SIMPLIFIED WAVELET-DOMAIN HIDDEN MARKOV MODELS USING CONTEXTS

Matthew S. Crouse and Richard G. Baraniuk

Rice University
Department of Electrical and Computer Engineering
Houston, Texas 77005

ABSTRACT

Wavelet-domain Hidden Markov Models (HMMs) are a potent new tool for modeling the statistical properties of wavelet transforms. In addition to characterizing the statistics of individual wavelet coefficients, HMMs capture the salient interactions between wavelet coefficients. However, as we model an increasing number of wavelet coefficient interactions, HMM-based signal processing becomes increasingly complicated. In this paper, we propose a new approach to HMMs based on the notion of *context*. By modeling wavelet coefficient inter-dependencies via contexts, we retain the approximation capabilities of HMMs, yet substantially reduce their complexity. To illustrate the power of this approach, we develop new algorithms for signal estimation and for efficient synthesis of nonGaussian, long-range-dependent network traffic.

1. INTRODUCTION

Wavelets present a powerful alternative to classical time-domain and frequency-domain approaches to statistical signal processing. In many instances, wavelets provide a compact, easy-to-model signal representation [1, 2].

For statistical applications ranging from compression to estimation to detection, the key to successful wavelet-based algorithms is an accurate joint probability model for the wavelet coefficients of our signals of interest. A complete model for the joint probability density function $f_{\mathbf{W}}(\mathbf{w})$, with $\mathbf{w}$ the vector of wavelet coefficients, is one possibility. However, such a characterization is intractable in practice, from both a computation and a robust estimation viewpoint. At the other extreme, modeling the wavelet coefficients as statistically independent, with $f_{\mathbf{W}}(\mathbf{w}) = \prod_i f_{W_i}(w_i)$, is simple but disregards the inter-coefficient probabilistic dependencies. To strike a balance between these two extremes, we must model the key wavelet coefficient dependencies, and only the key dependencies.

By design, wavelet-domain Hidden Markov models (HMMs) focus on the key wavelet coefficient dependencies, learning them via maximum-likelihood-based training [3, 4]. Hence, HMMs provide a natural setting for exploiting the structure inherent in real-world signals and images for signal estimation, detection, classification, prediction and filtering, and synthesis.

In this paper, we propose a new wavelet-domain signal modeling framework based on *contextual HMMs*. Contexts provide flexible conditional probability models for efficiently learning and expressing the dependencies in wavelet transforms. Before we develop these new models, we sketch some background on wavelets, mixture models, and wavelet-domain HMMs.

2. BACKGROUND

2.1 The Wavelet Transform. The discrete wavelet transform (DWT) represents a one-dimensional signal $z(t)$ in terms of shifted versions of a lowpass scaling function $\phi(t)$ and shifted and dilated versions of a prototype bandpass wavelet function $\psi(t)$ [5]. For special choices of the wavelet and scaling functions the atoms

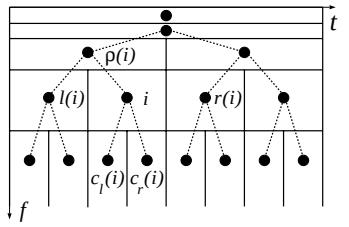

Figure 1: *Tiling of the time-frequency plane by the atoms of the wavelet transform. Each box depicts the idealized support of a scaling atom ϕ_i (top row) or a wavelet atom ψ_i (other rows) in time-frequency; the solid dot at the center corresponds to the scaling coefficient u_i or wavelet coefficient w_i. The figures also illustrates our tree notation for indexing neighboring coefficients.*

$\psi_{J,K}(t) \equiv 2^{-J/2}\,\psi\big(2^{-J}t - K\big)$, $\phi_{J,K}(t) \equiv 2^{-J}\,\phi\big(2^{-J}t - K\big)$, $J, K \in \mathbb{Z}$, form an orthonormal basis, and we have the signal representation [5]

$$z(t) \;=\; \sum_K u_{J_0,K}\,\phi_{J_0,K}(t) \;+\; \sum_{J=-\infty}^{J_0} \sum_K w_{J,K}\,\psi_{J,K}(t), \quad (1)$$

with $u_{J,K} \equiv \int z(t)\,\phi^*_{J,K}(t)\,dt$ and $w_{J,K} \equiv \int z(t)\,\psi^*_{J,K}(t)\,dt$.

