IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

COMBINING WAVELET-DOMAIN HIDDEN MARKOV TREES WITH HIDDEN MARKOV MODELS

Katrin Keller   Souheil Ben-Yacoub   Chafic Mokbel

IDIAP-RR 99-14

August 1999

SUBMITTED TO

Workshop on Automatic Speech Recognition and Understanding,
December 12-15, 1999, Keystone, Colorado

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP-RR 99-14

COMBINING WAVELET-DOMAIN HIDDEN MARKOV TREES WITH HIDDEN MARKOV MODELS

Katrin Keller   Souheil Ben-Yacoub   Chafic Mokbel

August 1999

SUBMITTED TO

Workshop on Automatic Speech Recognition and Understanding,
December 12-15, 1999, Keystone, Colorado

Abstract: In this paper, the concept of Wavelet-domain Hidden Markov Trees (WHMT) is introduced to Automatic Speech Recognition. WHMT are a convenient means to model the structure of wavelet feature vectors, as wavelet coefficients can be interpreted as nodes in a binary tree. By the introduction of hidden states in each node, non-Gaussian statistics inherent in wavelet features can be modeled. At the same time, correlations between neighboring coefficients in the time-frequency plane are accommodated. Phoneme probabilities obtained using the WHMT and wavelet features are then combined at the state level with those obtained by Gaussian distributions in conjunction with MFCCs, and fed into conventional Hidden Markov Models. Preliminary experiments show the potential advantages of this novel approach.

1 INTRODUCTION

In recent ASR research, it has been shown that supplementary information obtained on different timescales or resolution levels could improve recognition accuracy (e.g., [6][7][9]). As features with inherent multiresolution characteristics, wavelet coefficients offer an implicit way to exploit information on multiple timescales, since the timescale of the analysis varies with frequency. At the same time as they provide a higher temporal resolution for higher frequencies, a good frequency resolution for lower frequencies is obtained. However, the wavelet transformation as it is does not provide much information on a timescale large enough to capture the long-term dynamics of speech.

Wavelet coefficients have successfully been applied, e.g., in the field of image processing. Recent advances take advantage of inter-coefficient dependencies[1] by modeling wavelet feature vectors with a special kind of HMM: the Wavelet-domain Hidden Markov Model [4].

While the wavelet transformation is expected to decorrelate a signal, there still remain some major statistical dependencies. Particularly, adjacent coefficients in the time-frequency plane show a similar behavior, as can be seen from a speech sample in Figure 1. It becomes obvious that coefficients are correlated across time (horizontal axis) as well as scale (vertical axis). Those correlations are important properties of the wavelet transform, called clustering and persistency respectively.

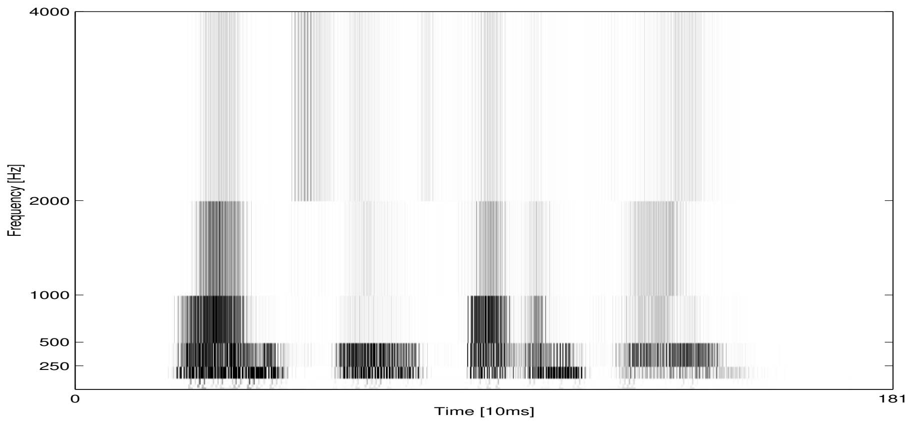

Figure 1: Wavelet data obtained from N95 database: The words pronounced are "one two seven three". Dark/light regions correspond to high/low energy coefficients.

