Wavelet-Based Denoising Using Hidden Markov Models

ELEC-631 Course Project Report

Mohammad Jaber Borran

Abstract

Hidden Markov models have already been used for wavelet-based statistical signal processing. In these models, Gaussian mixture distributions are used for the wavelet coefficients and the correlation between the magnitudes of the wavelet coefficients within each scale and/or across the scales are tried to be captured using hidden Markov models. In this work, using one-sided distributions as the components of the mixture distributions for the individual wavelet coefficients in a hidden Markov tree model, we will try to capture the correlation between the signs of the wavelet coefficients across the scales. Then we will use this model for denoising the signals corrupted by additive white Gaussian noise. Using some examples with standard test signals, we will show that this method can achieve better mean squared error, and the resulting denoised signals are generally much smoother.

I. Introduction

In most of the 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 considered in [1], [3].

In [1] a new 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 Hidden Markov Models (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 use of a mixture Gaussian distribution is suggested. 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 these two states.

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, 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 is decided, in [1] the use of Expectation Maximization (EM) algorithm is suggested to estimate the parameters of the model. 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, and further, performance of the algorithm can be adjusted by choosing the right number of iterations.

In Section II, first we will suggest a new probabilistic model for the individual wavelet coefficients in which, instead of Gaussian distributions, one-sided distributions are used as the components of the mixture distributions. Then we will use a hidden Markov tree model to capture the dependencies between the magnitudes and signs of the wavelet coefficients in adjacent scales.

In Section III, we will explain a method to train the HMT model using a noisy observation of the signal. For this, we will use the EM algorithm, which results in an iterative method for maximizing the log-likelihood function of the unknown parameters given the observation and has an acceptable amount of complexity.

In Section IV, we will use Maximum A Posteriori (MAP) and Conditional Mean (CM) estimators to find the wavelet coefficients of the original signal from the wavelet coefficients of the noisy observations. Some sample signals will be used to show the performance of the method for denoising different signals using different wavelets. Finally we will bring the conclusions in Section V.

II. NEW MODEL

The persistence property of the wavelet transform suggests using an HMT and implies that it is very likely that the transition probability from the high/low variance state in the parent to the high/low variance state in the child turns out to be much bigger than the transition probability to the low/high variance state. These transition probabilities will model the correlation between the magnitudes of the wavelet coefficients in adjacent scales.

In Fig. 3, assuming that the signal under consideration has only one rising/falling edge in the support of a wavelet in some 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, so is the sign of the wavelet coefficient in the finer resolution (or the coefficient is zero). This high 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. exponentials. 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, given the state variable, at node $i$, are given as follows:

$$
\begin{align*}
\text{If } m \text{ is even: } f_{W_i|s_i}(w|m) &= \begin{cases} 
\lambda_{i,m}e^{-\lambda_{i,m}w} & w \geq 0 \\
0 & w < 0
\end{cases} \\
\text{If } m \text{ is odd: } f_{W_i|s_i}(w|m) &= \begin{cases} 
\lambda_{i,m}e^{\lambda_{i,m}w} & w \leq 0 \\
0 & w > 0
\end{cases}
\end{align*}
$$

(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 root node, $p_{S_0}(m)$, for $m = 1, \cdots, M$
- transition probabilities $e_{i,\rho(i)}^m$ (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, $\theta$.

In the next section we will explain a method for estimating these parameters (except for the first one, 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, we expect that the mentioned correlation will increase as we go from the lower resolutions to higher resolutions.
III. 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 node. Assuming that the original signal is corrupted by an additive white Gaussian noise with variance $\sigma^2$, and noting that the wavelet transform is orthonormal, the noisy wavelet coefficient at node $i$, $y_i$, can be written as follows:

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

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 conditional probability density functions for the noisy wavelet coefficient at node $i$ are given as follows:

$$f_{y_i|y_i}(y|i) = \begin{cases} 
\lambda_i m e^{-\lambda_i y + \frac{1}{2} \lambda_i^2} \frac{e^{-\frac{y^2}{2 \sigma^2}}}{\sqrt{2 \pi \sigma^2}} & \text{if } m \text{ is even} \\
\lambda_i m e^{\lambda_i y + \frac{1}{2} \lambda_i^2} \frac{e^{-\frac{y^2}{2 \sigma^2}}}{\sqrt{2 \pi \sigma^2}} & \text{if } m \text{ is odd}
\end{cases}$$

Now, if we denote by $Y$, $S$, and $\theta$, 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:

$$\hat{\theta} = \arg \max_{\theta} \log f_Y(y|\theta).$$

In general, this problem is a very complicated and difficult to solve because, in this estimation process, we are also characterizing the states $S$ 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] and [6], which jointly estimates
both the model parameters \( \theta \), and probabilities for the hidden states \( S \), given the observed noisy wavelet coefficients, \( Y \).