The *wavelet coefficient* $w_{J,K}$ measures the signal content around time $2^J K$ and frequency $2^{-J} f_0$. The scaling coefficient $u_{J,K}$ measures the local mean around time $2^J K$. The DWT (1) employs scaling coefficients only at scale J_0; scaling coefficients at scales $J < J_0$ represent higher resolution approximations to the signal. Any filterbank or lifting DWT implementation produces all of the scaling coefficients $u_{J,K}$, $J < J_0$ as a natural byproduct [5].

To keep the notation manageable in the sequel, we will adopt an abstract index system for the DWT coefficients: $u_{J,K} \to u_i$, $w_{J,K} \to w_i$, with $J(i)$ the scale of the coefficient i. We will also use $\mathbf{w}$ to denote the vector of all wavelet coefficients.

The DWT has a natural interpretation in terms of a tree structure in the time-frequency domain (see Figure 1). In order to describe the relationships between wavelet coefficients, we will use standard tree notation for the parent $\rho(i)$, left $l(i)$ and right $r(i)$ neighbors, and left $c_l(i)$ and right $c_r(i)$ children of a node i.[1]

2.2 Gaussian Mixture Models. The DWTs of many real-world signals tend to be sparse, with just a few non-zero coefficients containing most of the signal energy [2]. Hence, the marginal density[2] $f_{W_i}(w_i)$ of each wavelet coefficient is typically described by a peaky (at $w_i = 0$) and heavy-tailed nonGaussian density.

Such densities are well approximated by *Gaussian mixture models* [6]. To each wavelet coefficient W_i, we associate a discrete hidden state S_i that takes on values $m = 1, \ldots, M$ with pmf $p_{S_i}(m)$.

This work was supported by NSF, grant no. MIP–9457438, by ONR, grant no. N00014–95–1–0849, and by DARPA, through AFOSR grant no. F49620-97-1-0513. Email: mcrouse@rice.edu, richb@rice.edu; Web: http://www.dsp.rice.edu/

[1] For clarity, we will assume throughout this paper that the length L of the signal is a power of two and furthermore that we take the maximum number of scales $J = \log_2 L$ in the DWT. However, all results extend to signals of arbitrary length, as well as to DWTs with fewer than the maximum possible number of scales (in which case, we have a forest of wavelet trees [4]).

[2] We will use $p_S(s)$ to denote the probability mass function (pmf) of the discrete random variable S and $f_W(w)$ to denote the probability density function (pdf) of the continuous random variable W.

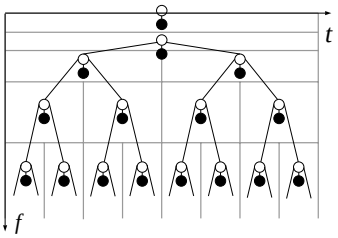

Figure 2: *Statistical models for the wavelet transform. We model each coefficient as a Gaussian mixture with a hidden state variable. Each black node represents a continuous wavelet coefficient W_i. Each white node represents the hidden mixture state variable S_i. Connecting the states vertically across scale yields the Hidden Markov Tree (HMT) model. Removing these links yields the Independent Mixture (IM) model.*

Conditioned on $S_i = m$, W_i is Gaussian with mean $\mu_{i,m}$ and variance $\sigma^2_{i,m}$. Thus, its overall pdf is given by

$$f_{W_i}(w_i) \;=\; \sum_{m=1}^{M} p_{S_i}(m)\, f_{W_i|S_i}(w_i|S_i = m). \qquad (2)$$

To generate a realization of W_i using the mixture model, we first draw a state value s_i according to $p_{S_i}(s_i)$ and then draw an observation w_i according to $f_{W_i|S_i}(w_i|S_i = s_i)$.

2.3 Hidden Markov Models. One simple approach to approximating the joint density $f_{\mathbf{W}}(\mathbf{w})$ would treat the wavelet coefficients as independent Gaussian mixtures. The result — the Independent Mixture (IM) model — has proven useful for signal estimation applications [6]. The primary motivation for this model lies in the fact that the DWT acts as an approximate Karhunen-Loève transform for a wide class of signals, and therefore the wavelet coefficients are approximately decorrelated.

