Signal Estimation Using Wavelet-Markov Models

Matthew S. Crouse and *Richard G. Baraniuk* *

Dept. of Electrical and Computer Engineering
Rice University
Houston, TX 77005

Robert D. Nowak

Dept. of Electrical Engineering
Michigan State University
East Lansing, MI 48824

Abstract

Current wavelet-based statistical signal and image processing techniques such as shrinkage and filtering treat the wavelet coefficients as though they were statistically independent. This assumption is unrealistic; considering the statistical dependencies between wavelet coefficients can yield substantial performance improvements. In this paper, we develop a new framework for wavelet-based signal processing that employs hidden Markov models to characterize the dependencies between wavelet coefficients. To illustrate the power of the new framework, we derive a new algorithm for signal estimation in nonGaussian noise.

1 Introduction

Wavelets have emerged as an exciting new tool for statistical signal and image processing. The wavelet transform is an atomic decomposition that represents a signal $z(t)$ in terms of its projections w_i onto shifted and dilated versions $\psi_i(t)$ of a prototype bandpass wavelet function $\psi(t)$. The w_i are referred to as the *wavelet coefficients* and measure the content of the signal at various locations in time and frequency (see Figure 1).

The joint time-frequency analysis effected by the wavelet transform has some attractive properties that make it natural for statistical applications, including estimation [1, 2, 3], detection, and classification. We call these the *primary properties* of the wavelet transform:

Locality: Each wavelet atom ψ_i is localized simultaneously in time and frequency. Therefore, wavelets can match a wide range of different signal components, from transients to harmonics.

Multiresolution: Wavelet atoms compress and dilate to analyze at a nested set of scales. This allows the transform to match both short-duration and long-duration signal structures.

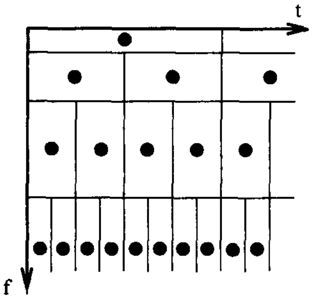

Figure 1. *Tiling of the time-frequency plane by the atoms of the wavelet transform. Each box depicts the idealized support of a wavelet atom ψ_i in time-frequency; the black dot at the center corresponds to the wavelet coefficient w_i. Each different row of wavelet atoms corresponds to a different* scale or frequency band. *(We run the frequency axis down rather than up for later convenience.)*

Compression: The wavelet transforms of real-world signals and images tend to be sparse.

Attention has focused on *scalar* processing of the wavelet coefficients [1]. Scalar wavelet processing algorithms are based on the primary properties above plus an interpretation of the transform as a "decorrelator" that attempts to make each wavelet coefficient statistically independent of all others. If this were possible for all signals and images, then simple scalar processing in the wavelet domain would be optimal.

However, the wavelet transform cannot completely decorrelate real-world signals and images — a *residual dependency structure* always remains between the wavelet coefficients. In words, we have following *secondary properties* of the wavelet transform:

Clustering: If a particular wavelet coefficient is large/small, then neighboring coefficients are very likely to also be large/small.

Persistence across Scale: Large/small values of wavelet coefficients tend to propagate across scales.

Both of these empirical observations have been exploited with tremendous success by the compression community [4]. Our goal is to do the same for signal processing.

*This work was supported by the National Science Foundation, grant no. MIP–9457438, and the Office of Naval Research, grant no. N00014–95–1–0849.

Email: mcrouse@rice.edu, richb@rice.edu, rnowak@egr.msu.edu
Web: http://www-dsp.rice.edu, http://www.egr.msu.edu/spc/

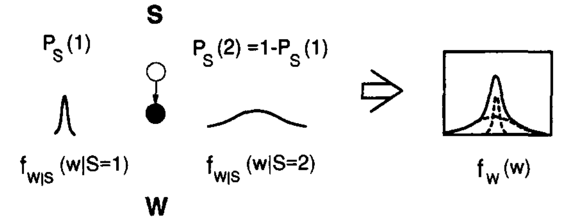

Figure 2. *A two-state Gaussian mixture model for a random variable W. We denote the state variable S with a white dot, the random variable W with a black dot. Illustrated are the Gaussian conditional pdf's for $W|S$ as well as the overall mixture pdf for W. In our application, we model each wavelet coefficient W_i (each black dot in Figure 1) in this way.*

In this paper, we introduce the concept of probabilistic graphs (specifically Hidden Markov models) for characterizing the dependencies between the coefficients of the wavelet transform. Our marriage of wavelet transforms and Hidden Markov models yields a flexible framework for statistical signal and image processing that both matches the properties of the wavelet transform and exploits the structure inherent in real-world signals and images. This framework provides a natural setting for signal estimation, detection, classification, and even synthesis. In particular, we will use this new theory to develop a new algorithm for signal estimation in nonGaussian noise.

