N$) and $\nu > 8$. The signal is reconstructed on the best-basis from coefficients whose magnitudes are larger than ζ. The threshold ζ is larger than $\sigma\sqrt{3\ln N}$ by at least a factor of $8\sqrt{2/3}$. Thus, criterion (18) imposes a larger penalty on nonzero coefficients, but none is associated with the complexity of the expansion-tree (see Eq. (12)).

REFERENCES

[1] D. L. Donoho and I. M. Johnstone, "Ideal denoising in an orthonormal basis chosen from a library of bases", *Comptes Rendus Acad. Sci., Ser. I*, Vol. 319, 1994, pp. 1317–1322.

[2] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion", in: E. Foufoula and P. Kumar, eds., *Wavelets in Geophysics*, Academic Press, 1994, pp. 299–324.

[3] H. Krim, and J.-C. Pesquet, "On the statistics of best bases criteria", in: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Springer-Verlag, 1995, pp. 193–207.

[4] P. Moulin, "Signal estimation using adapted tree-structured bases and the MDL principle", *Proc. IEEE Int. Sym. Time-Freq. Time-Scale Anal.*, Paris, France, 18–21 June 1996, pp. 141–143.

[5] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection", *IEEE Trans. Inform. Theory*, Vol. 38, No. 2, Mar. 1992, pp. 713–718.

[6] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising", in: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Springer-Verlag, 1995, pp. 125–150.

[7] N. A. Whitmal, J. C. Rutledge and J. Cohen, "Wavelet-based noise reduction", *Proc. ICASSP-95*, Detroit, Michigan, 8–12 May 1995, pp. 3003–3006.

[8] J.-C. Pesquet, H. Krim and H. Carfantan, "Time-invariant orthonormal wavelet representations", *IEEE Trans. Sig. Proc.*, Vol. 44, Aug. 1996, pp. 1964–1996.

[9] I. Cohen, S. Raz and D. Malah, "Orthonormal shift-invariant wavelet packet decomposition and representation", *Sig. Proc.*, Vol. 57, Mar. 1997, pp. 251–270.

[10] I. Cohen, S. Raz and D. Malah, "Eliminating interference terms in the Wigner distribution using extended libraries of bases", *Proc. ICASSP-97*, Germany, 20–24 Apr. 1997, pp. 2133–2136.

[11] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[12] I. Cohen, S. Raz and D. Malah, "Translation-invariant denoising using the minimum description length criterion", Tech. Rep., CC PUB No. 246, Dept. of Elect. Eng., Technion - IIT, Israel, June 1998.

[13] I. Cohen, S. Raz and D. Malah, "Adaptive time-frequency distributions via the shift-invariant wavelet packet decomposition", *Proc. IEEE Int. Sym. Time-Freq. Time-Scale Analysis*, , Pittsburgh, PA, 6–9 Oct. 1998.

[14] S. Chen and D. L. Donoho, "Atomic decomposition by basis pursuit", Tech. Rep., Dept. of Statistics, Stanford Univ., Feb. 1996.