However, the wavelet coefficients of real-world signals are not statistically independent in general. For instance, neighboring wavelet coefficients are often highly dependent — large/small coefficient values tend to propagate both within and across scales, creating clusters of large/small coefficients [3, 4].

Wavelet-domain Hidden Markov models (HMMs) are multidimensional mixture models in which the hidden states have a Markov dependency structure. The idea is to capture the dependencies in the wavelet coefficients through their hidden states. For example, the Hidden Markov Tree (HMT) model places a tree structure on the hidden states to capture wavelet dependencies across scale (See Figure 2) [3,4]. The HMT model is specified via the mixture parameters $\mu_{i,m}, \sigma^2_{i,m}$ and transition probabilities $p_{S_i|S_{\rho(i)}}(m|n)$.

Before we process signals using a wavelet-domain HMM, we first must train the model to capture the wavelet-domain properties of the signals of interest. That is, we determine the wavelet-domain HMM parameters that best characterize our observed wavelet coefficients. This standard HMM training problem can be efficiently accomplished (in linear time per iteration) using the iterative Expectation Maximization (EM) algorithm [7].[3]

Although the HMT model is powerful and relatively simple, in certain applications it is crucial to model more and different dependencies between the wavelet coefficients (such as across time and across scale simultaneously). More sophisticated dependency structures for the hidden states can be formulated using the theory of probabilistics graphs [4, 8], but the analysis and training of more complicated HMMs becomes extremely difficult [8]. For example, graphs with links that form cycles cannot be modeled using transition probabilities due to lack of a causal direction.

[3] EM algorithm intuition: If the values of the states S_i were known, then maximum-likelihood parameter estimation would be simple. Therefore, we iterate between estimating the probabilities for the states (Expectation) and updating our model given the state probabilities (Maximization). Under mild conditions, this iteration converges to a local maximum of the likelihood function.

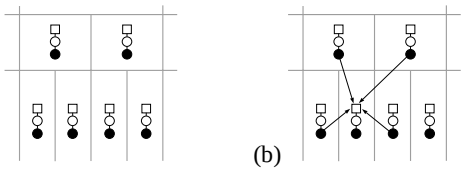

Figure 3: *Context-based models for the DWT. (a) To each wavelet coefficient-hidden state pair (W_i, S_i), we augment a (square) context node $\mathbf{V}_i$. The context vector is a function of the other wavelet and scaling coefficients. (b) Example context $\mathbf{V}_i$ formed using four wavelet coefficients neighbouring W_i.*

3. CONTEXT-BASED HIDDEN MARKOV MODELS

In this paper, we will use *contexts* to efficiently incorporate dependencies into our HMMs. We define the context for W_i as a length-P vector $\mathbf{V}_i \equiv [V_{i,1}, V_{i,2}, \ldots, V_{i,P}]$ formed as a function of the wavelet or scaling coefficients (see Figure 3). We condition S_i on $\mathbf{V}_i$ to predict W_i. The idea is for $\mathbf{V}_i$ to provide supplementary information to the HMM, so that given the context, we can treat the wavelet coefficients as independent.

By conditioning (2) on $\mathbf{V}_i$ (with the added assumption that $\mathbf{V}_i$ and W_i are independent given S_i), we have the context-based mixture model for the wavelet coefficients:

$$f_{W_i|\mathbf{V}_i}(w|\mathbf{v}_i) = \sum_{m=1}^{M} p_{S_i|\mathbf{V}_i}(m|\mathbf{v}_i)\, f_{W_i|S_i}(w|S_i = m). \qquad (3)$$

In this case, the mixing probabilities depend on the value of the context $\mathbf{V}_i$. If $\mathbf{V}_i$ is highly correlated with W_i, then (3) will provide a much more accurate characterization of the distribution of W_i than (2). In practice we do not specify $p_{S_i|\mathbf{V}_i}(m|\mathbf{v}_i)$ directly, but rather specify $p_{\mathbf{V}_i|S_i}(\mathbf{v}|m)$ and apply Bayes rule[4]

$$p_{S_i|\mathbf{V}_i}(m|\mathbf{v}_i) = \frac{p_{S_i}(m)\; p_{\mathbf{V}_i|S_i}(\mathbf{v}_i|m)}{\sum_{m=1}^{M} p_{S_i}(m)\; p_{\mathbf{V}_i|S_i}(\mathbf{v}_i|m)}. \qquad (4)$$

