WAVELET-BASED DENOISING USING HIDDEN MARKOV MODELS

M. Jaber Borran and Robert D. Nowak

ECE Department, MS-366
Rice University
Houston, TX 77005-1892

ABSTRACT

Hidden Markov models have been used in a wide variety of wavelet-based statistical signal processing applications. Typically, Gaussian mixture distributions are used to model the wavelet coefficients and the correlation between the magnitudes of the wavelet coefficients within each scale and/or across the scales is captured by a Markov tree imposed on the (hidden) states of the mixture. This paper investigates correlations directly among the wavelet coefficient *amplitudes* (sign$\times$magnitude), instead of magnitudes alone. Our theoretical analysis shows that the coefficients display significant correlations in sign as well as magnitude, especially near strong edges. We propose a new wavelet-based HMM structure based on mixtures of one-sided exponential densities that exploits both sign and magnitude correlations. We also investigate the application of this for denoising the signals corrupted by additive white Gaussian noise. Using some examples with standard test signals, we show that our new method can achieve better mean squared error, and the resulting denoised signals are generally much smoother.

1. INTRODUCTION

In most wavelet-based statistical signal processing techniques, the wavelet coefficients are assumed to be either independent or jointly Gaussian. This assumption is unrealistic for many real-world signals. Non-Gaussian statistics of the wavelet coefficients are considered in [1, 2]. Statistical dependencies between the wavelet coefficients are discussed in [1, 3], and the wavelet-based hidden Markov model (HMM) is proposed for statistical signal processing and analysis. The wavelet-based HMM captures correlations between the magnitudes of neighboring (in space/time) wavelet coefficients across scales.

Despite the success of the wavelet-based HMM in denoising applications [1], it does have one significant deficiency, which we address in this paper. Our theoretical analysis in Section 3 shows that there is a strong correlation in signs as well as the magnitudes of coefficients. Sign (or phase) correlation is not considered in the original wavelet-based HMM. We investigate a new model that account for both sign and magnitude correlations and apply the model to signal denoising. It is shown that the new model generally performs as well as or better than the wavelet-based HMM of [1]. In some cases, the improvement is quite significant.

The paper is organized as follows. In Section 2, we briefly review the wavelet-based HMM's. In Section 3, first we propose a new probabilistic model for the individual wavelet coefficients in which, instead of Gaussian distributions, one-sided exponential distributions are used as the components of the mixture distributions. We then use a hidden Markov tree model to capture the dependencies between the magnitudes and signs of the wavelet coefficients in adjacent scales. In Section 4, we devise a novel Expectation-Maximization (EM) algorithm to train the HMT model using a noisy observation of the signal. Since our new model involves non-Gaussian component densities, the EM algorithm is significantly different than that devised in [1]. In Section 5, we employ Maximum *A Posteriori* (MAP) and Conditional Mean (CM) estimators based on our new model for signal denoising. Standard test signals are used to demonstrate the performance of our new method, and the results show our new approach generally performs as well or better than the wavelet-based HMM in [1]. Finally we draw some conclusions in Section 6.

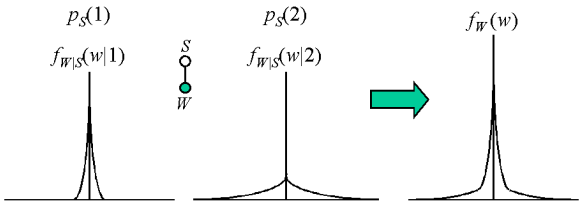


Fig. 1. Two-state, zero-mean Gaussian mixture model

2. REVIEW OF WAVELET-BASED HMM'S

In [1], a framework for statistical signal processing was developed in which the non-Gaussian statistics and statistical dependencies of the wavelet coefficients encountered in the real-world signals are concisely modeled using wavelet-domain HMM's [4]. In the design of these HMM's, *primary* and *secondary properties* of the wavelet transform are taken into account. The *primary properties* of the wavelet transform are *locality*, *multiresolution*, and *compression*. The last property states that the wavelet transforms of real-world signals tend to be sparse. In order to take into account this property, in [1] the a mixture Gaussian distribution is proposed to model wavelet coefficients. This model, shown in Fig. 1, consists of two Gaussian distributions with zero mean and two different variances, each one selected according to some probability mass function assigned to the two states, which serve as indicators for the two component densities.