Crouse et al. [4] introduce three types of Wavelet-domain Hidden Markov Models taking account of different correlations: Independent Mixtures treating each coefficient as statistically independent of all others (in this case, no correlations are considered), Markov Chains regarding only correlations across time, and Markov Trees. This article focuses on the latter model, thereby being mainly concerned with exploiting the dependencies across scale. On the basis of [2][4], Wavelet-domain Hidden Markov Trees (WHMT) are developed for application to ASR, and integrated into a system combining the conventional HMM approach with this new technology.

1. Exploiting correlations within and between feature vectors has recently shown some success in ASR, see, e.g., [1].

2 WAVELET TRANSFORMATION

The wavelet transformation is calculated using shifted versions of a low-pass scaling function Φ(t) and shifted and dilated versions of a bandpass wavelet function Ψ(t). These functions, if chosen reasonably, form an orthonormal basis, as, e.g., those proposed by Daubechies [5].

A way of interpreting wavelet coefficients is to regard their position in the time-frequency plane, where they are precisely localized. The width of the filterbanks increases with frequency, as does the number of the coefficients. At the same time, the length of the filters decreases in the same order. Thus, the highest resolution level L, spanning the upper half of the entire frequency band, consists of 2^{L-1} coefficients which each describes a well-defined timeslot within the analysis window. Those coefficients provide the most detail about the signal, while the lower resolution levels are the signal's subsequent approximations. The lowest resolution level does account for only one coefficient in a very low and narrow frequency band. From this, two major properties of the wavelet transform are apparent: locality (given the precise position of a coefficient in the time-frequency plane) and multiresolution (given the varying window size and different number of coefficients per resolution level). These characteristics make them attractive features in the area of speech recognition.

3 WAVELET-DOMAIN HIDDEN MARKOV TREES

3.1 General Concepts

Starting from the notion of a binary tree, we can model quite well wavelet coefficients using multiple Gaussian distributions within the nodes. A mixture of two Gaussians will accommodate the fact that wavelet data (for image processing) usually consist of a low number of high coefficients and a high number of low coefficients. However, this model does not consider dependencies between coefficients.

To introduce those dependencies, we assign to each node of the tree two states with one Gaussian distribution each. For a given data, we only observe its probability and not its state assignment, so the state distribution remains hidden. Each state is connected by an arc to its two possible parent states and to all of its child states (see Figure 2). These arcs model first order dependencies. We thereby obtain a Hidden Markov Tree for

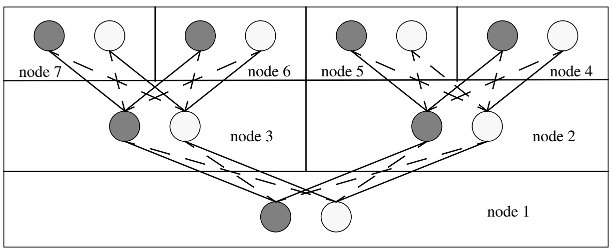

Figure 2: Part of a WHMT, showing the 3 coarsest resolution levels. The circles of different colors correspond to the different states (named H and L respectively). Full arcs show connections between similar states, dashed ones transitions between H and L or vice versa.

wavelet data with the following model parameters λ: the probability mass distribution of the two states in the root node (equivalent to the initial probabilities in a classical HMM) denoted π_i, the transition probabilities between states of adjacent resolution levels:

$$a_{s_n, s_{2n+i}},\ i \in \{0, 1\};\ s_n, s_{2n+i} \in \{H, L\}$$

as well as means and variances of the Gaussian of each state:

$$N_{s_n} = (\mu_{s_n}, \sigma_{s_n}),\ s_n \in \{H, L\}$$

Given an observed wavelet data W, the question arises of how to determine the likelihood of a known WHMT model. This can not be done directly since we do not observe the hidden states. Thus, to compute the likelihood we should develop the sum over all the possible paths S (expectation):

$$p(W|\lambda) = \sum_S p(W, S|\lambda) = \sum_S p(S|\lambda) p(W|S, \lambda) \tag{1}$$

$$= \sum_S \pi_{s_0} \cdot N_{s_0}(w(0)) \prod_{l=1}^{L-1} \prod_{n=2^{l-1}}^{2^l-1} \prod_{i=0}^{1} a_{s_n, s_{2n+i}} N_{s_{2n+i}}(w(2n+i))$$