Defining $\epsilon_{i,m} \equiv p_{S_i}(m)$ and $\alpha_{i,\mathbf{v},m} \equiv p_{\mathbf{V}_i|S_i}(\mathbf{v}_i|m)$, the context-based HMM (CHMM) is parameterized by the vector $\mathbf{\Theta} = \left\{\mu_{i,m},\ \sigma^2_{i,m},\ \epsilon_{i,m},\ \alpha_{i,\mathbf{v},m}\right\}$. Given an observation of wavelet data $\mathbf{w}$, we estimate $\mathbf{\Theta}$ using the EM algorithm below. When only a single signal observation is available, we make the standard assumption that the wavelet coefficients in each scale are identically distributed. Multiple signal observations, multiple wavelet trees, as well as models for the scaling coefficients, can be handled as in [4].

EM Algorithm for CHMMs

Initialize: Choose $\mathbf{\Theta}^0$ and set $I = 0$.

Expectation (E): Given $\mathbf{\Theta}^I$, calculate (Bayes rule)

$$p_{S_i|\mathbf{V}_i,W_i}(m|\mathbf{v}_i, w_i) \;=\; \frac{\epsilon_{i,m}\;\alpha_{i,\mathbf{v}_i,m}\; f_{W_i|S_i}(w_i|m)}{\sum_{m=1}^{M} \epsilon_{i,m}\;\alpha_{i,\mathbf{v}_i,m}\; f_{W_i|S_i}(w_i|m)}.$$

[4] Here, we assume that the context is discrete-valued. We can model a continuous-valued $\mathbf{V}_i$ as an M-component Gaussian mixture of its own, replacing $p_{\mathbf{V}_i|S_i}(\mathbf{v}_i|m)$ in (4) with $f_{\mathbf{V}_i|S_i}(\mathbf{v}_i|m)$. We will find this useful in Section 4.2.

Maximization (M): Compute the elements of $\boldsymbol{\Theta}^{I+1}$

$$\begin{aligned}
\epsilon_{i,m} &= \sum_{k \ s.t. \ J(k)=J(i)} p_{S_i|\mathbf{V}_i,w_i}(m|\mathbf{v}_i, w_i), \\
\mu_{i,m} &= \frac{1}{2^{J_0-J(i)}\epsilon_{i,m}} \times \\
&\quad \sum_{k \ s.t. \ J(k)=J(i)} w_k \ p_{S_k|\mathbf{V}_k,W_k}(m|\mathbf{v}_k, w_k), \\
\sigma^2_{i,m} &= \frac{1}{2^{J_0-J(i)}\epsilon_{i,m}} \times \\
&\quad \sum_{k \ s.t. \ J(k)=J(i)} (w_k - \mu_{i,m})^2 \ p_{S_k|\mathbf{V}_k,W_k}(m|\mathbf{v}_k, w_k), \\
\alpha_{i,\mathbf{v},m} &= \frac{1}{\epsilon_{i,m}} \sum_{k \ s.t. \ J(k)=J(i), \mathbf{v}_k=\mathbf{v}_i} p_{S_i|\mathbf{V}_i,W_i}(m|\mathbf{v}_i, w_i).
\end{aligned}$$

Iterate: Increment $I \to I+1$. Apply E and M until converged.

In contrast to the HMT E step [4], the CHMM E step is very straightforward. To ensure fast and robust training, we keep the number of free parameters in each context vector to a minimum.

4. APPLICATIONS

To illustrate the flexibility of the CHMM framework, we now apply these models to two distinctly different problems: signal denoising and synthesis of long-range-dependent data network traffic.

4.1 Denoising. DWT methods have proved remarkably successful for estimating signals corrupted by additive white Gaussian noise (WGN) [2–4, 6]. The superior results of HMT model denoising have demonstrated that significant performance gains can be achieved by exploiting dependencies between wavelet coefficients [4]. Using a CHMM, we seek similar gains, but with reduced complexity.

