An EM Algorithm to Learn Sequences in the Wavelet Domain

Diego H. Milone and Leandro E. Di Persia

Signals and Computational Intelligence Laboratory - CONICET
Department of Informatics, Faculty of Engineering and Water Sciences
National University of Litoral, Ciudad Universitaria, Santa Fe, Argentina
{dmilone,ldipersia}@fich.unl.edu.ar
http://fich.unl.edu.ar/sinc

Abstract. The wavelet transform has been used for feature extraction in many applications of pattern recognition. However, in general the learning algorithms are not designed taking into account the properties of the features obtained with discrete wavelet transform. In this work we propose a Markovian model to classify sequences of frames in the wavelet domain. The architecture is a composite of an external hidden Markov model in which the observation probabilities are provided by a set of hidden Markov trees. Training algorithms are developed for the composite model using the expectation-maximization framework. We also evaluate a novel delay-invariant representation to improve wavelet feature extraction for classification tasks. The proposed methods can be easily extended to model sequences of images. Here we present phoneme recognition experiments with TIMIT speech corpus. The robustness of the proposed architecture and learning method was tested by reducing the amount of training data to a few patterns. Recognition rates were better than those of hidden Markov models with observation densities based in Gaussian mixtures.

Key words: EM algorithm, Hidden Markov Models, Hidden Markov Trees, Speech Recognition, Wavelets.

1 Introduction

Hidden Markov trees (HMT) have been recently introduced [1] to model the statistical dependencies at different scales and the non-Gaussian distributions of the wavelet coefficients [2]. In the last years, the HMT model was improved in several ways, for example, using more states within each HMT node and developing more efficient algorithms for initialization and training [3, 4].

Discrete and continuous hidden Markov models (HMM) have been used in applications of machine learning and pattern recognition, such as computer vision, bioinformatics, speech recognition, medical diagnosis and many others [5–8]. A well-known model for continuous observation densities is the Gaussian mixture model (GMM) [9]. The HMM-GMM architecture is a widely used model, for

example, in speech recognition [10]. Nevertheless, more accurate models have been proposed for the observation densities [11]. In both, discrete and continuous models, the most important advantage of the HMM lies in that they can deal with sequences of variable length. However, if the whole sequence is analyzed by the discrete wavelet transform (DWT), like in the case of HMT, a representation whose structure is dependent on the sequence length is obtained. Therefore, the learning architecture should be trained and used only for this sequence length. On the other hand, in HMM modeling, stationarity is generally assumed withing each observation in the sequence. This stationarity assumption can be removed when observed features are extracted by the DWT, but a suitable statistical model for learning this features in the wavelet domain would be needed.

Fine et al. [12] proposed a recursive hierarchical generalization of discrete HMM. They apply the model to learn the multiresolution structure of natural English text and cursive handwriting. Some years later, Murphy and Paskin [13] derived a simpler inference algorithm by formulating the hierarchical HMM as a special kind of dynamic Bayesian network. A wide review about multiresolution Markov models was provided in [14], with special emphasis on applications to signal and image processing. Dasgupta et al. [15] proposed a dual-Markov architecture, trained by means of an iterative process where the most probable sequence of states is identified, and then each internal model is adapted with the selected observations. A similar approach applied to image segmentation is proposed in [16]. However, in these cases the model consists of two separated and independent entities, that are forced to work in a coupled way. By the contrary, Bengio et al. derived a training algorithm for the full model [17], composed of an external HMM in which for each state an internal HMM provides the observation probability distribution [18]. In the following, we derive an EM algorithm for a composite HMM-HMT architecture that observes sequences of DWTs in $\mathbb{R}^N$. This algorithm can be easy generalized to sequences in $\mathbb{R}^{N\times N}$ with 2-D HMTs like the used in [19] or [20].

In the next section we introduce the notation for HMM and HMT. Using this notation we review the training algorithms by defining the joint likelihood and then deriving the reestimation formulas. In Section 3, the experimental results for speech recognition using real data are presented and discussed. Two different alternatives for the feature extraction with DWT are tested. The experiments are carried out by training the proposed model and the HMM-GMM with a variable amount of data to test their robustness with a few training patterns. In the last section we present the main conclusions and some ideas for future works.