An explicit application of Eq. 1 is very costly. Thus, decoding is done using the forward-backward algorithm, similar to the one used in classical HMMs. However, the forward algorithm in this case is not straightforward. For a node in the WHMT and a state in that node, the forward variable represents the probability of observing the whole data from the root except the part of the data relative to the subtree started at that node. We define the subtree T_n relative to a node n as the set of possible nodes k such that $k = 2^l n+i,\ i = 0,..,2^l$-1. The forward variable can be written for node n and state k:

$$\alpha_n(k) = p(S_n = k, W(m, m \notin T_n) | \lambda) \tag{2}$$

The second part of the algorithm is much simpler since the backward variable of a given couple node/state designates the probability of emitting the data of the subtree attached to that node given that state:

$$\beta_n(k) = p(W(m, m \in T_n) | S_n = k, \lambda) \tag{3}$$

Starting from the leaves of the tree, the backward variables can be determined recursively. Even if they are sufficient to compute the likelihood, we might need the forward variables (e.g., for training) which can be computed given the backward variables.

The WHMT models are trained using an adapted version of the Expectation Maximization (EM) algorithm since a training population does not form a sufficient statistic to estimate the WHMT parameters. The EM is an iterative algorithm that ensures a likelihood increase with each iteration, thus converging to a local optimum. Every iteration consists of two steps: “Expectation” (E) and “Maximization” (M). During the E step, the forward and backward variables are computed for every single example of the training population given the preceding set of parameters. Then, in the M step the WHMT parameters can be reestimated. The computation of the likelihood involves the multiplication of 2^L likelihoods and 2^L probabilities. This might result in some numerical problems which can be avoided by the use of scaling factors. However, as these factors depend on the WHMT, the likelihoods become incomparable. We propose to compensate this scaling in the logarithmic domain after the decoding. Please refer to [4] for more details on this problem as well as on the reestimation functions of the forward and backward variables.

As one tree as described above might not be able to account for all potential variability in the pronunciation of a phoneme, some mechanism equivalent to having multiple Gaussians per state in a conventional HMM should be introduced. There are two obvious choices: we can either augment the number of states in each node or employ several WHMTs per phoneme in parallel. In our recent work, the multiple tree approach was used.

3.2 Combination with Conventional Systems

The WHMT models work on top of the HMMs as usually applied in ASR. In this framework HMM system, each phoneme is modeled by 3 states, for which the phoneme likelihood calculation is performed by the same WHMT[1]. Apart from the average phoneme durations (reflected by the HMM transition probabilities), only information present in the wavelet representation is used.

Unfortunately, with this rather simplistic approach we are not able to incorporate the long-term dynamics of the speech signal. As the well-established HMMs using MFCCs as feature vectors are at hand, our first choice was to use them as a parallel system to supply additional information. We combined likelihoods calculated by our WHMT models on wavelet data with those obtained from single-Gaussian HMMs on MFCCs at the frame level. They were then fed to the framework HMM system (see Figure 3). The EM algorithm might also be used to estimate the parameters of the combined HMM-WHMT. Starting from an initial set of parameters, the measured wavelet coefficients are first assigned statistically to the hidden states of the HMM and the hidden states of the WHMT during the E step. Based on this assignment, new values for the parameters of the combined HMM-WHMT model can be computed during the M step. This is continued iteratively until convergence.