Since the orthogonal DWT of zero-mean WGN is again zero-mean WGN of the same power, the signal estimation problem can be posed in the wavelet domain as: Estimate the wavelet coefficients y_i of a signal given the noisy measurements $w_i = y_i + n_i$, with $\{n_i\}$ a WGN process of variance σ_n^2. As in [4], we adopt an "empirical" Bayesian approach and model the signal wavelet coefficients Y_i using a two-component Gaussian mixture ($M = 2$) with $\mu_{i,1} = \mu_{i,2} = 0$.

If we knew the hidden state S_i of Y_i, then the minimum-mean-squared-error (MMSE) estimate would be the conditional mean estimate of a Gaussian signal in Gaussian noise

$$E[Y_i|w_i, S_i = m] = \frac{\sigma^2_{i,m}}{\sigma^2_{i,m} + \sigma_n^2} \ w_i. \tag{5}$$

Given probability estimates for the hidden states S_i, we estimate Y_i as the conditional mean

$$E[Y_i|w_i, \mathbf{v}_i] = \sum_{m=1}^{2} p_{S_i|w_i,\mathbf{v}_i}(m|w_i, \mathbf{v}_i) \ E[Y_i|w_i, S_i = m]. \tag{6}$$

If Y_i is a mixture of zero-mean Gaussians, then W_i is also a mixture of zero-mean Gaussians — the addition of zero-mean independent Gaussian noise increases the variance of each mixture component by σ_n^2, but leaves the state S_i unaffected. Hence, we train our CHMM on the the noisy wavelet data $\mathbf{W}$ to estimate the hidden state probabilities of the signal $p_{S_i|w_i,\mathbf{V}_i}(m|w_i, \mathbf{v}_i)$ and (by subtracting σ_n^2) the signal mixture variances $\sigma^2_{i,m}$. We then calculate the estimates (6) and invert the DWT to obtain the denoised signal. (See [4] for more details on a similar approach based on the HMT model.)

Table 1: *Denoising results for Donoho and Johnstone's length-1024 test signals [2]. Noise variance* $\sigma_n^2 = 1$.

Method	Mean-squared error			
	Bumps	Blocks	Doppler	Heavisine
SureShrink [2]	0.683	0.222	0.228	0.095
Bayesian [6]	0.350	0.099	0.165	0.087
IM	0.335	0.105	0.170	0.080
HMT	0.268	0.079	0.132	0.081
Context 1	0.252	0.101	0.141	0.081
Context 2	0.249	0.099	0.141	0.079

What remains is to specify contexts that are simple, yet effective, for gleaning information on the hidden states. Two simple discrete contexts that exploit clustering of signal energy in the wavelet domain [4] illustrate our approach. Define q_i as the quantized value of the wavelet coefficient w_i: Set $q_i = 1$ if $|w_i|^2$ is greater than the average energy in its scale, otherwise, set $q_i = 0$. The first context contains quantized values of the neighboring wavelet coefficients

$$\mathbf{V}_i^{(1)} = \left[q_{\rho(i)}, \ q_{l(i)}, \ q_{r(i)}, \ q_{c_{l(i)}}, \ q_{c_{r(i)}}\right], \tag{7}$$

and thus conveys gross information about the size of the neighboring coefficients. Our intuition is that if $w_{\rho(i)}$ and $w_{c_{l(i)}}$ are large, then there is a good chance that w_i will be large as well. To encode such information ("large" vs. "small"), even crudely quantized information is sufficient. The second context combines elements of $\mathbf{V}_i^{(1)}$ using logical *or* operations "$\vee$"

$$\mathbf{V}_i^{(2)} = \left[q_{\rho(i)}, \ \ q_{l(i)} \vee q_{r(i)}, \ \ q_{c_l(i)} \vee q_{c_r(i)}\right]. \tag{8}$$

To further reduce complexity, we also assume that the context probabilities factor as $p_{\mathbf{V}_i|S_i}(\mathbf{v}_i|m) = \prod_{j=1}^{P} p_{V_{i,j}|S_i}(v_{i,j}|m)$.