The *secondary properties* of the wavelet transform are *clustering* and *persistence*. The *clustering* property states that if a particular wavelet coefficient is large/small, then adjacent coefficients are very likely to also be large/small. The *persistence* property states that large/small values of wavelet coefficients tend to propagate across scales. In order to take into account these two properties,

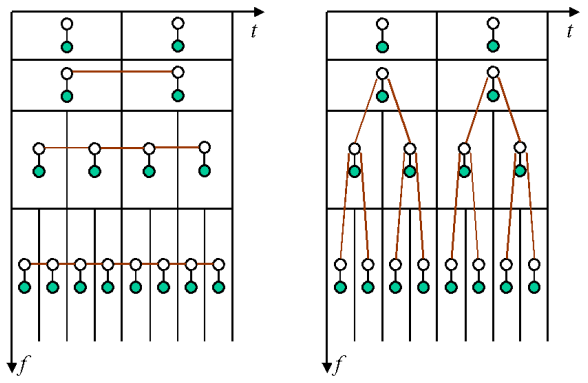


Fig. 2. Hidden Markov chain (left) and hidden Markov tree (right) models

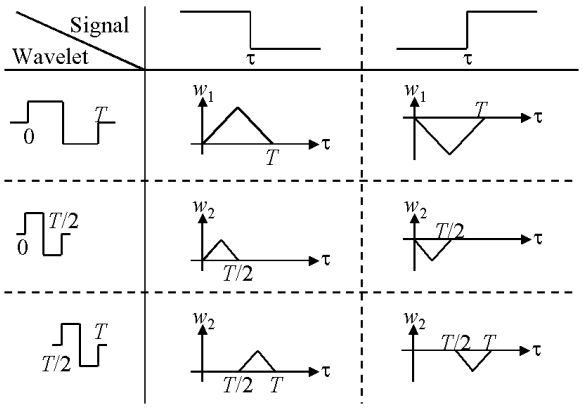


Fig. 3. Wavelet coefficients for adjacent scales

in [1] the use of a probabilistic graph that links the wavelet state variables either across time using a chain, or across scale using a tree, is suggested. These models are called *Hidden Markov Chain Model* and *Hidden Markov Tree (HMT) Model*, respectively, and are shown in Fig. 2.

Once the model structure is specified, the *Expectation Maximization* (EM) algorithm can be used to estimate the parameters of the model [1]. Using this algorithm, the multi-dimensional maximum likelihood estimation problem can be decomposed into several one-dimensional problems with an iterative nature. This way, the complexity of the maximum likelihood estimator is strikingly reduced, yet acceptable performance can be achieved.

3. NEW MODEL

While the original wavelet-based HMT [1] adequately captures the persistence of large/small magnitude (or energy) coefficients across scale, it does not reflect correlation in signs of the coefficients. However, it turns out that the signs of wavelet coefficients can be strongly related, as demonstrated in the following analysis. In Fig. 3, assuming that the signal under consideration has only one rising/falling edge in the support of a wavelet at a particular scale, the Haar wavelet coefficients in that scale and the next scale are compared. According to this figure, it can be observed that with the above assumption, the signs of the wavelet coefficients in these two neighboring scales are highly correlated. In fact, if the sign of the wavelet coefficient in the coarser resolution is positive/negative, then so is the sign of the wavelet coefficient in the finer resolution (or the coefficient is zero). The high degree of correlation between the signs of the wavelet coefficients is a motivation for us to consider a mixture distribution for the wavelet coefficients that consists of one-sided distributions, e.g., exponential densities. This way, we will be able to capture the correlation between the signs of the wavelet coefficients in the adjacent scales, and according to the observed high correlation, achieve better performance in denoising the noisy signals.