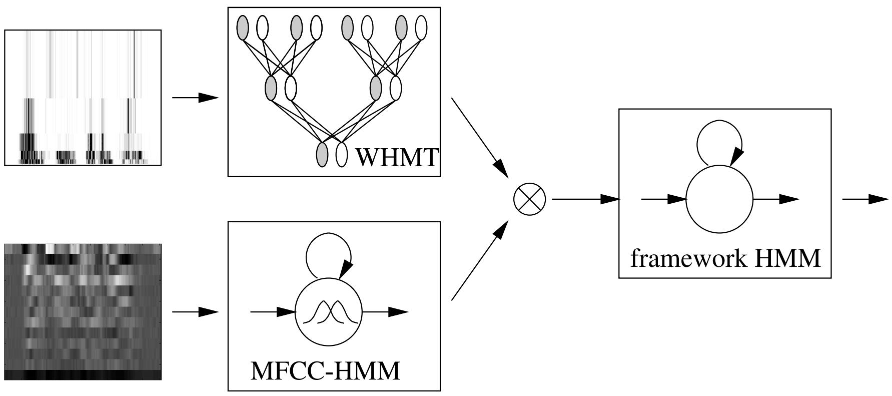

Figure 3: Combination of WHMTs and conventional HMMs (taking as features wavelet coefficients and MFCCs respectively), and integration into framework HMM system.

4 PRELIMINARY EXPERIMENTS

To run preliminary experiments, we developed a software implementing the WHMT models (on the basis of [2]) as well as a tool to combine probabilities of different streams. A modified version of the HTK toolkit [10] was used for the decoding part of the experiments. As database, Numbers95 [3] with its standard train and development test sets was used throughout. Feature extraction was performed using a front-end developed at IDIAP. Wavelet coefficients were calculated on 32ms windows of speech, shifted by 10ms, yielding large feature vectors of 256 components. Monophone WHMT models were trained separately for each phoneme on the base of the hand-segmented Numbers95 training data. For the second testing phase, 13 MFCC coefficients (including energy, calculated on the same signal windows) were used to train single Gaussians.

1. In some way, this is similar to the probability calculation by ANNs as employed by hybrid systems.

First test determined which kind of wavelet transformation was to be used. Table 1 shows the error rates on frame and segment level of the Daubechies wavelet transformation of different degrees.

Error Rate	Train Set (frame)	Train Set (segment)	Dev. Test (frame)	Dev. Test (segment)
20 coeff.	73.6	68.4	73.4	68.5
12 coeff.	72.7	68.4	73.0	69.1
4 coeff.	71.1	67.8	71.0	67.8

Table 1: Frame and segment level error rates for Daubechies wavelet transform of different degrees.

Clearly, the Daubechies transform with 4 coefficients performs best. Thus, and as the number of training iterations was comparable for all cases, the Daubechies-4 transformation was chosen for all following experiments. Looking at the confusion matrices of the above tests, a high percentage of errors can be stated for diphtons with resembling parts. For instance, “ey” was very often mistaken as “ay” or “iy”. This suggests that the employed models cannot handle the variations within one diphton over time. By the introduction of several parallel WHMTs per phoneme we hoped to circumvent this problem. Table 2 compares the error rates obtained on different numbers of parallel trees.

Error Rate	Train Set (frame)	Train Set (segment)	Dev. Test (frame)	Dev. Test (segment)
1 tree	71.1	67.8	71.0	67.8
2 trees	74.3	65.3	74.5	66.1
3 trees	70.1	59.3	70.9	60.8
4 trees	69.8	56.9	70.5	58.8
6 trees	70.0	55.1	71.2	58.6

Table 2: Frame and segment level error rates for Daubechies-4 wavelet transform employing several WHMTs in parallel.

Looking again at the confusion matrices, we can state that systems with different numbers of trees misrecognized different phonemes. In perspective, a sensible combination of these systems could be able to increase recognition performance. So we achieved some improvement in combining the 2-tree and the 6-tree systems, simply choosing a model on a per-phoneme bases as a function of the number of training examples. However, as defining suitable selection criteria is not a trivial task, we choose the 4-tree system for further experiments.

As our system is still in its infancy, we could not expect breathtaking recognition results. However, when combining the WHMT system with a basic conventional HMM system, we gained some improvement compared to either system working separately. After only the initialization step (thus, no EM training was performed yet within the framework HMM), we achieved a word accuracy of 49.70, which compares to 32.25 on the WHMT and 48.05 on the HMM system separately.

5 FUTURE RESEARCH DIRECTIONS

We would again like to emphasize that WHMTs have been developed only very recently, and even more so their application to speech recognition. Therefore, currently they do not compare by far to state-of-the-art speech recognition technology. However, they leave a lot of possibilities for further improvement. Some promising directions for future research are outlined below.