2 The HMM-HMT Model

The architecture proposed in this work is a composition of two Markov models: the long term dependencies are modeled with an external HMM and each pattern in the local context is modeled with an HMT.

2.1 Basic Definitions

To model a sequence $\mathbf{W} = \mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^T$, with $\mathbf{w}^t \in \mathbb{R}^N$, a continuous HMM is defined with the structure $\vartheta = \langle \mathcal{Q}, \mathbf{A}, \boldsymbol{\pi}, \mathcal{B} \rangle$, where:

i) $\mathcal{Q} = \{Q \in [1 \ldots N_Q]\}$ is the set of states.
ii) $\mathbf{A} = \left[a_{ij} = \Pr\left(Q^t = j \left| Q^{t-1} = i\right.\right)\right], \forall i, j \in \mathcal{Q}$, is the matrix of transition probabilities, where $Q^t \in \mathcal{Q}$ is the model state at time $t \in [1 \ldots T]$, $a_{ij} \geq 0\ \forall i, j$ and $\sum_j a_{ij} \stackrel{\circ}{=} 1\ \forall i$.
iii) $\boldsymbol{\pi} = [\pi_j = \Pr(Q^1 = j)]$ is the initial state probability vector. In the case of left-to-right HMM this vector is $\boldsymbol{\pi} = \boldsymbol{\delta}_1$.
iv) $\mathcal{B} = \{b_k(\mathbf{w}^t) = \Pr(\mathbf{W}^t = \mathbf{w}^t | Q^t = k)\}, \forall k \in \mathcal{Q}$, is the set of observation (or emission) probability distributions.

Let be $\mathbf{w} = [w_1, w_2, \ldots, w_N]$ resulting of a DWT analysis with J scales and without including w_0, the approximation coefficient at the coarsest scale (that is, $N = 2^J - 1$). The HMT can be defined with the structure $\theta = \langle \mathcal{U}, \mathcal{R}, \boldsymbol{\pi}, \boldsymbol{\epsilon}, \mathcal{F} \rangle$, where:

i) $\mathcal{U} = \{u \in [1 \ldots N]\}$ is the set of nodes in the tree.
ii) $\mathcal{R} = \{R \in [1 \ldots NM]\}$ is the set of states in all the nodes of the tree, denoting with $\mathcal{R}_u = \{R_u \in [1 \ldots M]\}$ the set of states in the node u.
iii) $\boldsymbol{\epsilon} = \left[\epsilon_{u,mn} = \Pr(R_u = m | R_{\rho(u)} = n)\right], \forall m \in \mathcal{R}_u, \forall n \in \mathcal{R}_{\rho(u)}$, is the array whose elements hold the conditional probability of node u being in state m given that the state in its parent node $\rho(u)$ is n, where $\sum_m \epsilon_{u,mn} \stackrel{\circ}{=} 1$.
iv) $\boldsymbol{\pi} = [\pi_p = \Pr(R_1 = p)], \forall p \in \mathcal{R}_1$ are the probabilities for the root node being on state p.
v) $\mathcal{F} = \{f_{u,m}(w_u) = \Pr(W_u = w_u | R_u = m)\}$ are the observation probability distributions. This is, $f_{u,m}(w_u)$ is the probability of observing the wavelet coefficient w_u with the state m (in the node u).

In the following, we will simplify the notation for random variables. For example, we write $\Pr(w_u | r_u)$ instead of $\Pr(W_u = w_u | R_u = r_u)$.