2 Wavelet-Markov Models

Recall the Compression property of the wavelet transform. The transform of a typical signal or image consists of a small number of large coefficients and a large number of small coefficients. Thus, we can roughly model each coefficient as being in one of two states: "high" or "low." If we associate with each state a pdf — say a high-variance, zero-mean density for the "high" state and a low-variance, zero-mean density for the "low" state — the result is a two-state mixture model for each wavelet coefficient.

In this paper, we will model each wavelet coefficient as a random variable W_i with a two-state (zero-mean) Gaussian mixture density. Empirically, this model has proven both effective and convenient [2, 3]. As we see from Figure 2, this simple model is completely parameterized by the pmf of the state variable S_i, $p_{S_i}(1)$, $1-p_{S_i}(1)$, and the variances of the Gaussian pdf's corresponding to each state, $\sigma^2_{i,1}$, $\sigma^2_{i,2}$. We say that the state variables are *hidden*, because their values are not observed directly, but rather are gleaned from the observed wavelet coefficients.

Based on the wavelet Clustering and Persistence Across Scales properties, we expect probabilistic coupling between the state variables. Simply put, these two properties suggest that the state of a given wavelet coefficient is likely to be high (low) if its neighbors across time and scale are high (low). To capture this behavior, we introduce a Markovian structure on the hidden states using a probabilistic graph [6, 7]. The Locality and Multiresolution properties of the wavelet transform suggest three simple graphs for characterizing the local dependencies between the wavelet coefficients of Figure 1. In Figure 3 we illustrate these graphs, which are formed by "connecting the dots" representing the wavelet state variables. We call these graphs *wavelet-Markov models.*

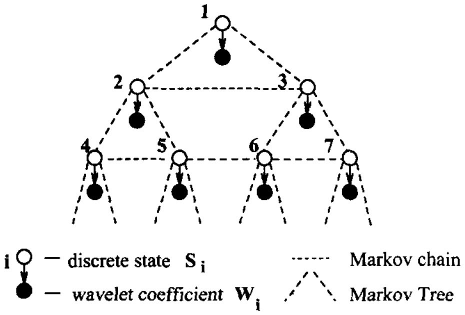

Figure 3. *Wavelet-Markov models for capturing the statistical dependencies of the coefficients of a wavelet transform. Each black dot represents a continuous wavelet coefficient W_i. Each white dot represents the (hidden) mixture state variable S_i for W_i. Removing all dashed connections corresponds to the Independent Mixture Model. Connecting white dots horizontally across time yields the Hidden Markov Chain Model. Connecting white dots diagonally across scales yields the Hidden Markov Tree Model.*

Independent Mixture Model (IMM): Removing all connections between state variables S_i in Figure 3 leads to the model presented in [2, 3]. It treats the wavelet state variables (and hence the wavelet coefficients) as independent.

Hidden Markov Chain Model: Connecting the state variables S_i horizontally in Figure 3 specifies a Markov chain [5] dependency between the state variables *within each scale.* This new model treats wavelet state variables as dependent within each scale, but independent from scale to scale.

Hidden Markov Tree Model: By connecting state variables *across scales* in Figure 3, we obtain a graph with tree-structured dependencies between state variables.

The Hidden Markov Tree Model matches both the Clustering and Persistence across Scale properties of the wavelet transform. Its structure is reminiscent of the zerotree wavelet compression system [4], which exploits tree-structured dependencies for substantial compression gains.

The Hidden Markov Tree Model has a natural parent-child dependency interpretation, which is defined formally by a directed tree graph [6, 7]. State variable dependencies are modeled via state transition probabilities from each parent state variable S_i to its "children," the two state variables connected to it from below (if they exist). For example, in

Figure 3, state variables S_4 and S_5 are both children of S_2, and hence causally dependent on S_2. Dependency is not simply limited to parent-child interactions, however. State variables S_4 and S_5 may be highly dependent due to their joint dependency with S_2.

Let $S_{\rho(i)}$ denote the parent of S_i. Using a zero-mean Gaussian mixture model for each wavelet coefficient value W_i, the parameters for the Hidden Markov Tree Model are:

1. $p_{S_1}(m)$, the pmf for the root S_1.
2. $\epsilon_{i,\rho(i)}^{mr} = p_{S_i|S_{\rho(i)}}(m|S_{\rho(i)} = r)$, the probability that S_i is in state m given $S_{\rho(i)}$ is in state r.
3. $\sigma_{i,m}^2$, the variance of the wavelet coefficient W_i given S_i is in state m.