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

\[
f_X(x|\theta) = f_{YS}(y,s|\theta) = f_Y(y|s,\theta)f_S(s|\theta),
\]  

where

\[
f_Y(y|s,\theta) = \prod_{i=1}^{N-1} f_{Y_i|S_i}(y_i|s_i),
\]  

and

\[
f_S(s|\theta) = p_{S_1}(s_1) \prod_{i=2}^{N-1} \epsilon_{i,\rho(i)}^{s_i,s_{\rho(i)}}.
\]

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

- **E-step**: Calculate \( U(\theta, \theta^t) = E_S \{ \log f_X(x|\theta)|y, \theta^t \} \)

- **M-step**: Find \( \theta^{(t+1)} = \arg \max_\theta U(\theta, \theta^t) \)

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

\[
U(\theta, \theta^t) = \sum_{S \in \{1,\cdots,M\}^{N-1}} f_{S|Y}(s|y, \theta^t) \log f_{YS}(y,s|\theta),
\]

or

\[
U(\theta, \theta^t) = \sum_{S \in \{1,\cdots,M\}^{N-1}} f_{S|Y}(s|y, \theta^t) \left[ \log p_{S_1}(s_1) + \sum_{i=2}^{N-1} \log \epsilon_{i,\rho(i)}^{s_i,s_{\rho(i)}} + \sum_{i=1}^{N-1} \log f_{Y_i|S_i}(y_i|s_i) \right].
\]
For calculating the $U$ function in (9), we need to know the state \textit{a posteriori} probabilities. These probabilities are calculated using the \textit{Upward-Downward} algorithm explained in [1] and [6].

At the $M$-step, the $U$ function calculated above, is maximized over the root state \textit{a priori} pmf, 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:

$$
\lambda_{i,m} = \arg \max_{\lambda} \sum_{j \in [i]} \Pr[S_j = m | y, \theta] \log f_Y | S_i (y|m),
$$

where $[i]$ denotes the set of indices of all nodes that are at the same scale as node $i$ is. For the case of mixture exponential distributions, this maximization problem has no analytic solution, however, it can be solved using numerical methods.

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. Some convergence criterion can be used to decide whether to continue the iterations or stop and consider the result of the $M$-step of the last iteration as the estimate of parameter vector.

\section*{IV. Denoising}

In this section, assuming that we have the model parameter vector (or a good estimate of that), we will try to find the wavelet coefficients of the original signal from the noisy wavelet coefficients. It can be easily shown that the \textit{a posteriori} probability density function of the original wavelet coefficients can be written as follows (for $s$ even):

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

The MAP estimate for $w$ is given by

$$
\hat{w} = \hat{w}_s,
$$

where

$$
\hat{s} = \arg \max_s p_s (s) f_{W|Y} (\hat{w}_s | y, s),
$$

and

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

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

$$
\hat{w} = \sum_{s=1}^{M} p_s (s) \hat{w}_s,
$$
where
\[
\hat{w}_s = E\{w|y, s\} = \begin{cases} 
\frac{\sigma}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}((\sigma \lambda - \frac{s}{2})^2)}}{Q(\sigma \lambda - \frac{s}{2})} + (y - \sigma^2 \lambda) & s \text{ even} \\
\frac{\sigma}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}((\sigma \lambda + \frac{s}{2})^2)}}{Q(\sigma \lambda + \frac{s}{2})} - (y + \sigma^2 \lambda) & s \text{ odd}
\end{cases}.
\]

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, 6, and 7 compare the performance of the proposed algorithm with the method given in [1], in denoising the standard test signals of 'Blocks', 'Doppler', and 'HeaviSine', 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.

V. CONCLUSIONS

We observed that there is a high correlation between the signs of 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 coefficients of the wavelet coefficients in adjacent scales. We used the lower complexity iterative expectations maximization algorithm to train the model with the noisy data. Using some example standard test signals, we showed that the proposed method achieves better MSE in denoising compared to the methods based on two-sided mixture distributions with the same number of states and complexity, and the resulting denoised signals are generally much smoother.

REFERENCES

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
Fig. 6. Denoising the 'Doppler' test signal

<table>
<thead>
<tr>
<th>Init. MSE = 2.429741</th>
<th>2 mix, $D_8$</th>
<th>4 mix, $D_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Mixture</td>
<td>0.471568</td>
<td>0.417795</td>
</tr>
<tr>
<td>Exponential Mixture</td>
<td>0.426488</td>
<td>0.397808</td>
</tr>
</tbody>
</table>
Init. MSE = 92.907059  2 mix, D₄  4 mix, D₈
Gaussian Mixture     8.442306  7.873508
Exponential Mixture  8.394187  7.862579

Fig. 7. Denoising the 'HeaviSine' test signal