At the feature extraction level, data from rather low frequencies are used, which usually are, in the case of telephone speech, disturbed by line effects [8]. In our case, the 4 lower levels of the wavelet data contain only information from frequencies below 250Hz, which might influence the recognition results in a negative way. This also causes the problem that there is no trustworthy information on analysis windows longer than 2ms, a problem which is aggravated by the fact that derivatives are not yet integrated. Furthermore, there is no appropriate energy measure and no normalization. With an ameliorated signal processing, wavelet data more adapted to the characteristics of (telephone) speech can be generated.

At the wavelet feature modeling level, some problems described above are reflected. Obviously, the system relies heavily on the 4 lowest resolution levels. Furthermore, a mechanism has to be introduced to incorporate derivatives into our system. This could be done by derivative WHMTs which are connected to the others to reflect dependencies between wavelet data and their derivatives. Some improvement might also be gained be extending the model to allow different numbers of parallel WHMTs or Gaussian mixture distributions. Moreover, the WHMT model might be rendered more flexible by introducing loops into the nodes/states, which, among other potential advantages, would allow to move the phoneme duration modeling from conventional HMMs directly to the WHMT models.

As for conventional HMMs to process MFCCs, so far a rather basic system has been applied which could be replaced by a state-of-the art one. Certainly, the probability combination mechanism could be improved with an appropriate weighting scheme. However, the motivation of introducing such a parallel system has been the present drawbacks of our WHMT system which should be investigated first. Nevertheless, having a second stream of information remains an interesting future research area also in the case of WHMTs. Furthermore, the framework HMM system could be improved by, e.g., introducing triphones and/or more emitting states per phoneme.

More generally, the idea of modeling features by means of special Markov models working on top of the conventional HMM mechanism can be extended to features other than wavelets. For example, filterbank coefficients or even MFCCs could be modeled by double Markov chains with cross-connected hidden states. Also in this case, considering the residual correlations between adjacent coefficients of a feature vector after signal processing seems a promising research direction.

6 CONCLUSIONS

In this article, Wavelet-domain Hidden Markov Trees are introduced to ASR. They exploit inherent properties of wavelet data, in particular the correlations between coefficients. Preliminary experiments in conjunction with conventional Hidden Markov Models show some potential of this novel approach. As shown in Chapter 5, there are still a lot of possibilities for improvement. Therefore, we see much more potential in the WHMT approach.

7 ACKNOWLEDGMENTS

The authors would like to thank H. Choi of Rice University for software forming the starting point of the WHMT system development. This work was partly supported by grant FN 2100-50742.97/1 from the Swiss National Science Foundation.

REFERENCES

[1] Jeff A. Bilmes. Buried Markov Models for Speech Recognition. *Proc. ICASSP*, II:713-716, March 1999.

[2] H. Choi and R. G. Baraniuk. Image Segmentation using Wavelet-domain Classification, *Proc. SPIE Technical Conference on Mathematical Modeling, Bayesian Estimation, and Inverse Problems*, pp. 306-320, Denver, July 1999.

[3] R. A. Cole, M. Noel, T. Lander, and T. Durham. New Telephone Speech Corpora at CSLU. *Proc. Eurospeech*, 1:821-824, September 1995.

[4] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Trans. on Signal Processing*, vol. 46, no. 4, pp. 886-902, April 1998.

[5] Ingrid Daubechies. *Ten Lectures on Wavelets.* CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[6] H. Hermansky and S. Sharma. Temporal Patterns (TRAPS) in ASR of Noisy Speech. *Proc. ICASSP*, I:289-292, March 1999.

[7] P. McCourt, S. Vaseghi, and N. Harte. Multi-Resolution Cepstral Features for Phoneme Recognition across Speech Sub-Bands. *Proc. ICASSP*, I:557-560, May 1998.

[8] C. Mokbel, D. Jouvet, and J. Monné. Deconvolution of Telephone Line Effects for Speech Recognition. *Speech Communication*, vol. 19, no. 3, pp. 185-196, September 1996.

[9] S. Wu, B. Kingsbury, N. Morgan, and S. Greenberg. Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition. *Proc. ICASSP*, II:721-724, May 1998.

[10] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book.* Cambridge University, 1995.