A theory exists for analyzing more complicated graphs [6], such as those obtained by linking state variables across both time and scale, but it is beyond the scope of this paper.

3 Model Training and Likelihood Determination

We have defined three probabilistic graphs for capturing the structure in a wavelet transform. To use these graphs for signal processing, two operations are of interest:

Model Training: Given a set of training data, estimate the model parameters to achieve a maximum-likelihood (ML) fit.

Likelihood determination: Given a fixed model, calculate the probability of the observed wavelet data using the model.

Training is fundamental to any application. Once we have trained the model on a signal or class of signals, we can apply it to tasks such as estimation, classification, prediction (useful for compression), and synthesis. Likelihood determination not only is useful for tasks such as detection and classification, but also is a key component of training.

We train our models by choosing parameters that maximize the likelihood of the observed wavelet coefficients.[1] These parameters are the state transition probabilities and conditional Gaussian variances. Unfortunately, the fact that we cannot observe the hidden state variables means that closed-form parameter estimates are unobtainable. We circumvent this obstacle using Expectation Maximization (EM) algorithms.

For each of the three graphs discussed above, it can be shown that a specialized EM-type algorithm converges to a local maximum of the likelihood function [5, 9, 10]. Moreover, for these graphs, the Expectation step is equivalent to likelihood determination. For details on the specific expectation and maximization steps for the three different graphs see [8].

4 Application to Signal Estimation

We now apply wavelet-Markov modeling to signal estimation in additive white *nonGaussian* noise, extending the work done in [11] for estimation in white Gaussian noise. The estimation problem is expressed in the wavelet domain as $w_i = \theta_i + n_i$, where w_i, θ_i, and n_i denote the wavelet coefficients of the observed data, the signal, and the noise, respectively. We assume the noise in the signal domain is independent identically distributed (iid) and independent of the signal. The structure of the wavelet transform leads to wavelet domain noise that is uncorrelated, identically distributed within each scale, and independent of the signal.

Using a wavelet-Markov model for the signal prior and an IMM for the noise prior, we apply an "empirical Bayesian" estimation approach that automatically learns the prior densities from the noisy data. The prior densities are used to find minimum mean-squared-error (MMSE) conditional mean estimates $E[\theta_i|w_1, w_2, \ldots, w_n]$ for each signal wavelet coefficient θ_i. The estimates are relatively straightforward to compute, since the signal and noise priors involve coefficients that are conditionally Gaussian [8]. Hence, the major task of our approach is estimating the prior signal and noise densities.

Since the wavelet domain noise is generally not identically distributed across scale, *different* IMM noise priors are required for each scale. If a noise-only observation is available, we use it to estimate an IMM noise prior at each scale. If only one noisy signal observation is at hand, we first estimate a noise IMM in the finest scale, where the signal energy is assumed negligible. Then, using the finest-scale noise IMM, we can easily deduce IMMs for the other scales [8].

Estimating the wavelet-Markov signal prior is a nontrivial task, since we do not directly observe the signal but rather signal in noise. We use a modified EM algorithm that maximizes the likelihood of the observed *signal plus noise* as a function of the wavelet-Markov signal model. Exact details of empirically estimating both the signal and the noise priors are provided in [8].

Laplacian Noise Example: NonGaussian noise can exhibit properties quite different from Gaussian noise of the same power — much more "spikiness," for example. Additionally, the wavelet transform of iid nonGaussian noise is distributed differently in each scale, with the noise in coarser scales tending towards Gaussian by the Central Limit Theorem. Hence, wavelet-based de-noising algorithms assuming iid Gaussian noise may perform poorly.

[1] To obtain reliable parameter estimates it is desirable to have multiple iid observations of the entire set of wavelet coefficients. Often, however, only a single realization is observed. To estimate the parameters in this situation we average over wavelet coefficients assumed to be statistically similar, a practice known as *tying* [5]. For details see [8].

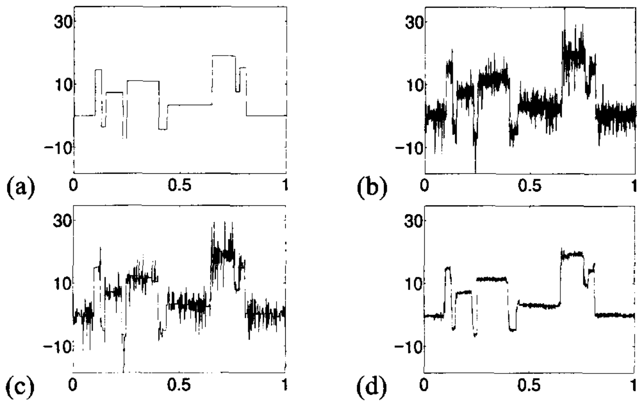

Figure 4. *Example of de-noising in white, nonGaussian noise. (a) Length-1024 Blocks signal. (b) In Laplacian noise, MSE = 8.0. (c) De-noised via SureShrink [1], MSE = 6.3. (d) De-noised via wavelet-based Bayesian estimation, MSE = 1.9. For both algorithms, the Haar wavelet filter was used to transform the signal, and the MSE's were averaged over 1000 Monte Carlo trials.*