In Table 1, we provide the MSE results for denoising Donoho and Johnstone's standard test signals [2] using CHMMs versus other state-of-the-art algorithms. Contexts 1 and 2 correspond to our proposed algorithm using the contexts defined in (7) and (8), respectively. Implementation details, such as the exact DWTs used, are provided in [4].

The key benchmarks for comparison are the IM and HMT models from [4]. IM denoising employs a mixture model that treats the signal wavelet coefficients as independent. Improvements over IM signify the context's ability to capture and exploit dependencies between coefficients. Overall, the MSE performance of the context-based approach is roughly comparable to the considerably more complicated HMT denoiser of [4].

4.2 Signal Synthesis. Recent studies have shown that data network traffic is statistically self-similar and exhibits the long-range dependence characteristic of slowly-decaying correlation functions [9]. These properties are difficult to model using classical traffic models involving Poisson or Markov processes. Complicating matters further is the fact that actual network inter-arrival times are nonGaussian, positive, and heavy-tailed [9]. Classical self-similar process models, such as fractional Brownian motion (fBm) can capture the long-range dependence of network traffic; however, fBm is a Gaussian process, and current methods for its synthesis are computationally intensive (up to $O(L^3)$ complexity for an L-point trace). New tools for analyzing and synthesizing very long traces of such data are important for network design and control, since classical models can severely overestimate network performance.

Our goal is to develop a fast wavelet-based synthesis algorithm consistent both with the long-range dependence and the positive, nonGaussian marginal statistics of network traffic. Our approach

will be to first train a CHMM on an actual traffic trace, and then synthesize artificial traffic with "equivalent" statistical properties. By characterizing how the wavelet coefficient variances change with scale, CHMMs can approximate the long-range dependence properties of the data. By using the Haar scaling coefficients as contexts, CHMMs can capture the positive, nonGaussian marginal properties of the traffic as we will show.

Using a Haar DWT [5], we will associate with each w_i ($w_{J,K}$ in the notation of Section 2.1) its corresponding scaling coefficient u_i ($u_{J,K}$ in the notation of Section 2.1). Since u_i corresponds to a local mean of the (positive) signal, we know that $u_i > 0$, $\forall i$. Moreover, since for the Haar DWT $u_{c_l(i)} = 2^{-1/2}(u_i + w_i)$ and $u_{c_r(i)} = 2^{1/2}(u_i - w_i)$, we must have $|w_i| < u_i$, $\forall i$.

Because of this clear dependence, we use the random variable $\mathbf{V}_i = U_i$ as the context for the random variable W_i. We model U_i as a Gaussian mixture, with the parameters $\widetilde{\mu}_{i,m}, \widetilde{\sigma}^2_{i,m}$ updated in the M step in a fashion similar to the updates for $\mu_{i,m}, \sigma^2_{i,m}$.

In essence, this procedure employs a mixture model to approximate the 2-d density for (U_i, W_i) and then uses the 2-d density to obtain a conditional density for W_i based on U_i. With enough mixture parameters, this approach in theory can approximate (U_i, W_i) to arbitrary precision, hence automatically learning the constraints $U_i > 0$ and $|W_i| < U_i$.

In practice, to simplify our modeling, we map the cone $U_i > 0$, $|W_i| < U_i$ to the plane through the invertible map $g : (U_i, W_i) \mapsto \big(\log(U_i),\ -\mathrm{sgn}(W_i)\,\log(1 - |W_i|/U_i)\big)$. By modeling $g(U_i, W_i)$ and then inverting to form (U_i, W_i), we automatically enforce the positivity constraints. To synthesize W_i given U_i, we map U_i to $\log(U_i)$, use it as a context to synthesize the transformed data, generate a realization, and then invert the map g to produce W_i.

To synthesize an entire wavelet transform $\mathbf{W}$, we work in "top-down" fashion starting from the root of the wavelet tree by synthesizing the single coarsest scale wavelet coefficient. (We assume its context, the global mean of the signal, is already specified.) We iterate down the tree using the fact that summing and differencing U_i and W_i provides the context information for synthesizing $W_{c_l(i)}$ and $W_{c_r(i)}$.