2.2 Joint Likelihood

Let be Θ an HMM like the one defined above but using a set of HMTs to model the observation densities within each HMM state:

$$b_{q^t}(\mathbf{w}^t) = \sum_{\forall \mathbf{r}} \prod_{\forall u} \epsilon^{q^t}_{u,r_u r_{\rho(u)}} f^{q^t}_{u,r_u}(w^t_u), \tag{1}$$

with $\mathbf{r} = [r_1, r_2, \ldots, r_N]$ a combination of hidden states in the HMT nodes. To extend the notation in the composite model, we have added a superscript in the HMT variables to make reference to the state in the external HMM. For example, $\epsilon^k_{u,mn}$ will be the conditional probability that, in the state k of the external HMM, the node u is in state m given that the state of its parent node $\rho(u)$ is n.

Thus, the complete joint likelihood for the HMM-HMT can be obtained as

$$\begin{aligned}\mathcal{L}_\Theta(\mathbf{W}) &= \sum_{\forall \mathbf{q}} \sum_{\forall \mathbf{R}} \prod_t a_{q^{t-1}q^t} \prod_{\forall u} \epsilon^{q^t}_{u,r^t_u r^t_{\rho(u)}} f^{q^t}_{u,r^t_u}(w^t_u) \\ &\triangleq \sum_{\forall \mathbf{q}} \sum_{\forall \mathbf{R}} \mathcal{L}_\Theta(\mathbf{W},\mathbf{q},\mathbf{R}),\end{aligned} \tag{2}$$

where we simplify $a_{01} = \pi_1 = 1$, $\forall \mathbf{q}$ is over all possible state sequences $\mathbf{q} = q^1, q^2, \ldots, q^T$ and $\forall \mathbf{R}$ are all the possible sequences of all the possible combinations of hidden states $\mathbf{r}^1, \mathbf{r}^2, \ldots, \mathbf{r}^T$ in the nodes of each tree.

2.3 Training Formulas

In this section we will obtain the maximum likelihood estimation of the model parameters. For the optimization, the auxiliary function can be defined as

$$\mathcal{D}(\Theta, \bar{\Theta}) \triangleq \sum_{\forall \mathbf{q}} \sum_{\forall \mathbf{R}} \mathcal{L}_\Theta(\mathbf{W},\mathbf{q},\mathbf{R}) \log\left(\mathcal{L}_{\bar{\Theta}}(\mathbf{W},\mathbf{q},\mathbf{R})\right) \tag{3}$$

and using (2)

$$\begin{aligned}\mathcal{D}(\Theta, \bar{\Theta}) = \sum_{\forall \mathbf{q}} \sum_{\forall \mathbf{R}} \mathcal{L}_\Theta(\mathbf{W},\mathbf{q},\mathbf{R}) \cdot \Bigg\{ &\sum_t \log(a_{q^{t-1}q^t}) + \\ &+ \sum_t \sum_{\forall u} \left[\log\left(\epsilon^{q^t}_{u,r^t_u r^t_{\rho(u)}}\right) + \log\left(f^{q^t}_{u,r^t_u}(w^t_u)\right) \right] \Bigg\}.\end{aligned} \tag{4}$$

For the estimation of the transition probabilities in the HMM, a_{ij}, no changes from the standard formulas will be needed. However, on each internal HMT we hope that the estimation of the model parameters will be affected by the probability of being in the HMM state k at time t.

Let be $q^t = k$, $r^t_u = m$ and $r^t_{\rho(u)} = n$. To obtain the learning rule for $\epsilon^k_{u,mn}$ the restriction $\sum_m \epsilon^k_{u,mn} \stackrel{\circ}{=} 1$ should be satisfied. If we use

$$\hat{\mathcal{D}}(\Theta, \bar{\Theta}) \triangleq \mathcal{D}(\Theta, \bar{\Theta}) + \sum_n \lambda_n \left(\sum_m \epsilon^k_{u,mn} - 1 \right), \tag{5}$$

the learning rule results

$$\epsilon^k_{u,mn} = \frac{\sum_t \gamma^t(k) \xi^{tk}_u(m,n)}{\sum_t \gamma^t(k) \gamma^{tk}_{\rho(u)}(n)}, \tag{6}$$

where $\gamma^t(k)$ is computed as usual for HMM and $\gamma^{tk}_{\rho(u)}(n)$ and $\xi^{tk}_u(m,n)$ can be estimated with the upward-downward algorithm [4].