Fig. 4 shows a mixture distribution consisting of four one-sided exponential distributions. The conditional probability density functions for the wavelet coefficients at node i, given the state variable are given as follows:

$$f_{W_i|S_i}(w|m) = \begin{cases} \lambda_{i,m} e^{-\lambda_{i,m} w} & w \geq 0 \\ 0 & w < 0 \end{cases}, \; m \text{ even}$$
$$\begin{cases} 0 & w > 0 \\ \lambda_{i,m} e^{\lambda_{i,m} w} & w \leq 0 \end{cases}, \; m \text{ odd} \qquad (1)$$

As mentioned earlier, we will use a hidden Markov tree model (shown in Fig. 2) for the wavelet transform. Assuming that we are using a full-scale wavelet transform for an N-point signal, this model is parameterized by the following parameters:

- number of the states at each node, M
- pmf of the states at root node, $p_{S_1}(m)$, for $m = 1, \cdots, M$
- transition probabilities $\epsilon^{m,r}_{i,\rho(i)}$ (probability of node i being in state m given that its parent, $\rho(i)$, is in state r), for $m, r = 1, \cdots, M$ and $i = 1, \cdots, N-1$
- conditional pdf's for wavelet coefficients given the state, $f_{i,m}(w) = f_{W_i|S_i}(w|m)$, (or $\lambda_{i,m}$), for $m = 1, \cdots, M$ and $i = 1, \cdots, N-1$

We will collect these parameters in a model parameter vector, θ.

In the next section we will explain a method for estimating these parameters (except for M, the number of states, which is usually chosen to be 2 or 4). It should be emphasized that in Fig. 3 it is assumed that the signal under consideration has only one rising/falling edge in the support of the wavelet in the coarser resolution. If this condition is satisfied, then the wavelet coefficients in the higher resolutions are highly correlated (in their sign) with the wavelet coefficient in this resolution. Therefore, because edges or singularities tend to be somewhat isolated signal features, we expect that the mentioned correlation will be most significant at higher resolutions (finer scales).

4. TRAINING THE HMT MODEL

As mentioned earlier, we will use the noisy observations to estimate the parameters of the model. In order to do this, we first need to find the conditional probability density functions of the noisy wavelet coefficients given the state parameter at the specific

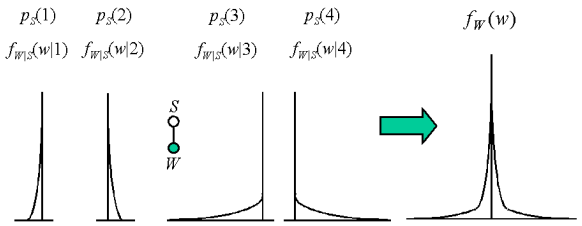


Fig. 4. Four-state, one-sided exponential mixture model

node. Assuming that the original signal is corrupted by an additive white Gaussian noise with *known* variance σ^2, and noting that the wavelet transform is orthonormal, the i-th noisy wavelet coefficient, y_i, can be written as follows:

$$y_i = w_i + n_i, \quad n_i \sim iid\left(\mathcal{N}(0,\sigma^2)\right), \tag{2}$$

where w_i is the wavelet coefficient of the (noise-free) signal. With this assumption, and using the conditional probability density functions for the wavelet coefficients of the original signal given in the previous section, it can be easily shown that the (state-)conditional probability density functions for the noisy wavelet coefficient at node i are given as follows:

$$f_{Y_i|S_i}(y|m) = \begin{cases} \lambda_{i,m} e^{-\lambda_{i,m} y + \frac{1}{2}\lambda_{i,m}^2 \sigma^2} Q\left(\sigma\lambda_{i,m} - \frac{y}{\sigma}\right), & m \text{ even} \\ \lambda_{i,m} e^{\lambda_{i,m} y + \frac{1}{2}\lambda_{i,m}^2 \sigma^2} Q\left(\sigma\lambda_{i,m} + \frac{y}{\sigma}\right), & m \text{ odd} \end{cases} \tag{3}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$. Now, denoting by $\mathbf{y}$, $\mathbf{s}$, and θ, the vectors of observed noisy wavelet coefficients, hidden states, and model parameters, respectively, and using the Maximum Likelihood (ML) criterion, the parameter estimation problem can be formulated as the following optimization problem:

$$\hat{\theta} = \arg\max_{\theta} \log f_{\mathbf{Y}}(\mathbf{y}|\theta). \tag{4}$$

In general, this problem is a very complicated and difficult one to solve, because in this estimation process, we are also characterizing the states $\mathbf{S}$, of the nodes, which are unobserved. However, given the values of the states, the ML estimation of the parameter vector is much simpler. Therefore, we use the iterative EM approach [5, 6], which jointly estimates both the model parameters θ, and probabilities for the hidden states $\mathbf{S}$, given the observed noisy wavelet coefficients, $\mathbf{Y}$.

First, we define the set of complete data, $\mathbf{X}$, as $\mathbf{X} = (\mathbf{Y}, \mathbf{S})$. Note that the likelihood function for the complete data can be expressed in terms of the conditional pdf of $\mathbf{Y}$given $\mathbf{S}$, and the pmf of $\mathbf{S}$, given the parameter vector θ, as follows:

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = f_{\mathbf{YS}}(\mathbf{y},\mathbf{s}|\theta) = f_{\mathbf{Y}}(\mathbf{y}|\mathbf{s},\theta) f_{\mathbf{S}}(\mathbf{s}|\theta), \tag{5}$$

where

$$f_{\mathbf{Y}}(\mathbf{y}|\mathbf{s},\theta) = \prod_{i=1}^{N-1} f_{Y_i|S_i}(y_i|s_i), \tag{6}$$

and

$$f_{\mathbf{S}}(\mathbf{s}|\theta) = p_{S_1}(s_1) \prod_{i=2}^{N-1} \epsilon_{i,\rho(i)}^{s_i, s_{\rho(i)}}. \tag{7}$$

Then, instead of maximizing the log-likelihood function of $\mathbf{Y}$, we maximize the log-likelihood function of $\mathbf{X}$. But since the states are unknown, we take the expected value of the log-likelihood with respect to the random variable $\mathbf{S}$, and since we need to know the parameter vector θ to calculate this expectation, we use the current estimate for the parameter vector, θ^l, in calculating the expectation. This results in an iterative algorithm with two steps in each iteration:

- *E-step*: Calculate $U(\theta,\theta^l) = E_{\mathbf{S}}\left\{\log f_{\mathbf{X}}(\mathbf{x}|\theta)|\mathbf{y},\theta^l\right\}$
- *M-step*: Find $\theta^{(l+1)} = \arg\max_\theta U(\theta,\theta^l)$

For our estimation problem, the *E-step* can be rewritten as follows:

$$U(\theta,\theta^l) = \sum_{\mathbf{S}\in\{1,\cdots,M\}^{N-1}} f_{\mathbf{S}|\mathbf{Y}}(\mathbf{s}|\mathbf{y},\theta^l) \log f_{\mathbf{YS}}(\mathbf{y},\mathbf{s}|\theta), \tag{8}$$

or

$$U(\theta,\theta^l) = \sum_{\mathbf{S}\in\{1,\cdots,M\}^{N-1}} f_{\mathbf{S}|\mathbf{Y}}(\mathbf{s}|\mathbf{y},\theta^l) \times \left[\log p_{S_1}(s_1) + \sum_{i=2}^{N-1} \log \epsilon_{i,\rho(i)}^{s_i,s_{\rho(i)}} + \sum_{i=1}^{N-1} \log f_{Y_i|S_i}(y_i|s_i)\right]. \tag{9}$$

For calculating the U function in (9), we need to know the *a posteriori* state probabilities. These probabilities are calculated using the *Upward-Downward* algorithm explained in [1] and [6].