As a test, we trained the CHMM synthesis algorithm on a portion of the Bellcore Ethernet data (the first 10^6 arrivals of the day-long trace started August 29, 1989) [9]. The model was equipped with ten mixture-components ($M = 10$) at each wavelet scale. In Figure 4, we compare, over different time scales, a random realization from our synthesis algorithm with the actual data. In Figure 5, we illustrate the histogram fit that our synthesis algorithm achieves over different time scales.

As is evident from the Figures, CHMM synthesis captures both the marginal properties of the traffic and, because of the match over a number of time scales, the long-range dependence as well. For synthesis applications, CHMMs are both accurate and fast ($O(L)$ operations), demonstrating the power of the context-based framework.

5. CONCLUSIONS

CHMMs have a number of potential advantages over conventional HMMs for exploiting the wavelet-domain structure inherent in real-world signals. First, CHMMs allow the user to characterize dependencies that may be too complex or even downright impossible to model using standard HMMs. Second, although efficient algorithms exist for HMMs based on trees, for more complicated graph structures (such as 2-d HMMs for images), the training procedure can become intractable. CHMMs deal naturally with noncausal information, yet retain the simplicity of a causal model. The explanation lies in the fact a CHMM consists essentially of a series of local models, each with a small number of parameters, that can be trained independently. More traditional HMM models, on the other hand, adjust their parameters to optimize a complicated global objective function.

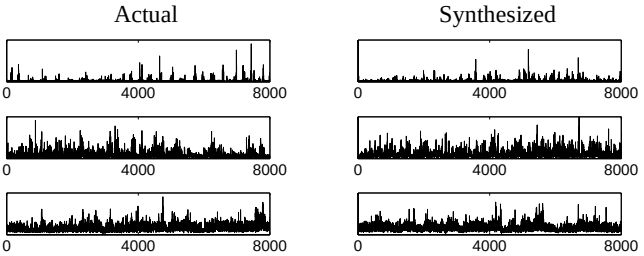

Figure 4: *Network data traffic synthesis via CHMM. Inter-arrival times as a function of packet group number plotted for (top) one, (middle) ten, and (bottom) one-hundred packets. The actual traces consist of approximately 10^6 packet arrivals, but only the inter-arrival times of the first groups of packets are shown.*

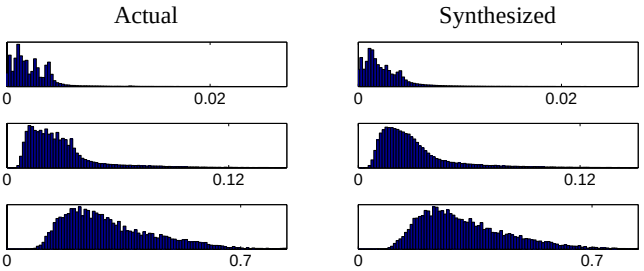

Figure 5: *Histograms of the inter-arrival times corresponding to the data from Figure 4 for groups of (top) one, (middle) ten, and (bottom) one-hundred packets.*

The primary disadvantage of the CHMM framework is that it lacks the feedback mechanism of more traditional HMMs that allow the model to propagate information from variables across the entire model, hence capturing dependencies from more than just neighboring wavelet coefficients. However, in many instances, we expect the convenience and efficiency of the context approach to outweigh this potential limitation.

6. REFERENCES

[1] G. W. Wornell, "A Karhunen-Loève like expansion for $1/f$ processes via wavelets," *IEEE Trans. on Inform. Theory*, vol. 36, pp. 859–861, Mar. 1990.

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

[3] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Signal estimation using wavelet-Markov models," in *IEEE Int. Conf. on Acoust., Speech, Signal Proc. — ICASSP ' 97*, (Munich), pp. 3429–3432, April 1997.

[4] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.*, 1998. To appear. Technical report at http//www.dsp.rice.edu/.

[5] I. Daubechies, *Ten Lectures on Wavelets*. New York: SIAM, 1992.

[6] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive Bayesian wavelet shrinkage," *Journal of the American Statistical Association*, vol. 92, 1997.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[8] H. Lucke, "Which stochastic models allow Baum-Welch training?," *IEEE Trans. Signal Proc.*, vol. 11, pp. 2746–2756, Nov. 1996.

[9] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic," *IEEE/ACM Trans. on Networking*, vol. 2, pp. 1–15, Feb. 1994.