For the observation distributions we use $f^{q^t}_{u,r^t_u}(w^t_u) = \mathcal{N}\left(w^t_u, \mu^{q^t}_{u,r^t_u}, \sigma^{q^t}_{u,r^t_u}\right)$. From (4) we have

$$\mathcal{D}(\Theta,\bar{\Theta}) = \sum_{\forall \mathbf{q}}\sum_{\forall \mathbf{R}} \mathcal{L}_\Theta(\mathbf{W},\mathbf{q},\mathbf{R}) \cdot \left[\sum_t \log(a_{q^{t-1}q^t}) + \sum_t\sum_{\forall u} \log\left(\epsilon^{q^t}_{u,r^t_u r^t_{\rho(u)}}\right) + \right.$$
$$\left. + \sum_t\sum_{\forall u}\left(-\frac{\log(2\pi)}{2} - \log\left(\sigma^{q^t}_{u,r^t_u}\right) - \frac{\left(w^t_u - \mu^{q^t}_{u,r^t_u}\right)^2}{2\left(\sigma^{q^t}_{u,r^t_u}\right)^2}\right)\right]. \qquad (7)$$

Thus, the training formulas result:

$$\mu^k_{u,m} = \frac{\sum_t \gamma^t(k)\gamma^{tk}_u(m) w^t_u}{\sum_t \gamma^t(k)\gamma^{tk}_u(m)} \quad \text{and} \quad (\sigma^k_{u,m})^2 = \frac{\sum_t \gamma^t(k)\gamma^{tk}_u(m)\left(w^t_u - \mu^k_{u,m}\right)^2}{\sum_t \gamma^t(k)\gamma^{tk}_u(m)}.$$

2.4 Multiple Observation Sequences

In practical situations we have a training set with a large number of observed data $\mathcal{W} = \{\mathbf{W^1}, \mathbf{W^2}, \ldots, \mathbf{W}^P\}$, where each observation consists of a sequence of evidences $\mathbf{W}^p = \mathbf{w}^{p,1}, \mathbf{w}^{p,2}, \ldots, \mathbf{w}^{p,T_p}$, with $\mathbf{w}^{p,t} \in \mathbb{R}^N$. In this case we define the auxiliary function

$$\mathcal{D}(\Theta,\bar{\Theta}) \triangleq \sum_{p=1}^{P} \frac{1}{\Pr\left(\mathbf{W}^p|\theta\right)} \sum_{\forall \mathbf{q}}\sum_{\forall \mathbf{R}} \mathcal{L}_\Theta(\mathbf{W}^p,\mathbf{q},\mathbf{R}) \log\left(\mathcal{L}_{\bar{\Theta}}(\mathbf{W}^p,\mathbf{q},\mathbf{R})\right) \qquad (8)$$

and replacing with (2)

$$\mathcal{D}(\Theta,\bar{\Theta}) = \sum_{p=1}^{P} \frac{1}{\Pr\left(\mathbf{W}^p|\theta\right)} \sum_{\forall \mathbf{q}}\sum_{\forall \mathbf{R}} \mathcal{L}_\Theta(\mathbf{W}^p,\mathbf{q},\mathbf{R}) \cdot \left\{\sum_{t=1}^{T_p} \log(a_{q^{t-1}q^t}) + \right.$$
$$\left. + \sum_{t=1}^{T_p}\sum_{\forall u}\left[\log\left(\epsilon^{q^t}_{u,r^t_u r^t_{\rho(u)}}\right) + \log\left(f^{q^t}_{u,r^t_u}(w^{p,t}_u)\right)\right]\right\}. \qquad (9)$$