In the *M-step*, the U function, calculated above, is maximized over the *a priori* pmf of the root state, the state transition probabilities, and the parameters of the conditional pdf's. The root state pmf and the transition probabilities can be calculated using Lagrange multipliers, and the results are given in [1].

Maximizing the function U with respect to the parameters of the conditional pdf's, $\lambda_{i,m}$, reduces to the following maximization problem:

$$\lambda_{i,m} = \arg\max_{\lambda} \sum_{j\in[i]} \Pr[S_j = m|\mathbf{y},\theta^l] \log f_{Y_i|S_i}(y_i|m), \tag{10}$$

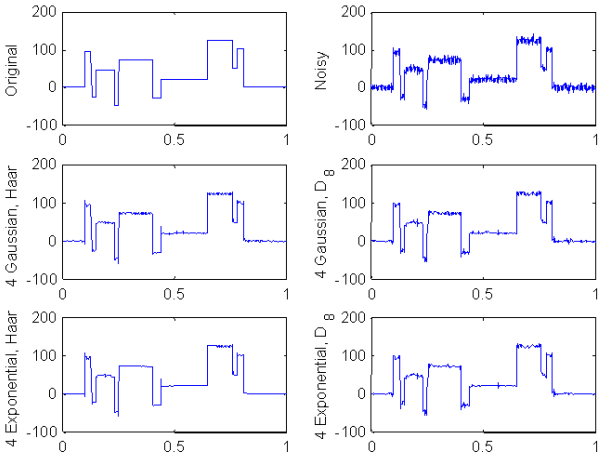
where $[i]$ denotes the set of indices of all nodes that are at the same scale as node i. Unlike the Gaussian mixture case (which admits a closed-form solution), in the case of mixture exponential distributions, this maximization problem has no analytic solution. However, the maximization above can be solved very efficiently using numerical methods (since it is simply a one-dimensional optimization problem).

As the result of the *M-step*, we will have a new and more reliable estimate for the parameter vector, and as we increase the number of the iterations, more and more reliable estimates can be achieved. A convergence criterion can be used to decide whether to continue the iterations or stop, in which case the result of the *M-step* of the last iteration is taken as the estimate of parameter vector.

5. DENOISING

In this section, assuming that we have the model parameter vector (or a good estimate of it), we will estimate the wavelet coefficients of the original signal from the noisy wavelet coefficients. It can be easily shown that the *a posteriori* probability density function of each (noise-free) wavelet coefficient can be written as follows (for s even):

$$f_{W|YS}(w|y,s) = \frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-\frac{1}{2}\sigma^2\lambda^2 + \lambda(y-w) - \frac{(y-w)^2}{2\sigma^2}}}{Q\left(\sigma\lambda - \frac{y}{\sigma}\right)}. \tag{11}$$



Init. MSE = 24.639723	4 mix, Haar	4 mix, D_8
Gaussian Mixture	3.078267	7.020152
Exponential Mixture	2.326472	7.030970

Fig. 5. Denoising the 'Blocks' test signal

We obtain a similar expression for the case s odd. The MAP estimate for w is given by

$$\hat{w} = \hat{w}_{\hat{s}}, \tag{12}$$

where

$$\hat{s} = \arg\max_{s} p_S(s) f_{W|YS}(\hat{w}_s|y,s), \tag{13}$$

$$\hat{w}_s = \arg\max_{w} f_{W|YS}(w|y,s) = \begin{cases} (y-\sigma^2\lambda)_+ & s \text{ even} \\ (y+\sigma^2\lambda)_- & s \text{ odd} \end{cases}, \tag{14}$$

and

$$(x)_+ = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}, \quad (x)_- = \begin{cases} x & x \leq 0 \\ 0 & x > 0 \end{cases}.$$