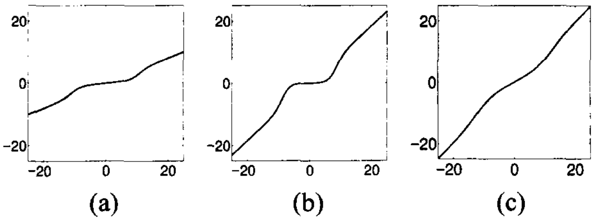

Figure 5. *Nonlinearities applied for de-noising Blocks of Figure 4 at the (a) 2nd, (b) 4th, and (c) 7th scales of a 7-scale wavelet transform, with the scale index increasing from fine resolution to coarse resolution.*

We illustrate this point in Figure 4, where we examine two approaches for estimating the "Blocks" signal in Laplacian noise. We compare our Bayesian approach using an IMM signal prior[2] and IMM noise priors (one at each scale) to Donoho's state-of-the-art SureShrink method [1], which is based on a Gaussian noise assumption. It is clear from the Figure that accurate modeling of the wavelet-domain noise leads to reduced mean-squared error and improved visual quality in the de-noised signal. The IMM noise priors lead to similar improvements over empirical Bayesian estimation with an iid Gaussian noise model.

Standard de-noising techniques [1] estimate the signal wavelet coefficients θ_i by thresholding the noisy wavelet coefficients w_i. Our Bayesian approach leads to threshold-like nonlinearities that vary across scale. For the Laplacian noise example, Figure 5 shows how these nonlinearities evolve as a function of scale, adjusting to match signal and noise properties such as heavier-tailed noise in the finer scales.

5 Conclusions

The wavelet transforms of real-world signals and images have residual structure that can be used to improve upon algorithms that process wavelet coefficients independently according to iid signal and/or iid Gaussian noise assumptions. In this paper, we have modeled the dependencies between wavelet coefficients that stem from the secondary properties of the wavelet transform. We can interpret our approach in the following way: The wavelet transform "almost decorrelates" the signal, removing all but the most local dependencies for the probabilistic graph model to handle. It is the fact that the wavelet transform can almost decorrelate so many signals that makes our approach feasible.

We feel that the graph-theoretic framework presented here could serve as a powerful new tool for wavelet-based statistical signal and image processing, with applications in signal estimation, detection, classification, compression, and even synthesis. A key to future work is tapping into the knowledge base that has already accumulated in statistics, speech recognition, artificial intelligence, and related fields.

References

[1] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

[2] J.-C. Pesquet, H. Krim, and E. Hamman, "Bayesian approach to best basis selection," in *IEEE Int. Conf. on Acoust., Speech, Signal Proc. — ICASSP '96*, (Atlanta), pp. 2634–2637, 1996.

[3] H. Chapman, E. Kolaczyk, and E. McCulloch, "Signal de-noising using adaptive Bayesian wavelet shrinkage," in *Proc. IEEE-SP Int. Symp. Time-Frequency and Time-Scale Analysis*, (Paris), pp. 225–228, June 1996.

[4] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3445–3462, Dec. 1993.

[5] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[6] P. Smyth, D. Heckerman, and M. Jordan, "Probabilistic independence networks for hidden Markov probability models," *Neural Comp.*, vol. 9, no. 1, To appear.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann, 1988.

[8] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Hidden Markov models for wavelet-based signal processing," Technical report #9612, Dept. ECE, Rice Univ., 1996.

[9] R. Redner and H. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, pp. 195–239, Apr. 1994.

[10] O. Ronen, J. Rohlicek, and M. Ostendorf, "Parameter estimation of dependence tree models using the EM algorithm," *IEEE Signal Proc. Lett.*, vol. 2, pp. 157–159, Aug. 1995.

[11] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Hidden Markov models for wavelet-based signal processing," in *Proc. 30th Asilomar Conf.*, (Pacific Grov, CA), Nov. 1996.

[2] Our IMM signal prior is a wavelet-Markov prior that assumes independence between signal wavelet coefficients. The set-up thus focuses on gains from improved noise modeling, rather than improved signal modeling, which was explored in [11].