The derivation of the training formulas is similar to the ones presented for a single sequence. We simply summarize here the main results:

$$a_{ij} = \frac{\sum_{p=1}^{P}\sum_{t=1}^{T_p} \xi^{p,t}(i,j)}{\sum_{p=1}^{P}\sum_{t=1}^{T_p} \gamma^{p,t}(i)}, \; (\sigma^k_{u,m})^2 = \frac{\sum_{p=1}^{P}\sum_{t=1}^{T_p} \gamma^{p,t}(k)\gamma^{p,tk}_u(m)\left(w^{p,t}_u - \mu^k_{u,m}\right)^2}{\sum_{p=1}^{P}\sum_{t=1}^{T_p} \gamma^{p,t}(k)\gamma^{p,tk}_u(m)},$$

$$\epsilon^k_{u,mn} = \frac{\sum_{p=1}^{P}\sum_{t=1}^{T_p}\gamma^{p,t}(k)\xi^{p,tk}_u(m,n)}{\sum_{p=1}^{P}\sum_{t=1}^{T_p}\gamma^{p,t}(k)\gamma^{p,tk}_{\rho(u)}(n)}, \; \mu^k_{u,m} = \frac{\sum_{p=1}^{P}\sum_{t=1}^{T_p}\gamma^{p,t}(k)\gamma^{p,tk}_u(m)w^{p,t}_u}{\sum_{p=1}^{P}\sum_{t=1}^{T_p}\gamma^{p,t}(k)\gamma^{p,tk}_u(m)}$$

3 Experimental Results and Discussion

In this section we test the proposed model in the context of automatic speech recognition with the TIMIT corpus [21]. In the following the data set for the experiments is briefly described. Subsection 3.2 details the first experiments with the standard DWT and comparing the performance of the proposed model and the others which are often used for this classification task. Next, we present and discuss the results obtained by introducing an improvement in the DWT feature extraction for speech data. In the Subsection 3.4 we present a set of tests aimed to evaluate the robustness of the models to the reduction of the amount of training data.

3.1 Data sets and implementation details

TIMIT is a well known corpus that has been used extensively for research in automatic speech recognition. From this corpus five phonemes that are difficult to classify were selected. The voiced stops /*b*/ and /*d*/ have a very similar articulation (bilabial/alveolar) and different phonetic variants according to the context (allophones). Vowels /*eh*/ and /*ih*/ were selected because their formants are very close. Thus, these phonemes are very confusable. To complete the selected phonemes, the affricate phoneme /*jh*/ was added as representative of the voiceless group [22]. Table 1 shows the number of train and test samples for each phoneme in all the dialectical regions of the TIMIT corpus.

Phonemes	/b/	/d/	/eh/	/ih/	/jh/
Train patterns	2181	3548	3853	5051	1209
Test patterns	886	1245	1440	1709	372

Table 1. Selected phonemes from TIMIT speech corpus.

Regarding practical issues, the training formulas were implemented in logarithmic scale to make a more efficient computation of products and to avoid underflow errors in the probability accumulators [4]. In addition, underflow errors are reduced because in the HMM-HMT architecture each DWT is in a lower dimension than the dimension resulting from an unique HMT for the whole sequence (like in [1] and [4]). All learning algorithms and transforms used in the experiments were implemented in C++ from scratch.

3.2 Standard DWT features

Frame by frame, each local feature is extracted using a Hamming window of width N_w, shifted in steps of N_s samples [10]. The first window begins N_o samples out (with zero padding) to avoid the information loss at the beginning of the sequence. The same procedure is used to avoid information loss at the end of the sequence. Then a DWT is applied to each windowed frame. The DWT was implemented by the fast pyramidal algorithm [23], using periodic convolutions and the Daubechies-8 wavelet [24]. Preliminary tests were carried out with other wavelets of the Daubechies and Splines families but not important differences in results were found.

In this first study, different recognition architectures are compared, but setting they to have the total number of trainable parameters in the same order of magnitude. A separate model is trained for each phoneme and the recognition is made by the conventional maximum-likelihood classifier. Table 2 shows the recognition rates (RR) for: GMM with 4 Gaussians in the mixture (2052 trainable parameters), HMT with 2 states per node and one Gaussian per state (2304 trainable parameters), HMM-GMM with 3 states and 4 Gaussians in each mixture (6165 trainable parameters), HMM-HMT with 3 HMM states, 2 states per HMT node and one Gaussian per node state (6921 trainable parameters)[1]

For HMM-GMM and HMM-HMT the external HMM have connections $i \to i$, $i \to (i+1)$ and $i \to (i+2)$. The last link allows to model the shortest sequences, with less frames than states in the model. In both, GMM and HMM-GMM, the Gaussians in the mixture are modeled with diagonal covariance matrices.