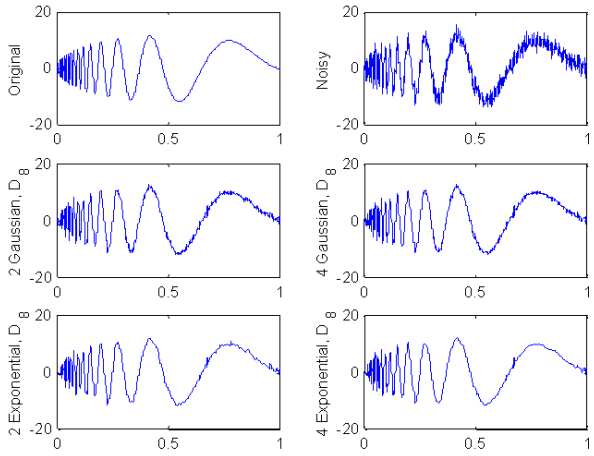
The conditional mean estimate for w is given by

$$\hat{w} = \sum_{s=1}^{M} p_S(s)\hat{w}_s, \tag{15}$$

where

$$\hat{w}_s = E\{w|y,s\} = \begin{cases} \frac{\sigma}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}\left(\sigma\lambda - \frac{y}{\sigma}\right)^2}}{Q\left(\sigma\lambda - \frac{y}{\sigma}\right)} + (y-\sigma^2\lambda) & s \text{ even} \\ \frac{\sigma}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}\left(\sigma\lambda + \frac{y}{\sigma}\right)^2}}{Q\left(\sigma\lambda + \frac{y}{\sigma}\right)} - (y+\sigma^2\lambda) & s \text{ odd} \end{cases} \tag{16}$$

Either of the above estimates can be used to find the denoised version of the wavelet coefficients from the noisy version, and then the denoised signal can be calculated as the inverse wavelet transform of the denoised wavelet coefficients. Figs. 5 and 6 compare the performance of the proposed algorithm with the method given in [1], in denoising the standard test signals of 'Blocks', and 'Doppler', respectively. As can be observed from these figures, in most of the cases the proposed algorithm has improved Mean Squared Error (MSE), and the resulting denoised signals are much smoother and have a significantly better visual quality with the same number of states and same wavelet.



Init. MSE = 2.429741	2 mix, D_8	4 mix, D_8
Gaussian Mixture	0.471568	0.417795
Exponential Mixture	0.426488	0.397808

Fig. 6. Denoising the 'Doppler' test signal

6. CONCLUSIONS

We observed that there is a high correlation between the signs of the wavelet coefficients of a signal in adjacent scales. We used one-sided distributions as components of a mixture distribution assigned for the individual wavelet coefficients, and then we used a hidden Markov tree model to capture the dependencies between the magnitudes and signs of the wavelet coefficients in adjacent scales. We used the iterative expectation-maximization algorithm to train the model directly from the noisy data. Using standard test signals, we showed that the proposed method achieves better MSE in denoising compared to the methods based on Gaussian mixture distributions with the same number of states and complexity, and the resulting denoised signals are generally much smoother.

7. REFERENCES

[1] R. G. Baraniuk M. S. Crouse, R. D. Nowak, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, April 1998.

[2] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Lecture Notes in Statistics*, pp. 291–308. Springer-Verlag, 1999.

[3] R. D. Nowak, "Multiscale hidden Markov models for Bayesian image analysis," in *Lecture Notes in Statistics*, pp. 243–265. Springer-Verlag, 1999.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1998.

[5] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, pp. 47–60, Nov. 1996.

[6] O. Ronen, J. R. Rohlicek, and M. Ostendorf, "Parameter estimation of dependence tree models using the EM algorithm," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 157–159, Aug. 1995.