The maximum number of reestimations used for all experiments was 10, but also, as finalization criteria, the training process was stopped if the average (log) probability of the model given the training sequences was improved less than 1%. In most of the cases, the training converges after 4 to 6 iterations, but HMM-GMM models experienced several convergence problems with the DWT data. When a convergence problem was observed, the model corresponding to the last estimation with an improvement in the average probability of the model given the sequences was used for testing.

Results in Table 2 were obtained with the dialectical region 1 of the TIMIT corpus (1145 phonemes for train and 338 for test). Recognition rates (RR) for HMT are higher than those achieved by GMM, mainly because HMT provides a better model for the structure in the wavelet coefficients. However, the capability of the HMM-GMM to model the dynamics of the sequences of frames allows to improve RR over HMT. But moreover, the combination of the two advantages in the HMM-HMT surpass all previous results.

The computational cost may be one of the major handicaps of the proposed approach, mainly because of the double Baum-Welch process required in the training. To provide an idea of the computational cost, results reported in Table 2, with $N_w = 256$ and $N_s = 128$, demand 30.20 s of training for the HMM-GMM whereas the same training set demands 240.89 s in the HMM-HMT[2].

[1] All these counts are for $N_w = 256$.

[2] Using a Intel Core 2 Duo E6600 processor.

Learning Architecture	Frame size N_w 128	256	Average RR%
GMM	28.99	29.88	29.44
HMT	31.36	36.39	33.88
HMM-GMM	35.21	37.87	36.54
HMM-HMT	47.34	39.64	43.49

Table 2. Recognition rates (RR%) for TIMIT phonemes (dialectical region 1) using models with a similar number of trainable parameters.

3.3 Delay-invariant DWT features

The next experiments were aimed to the comparison of the two main models related with this work, that is, HMM using observation probabilities provided by GMMs or HMTs. In this context, the best relative scenario for HMM-GMM is using $N_w = 256$ and $N_s = 128$ (see Table 2).

Furthermore, in this section we will study a problem that arises in the feature extraction with the DWT, applied in the context of a frame by frame analysis. When a quasi-periodic waveform is analyzed by DWT, it can be seen that the major peak is replicated within each scale. In a frame by frame analysis, the positions of these peaks are related to the difference between the starting time of the frame and the location of the maximum. Thus, this is an artifact not related with the identity of the phoneme. Undoubtedly, these artifacts make training data too confusable for any recognition architecture without translation-invariance. A wavelet representation for quasi-periodic signals was proposed in [25]. Other authors integrate all the coefficients within each scale, using only the subband energy information of the DWT. Then, they apply a principal component analysis [26] or the cosine transform [27] to decorrelate the frame. However, a simpler idea can be applied to avoid this information loss: the spectrum module by scale (SMS) can be used rather than the wavelet coefficients themselves. This solution preserves the DWT information within scales without a major complexity in the implementation and remove the artifact generated by the quasi-periodic behavior of the waveform. In the following experiments we will refer to this method for feature extraction as SMS-DWT.

In Table 3 we present a fine tuning for the HMM-GMM model, with DWT and SMS-DWT feature extraction. Note that comparable architectures for HMM-HMT are HMM-GMM with between 2 to 8 Gaussians in the mixtures, because they have a similar number of trainable parameters. In spite of all the tested HMM-GMM alternatives, the HMM-HMT is still providing the best recognition rates. The improvement obtained by the SMS method is very important for both models. The proposed model is still providing the best results, mainly because the structural relations between the wavelet scales (modeled by the HMTs) are preserved after the SMS post-processing.

Learning Architecture	Gaussians per GMM	Total of parameters	Recognition rates [%] DWT	SMS-DWT
HMM-GMM	2	3087	29.60	63.07
	4	6165	28.33	63.40
	8	12321	33.73	62.20
	16	24633	28.93	62.00
	32	49257	27.00	63.00
	64	98505	33.47	59.13
HMM-HMT	512	6921	37.93	66.27

Table 3. Recognition results for TIMIT phonemes applying the DWT directly to each frame and with the SMS post-processing. All dialectical region of TIMIT corpus were used in these experiments. Note that HMM-HMT Gaussians are in $\mathbb{R}^1$ whereas HMM-GMM Gaussians are in $\mathbb{R}^{256}$.

3.4 Robustness to the amount of training data

Fig. 1 shows the results when the training data is reduced to a few patterns. Each recognition rate on this figure is the average of 10 independent trials, with training patterns selected at random. From the testing set of TIMIT corpus, 300 patterns for each class were also selected at random for each trial.

This figure shows that, when there is a sufficient amount of training data, recognition rates grow up to the reported in Table 3. However, when the amount of training data is reduced, HMM-GMM performance is significantly more affected than the performance of HMM-HMT. In the case of SMS-DWT, the RR of HMM-GMM fall from 61.33% to 37.34% (about the same RR that the average for HMM-HMT with the standard DWT features). In contrast, RR for the HMM-HMT only falls from 64.67% to 52.51%.

On the other hand, it can be observed that with standard DWT features the performance of the HMM-GMM falls more slowly, from 30.11% to 21.18%. Nevertheless, HMM-HMT retain the RR up to approximately 20 training patterns, where begins a weak trend from 35.08% to 30.73%. This important robustness of the proposed model can be attributed to the better capability for modeling the relevant information of features in the wavelet domain.

4 Conclusions

The proposed algorithms for HMM-HMT allows learning from variable-length sequences in the wavelet domain. The training algorithms were derived using the EM framework, resulting in a set of learning rules with a simple structure.

The recognition rates obtained for classification were very competitive, even in comparison with the state-of-the-art technologies in this application domain. The proposed post-processing for the DWT feature extraction resulted in very important improvements of the recognition rates. In this empirical tests, the

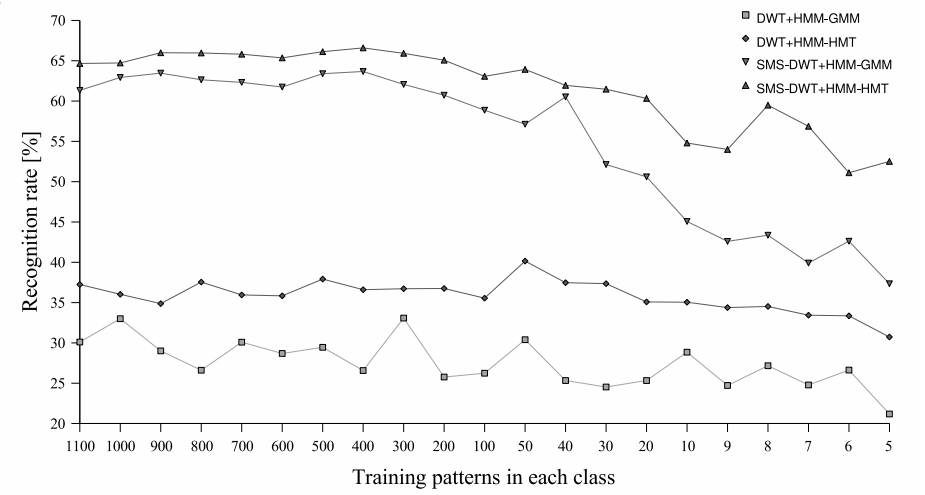

Fig. 1. Performance of proposed architecture for classification using different amounts of training patterns.

novel architecture and training algorithm demonstrated to be the most robust to the reduction of the amount of training data.

Future works will be oriented to reduce the computational cost of the training algorithms and to test the proposed model in continuous speech recognition and contaminating the speech with non-stationary noises.

Acknowledgment: This work is supported by the National Research Council for Science and Technology (CONICET), the National Agency for the Promotion of Science and Technology (ANPCyT-UNL PICT 11-25984 and ANPCyT-UNER PICT 11-12700), and the National University of Litoral (UNL, project CAID 012-72).

References

1. Crouse, M., Nowak, R., Baraniuk, R.: Wavelet-based statistical signal processing using hidden Markov models. IEEE Transactions on Signal Processing **46**(4) (1998) 886–902
2. Mallat, S.: A Wavelet Tour of signal Processing. 2nd edn. Academic Press (1999)
3. Fan, G., Xia, X.G.: Improved hidden Markov models in the wavelet-domain. IEEE Transactions on Signal Processing **49**(1) (2001) 115–120
4. Durand, J.B., Gonçalvès, P., Guédon, Y.: Computational methods for hidden Markov trees. IEEE Transactions on Signal Processing **52**(9) (2004) 2551–2560
5. Sebe, N., Cohen, I., Garg, A., Huang, T.: Machine Learning in Computer Vision. Springer (2005)
6. Baldi, P., Brunak, S.: Bioinformatics: The Machine Learning Approach. MIT Press, Cambrige, Masachussets (2001)

7. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambrige, Masachussets (1999)
8. Kim, S., Smyth, P.: Segmental Hidden Markov Models with Random Effects for Waveform Modeling. Journal of Machine Learning Research **7** (2006) 945–969
9. Bishop, C.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
10. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice-Hall, New Jersey (1993)
11. Bengio, Y.: Markovian Models for Sequential Data. Neural Computing Surveys **2** (1999) 129–162
12. Fine, S., Singer, Y., Tishby, N.: The Hierarchical Hidden Markov Model: Analysis and Applications. Machine Learning **32**(1) (1998) 41–62
13. Murphy, K., Paskin, M.: Linear time inference in hierarchical HMMs. In Dietterich, T., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14. Volume 14., Cambridge, MA, MIT Press (2002)
14. Willsky, A.: Multiresolution Markov models for signal and image processing. Proceedings of the IEEE **90**(8) (2002) 1396–1458
15. Dasgupta, N., Runkle, P., Couchman, L., Carin, L.: Dual hidden Markov model for characterizing wavelet coefficients from multi-aspect scattering data. Signal Processing **81**(6) (2001) 1303–1316
16. Lu, J., Carin, L.: HMM-based multiresolution image segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Volume 4. (2002) 3357–3360
17. Bengio, S., Bourlard, H., Weber, K.: An EM algorithm for HMMs with emission distributions represented by HMMs. Technical Report IDIAP-RR 11, Martigny, Switzerland (2000)
18. Weber, K., Ikbal, S., Bengio, S., Bourlard, H.: Robust speech recognition and feature extraction using HMM2. Computer Speech & Language **17**(2-3) (2003) 195–211
19. Bharadwaj, P., Carin, L.: Infrared-image classification using hidden Markov trees. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(10) (2002) 1394– 1398
20. Ichir, M., Mohammad-Djafari, A.: Hidden Markov models for wavelet-based blind source separation. IEEE Transactions on Image Processing **15**(7) (2006) 1887–1899
21. Zue, V., Sneff, S., Glass, J.: Speech database development: TIMIT and beyond. Speech Communication **9**(4) (1990) 351–356
22. Stevens, K.: Acoustic phonetics. MIT Press, Cambrige, Masachussets (1998)
23. Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **11**(7) (1989) 674–693
24. Daubechies, I.: Ten Lectures on Wavelets. Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia (1992)
25. Evangelista, G.: Pitch-synchronous wavelet representations of speech and music signals. IEEE Transactions on Signal Processing **41**(12) (1993) 3313–3330
26. Chan, C.P., Ching, P.C., Leea, T.: Noisy speech recognition using de-noised multiresolution analysis acoustic features. J. Acoust. Soc. Am. **110**(5) (2001) 2567–2574
27. Farooq, O., Datta, S.: Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition. IEEE Signal Processing Letters **8**(7) (2001)