CONTEXTUAL HIDDEN MARKOV MODELS

Mathieu Radenen, Thierry Artières

LIP6, Université Pierre et Marie Curie, Paris, France

ABSTRACT

Multiple works have proposed extensions of HMMs for handling variability. We focus here on the design of HMMs whose probability distribution on sequences depends on additional external variables that we call the context, which may stand for emotion features in speech recognition, physical features in gesture recognition, etc. We show experimentally the relevance of the method for handwriting recognition.

Index Terms— Hidden Markov Model, Handwriting Recognition, Context variables

1. INTRODUCTION

Hidden Markov Models are a famous class of probabilistic generative model known for their simplicity and robustness at classifying and labeling sequences. Despite their popularity, they rely on several assumptions and a limited expressive power. Among the principal shortcomings of the HMM is that their states are mutually exclusive. So, it requires N states to get N different output distributions.

The easiest way to handle variability in HMMs consists in increasing the number of states, in increasing the size of Gaussian mixtures, in using context dependent unit (e.g. phone) models. Such a strategy is easy to implement but leads to an increased number of parameters yielding estimation difficulty and overfitting. To overcome this difficulty the speech recognition community has focused on ways to share parameters between states, for instance by using tied mixtures where a pool of Gaussian is learned on all frames and where only mixture weights are learned for every state or by directly sharing states between different phone models [1].

Yet all these strategies allow handling *local* variability, at the state level, but not global variability factors that affect the whole sequence. The starting point of this work is that an important part of the variability between observation sequences may be the consequence of a few contextual variables (which may be hidden or observed) that remain fixed all along a sequence or that vary slowly with time. For instance a sentence may be uttered quite differently according to the speaker emotion. A gesture may have more amplitude if it is performed slower, and its overall shape depends on the weight and on the height of the performer. Such a global variability cannot always be removed through preprocessing or normalization and would benefit from a specific handling in HMMs.

Few researchers have tackled this problem, by designing a HMM whose probability distribution on sequences depends on external variables, that we call here the context of the sequence (that we note θ). [2] proposed Parametric Hidden Markov Models where the means of Gaussian distribution vary linearly as a function of the context. As the output distribution depends not only on the state but also on the context, a model may express many distribution with a limited number of additional parameters. [3], [4] and [5] investigated rather similar approaches. All these approaches differ by the nature of the dependency of HMM parameters to context variables, the ability to deal with dynamic context variables, the ability to infer context variables at test time.

We build here upon these pioneer works and propose a framework for conditioning the probability distribution of a HMM (means and covariance matrices) on a set of external variables, that may vary in time, and that may eventually be inferred at test time. In the following sections, we first introduce our modeling framework and detail the two main cases where the context remains fixed all along the sequence (time independent modeling) and when it is dynamic. Then we compare our approach with previous works. Finally we report experimental results on a handwriting recognition task.

2. CONTEXTUAL HIDDEN MARKOV MODEL (CHMM)

In the following we focus first on the case of single Gaussian CHMM when θ is static and remain fixed all along a sequence. Then we discuss variants including dealing with dynamic θ and using Gaussian mixtures.

2.1. Time independent CHMM

2.1.1. CHMM modeling

First, assume that we are given a set of external (contextual) variables θ (vector of dimension c) for any observation sequence $\mathbf{x} = (x_1, ..., x_T)$ where x_t's are d-dimensional feature vectors (e.g. θ may be the length of $\mathbf{x}$). We consider HMMs where means and covariance matrices depend on θ. Considering first single Gaussian models, we define the mean $\hat{\mu}_j$ (d-dimensional vector) and the covariance matrix $\hat{\Sigma}_j$ ($d \times d$

matrix) of the Gaussian distribution in state j as:

$$\hat{\mu}_j(\theta) = W_j^{\mu}\theta + \bar{\mu}_j$$
$$\hat{\Sigma}_j(\theta) = D_j(\theta) \times \bar{\Sigma}_j \times D_j(\theta)$$
$$\text{with } D_j(\theta) = diag(exp(W_j^{\Sigma}\theta + \widetilde{\Sigma}_j))$$

with W_j^{μ} and W_j^{Σ} two $d \times c$ matrices of linear transform for μ and Σ parameterization, and $\bar{\mu}_j$ and $\widetilde{\Sigma_j}$ their offset coefficients vector. Also we note the exponential of a matrix A, $exp(A)$, to be the matrix of the exponential function applied componentwize to all elements of A, and we note $diag$ the function transforming a vector to a diagonal matrix. The use of the exponential function ensures elements of $D_j(\theta)$ to be strictly positive, which makes $\hat{\Sigma}_j(\theta)$ a valid covariance matrix provided $\bar{\Sigma}_j$ is one (note that $D_j(\theta)$, $\hat{\Sigma}_j(\theta)$ and $\bar{\Sigma}_j$ are all $d \times d$ matrices).

Actually the shape of $\hat{\mu}_j(\theta)$ makes it linearly dependent on θ while the shape of the covariance matrix makes the term at u^{th} row and v^{th} column equal to:

$$\hat{\Sigma}_j(\theta)(u,v) = D_j(u,u) \times D_j(v,v) \times \bar{\Sigma}_j(\theta)(u,v)$$

Figure 1 shows the effect of such a parameterization on the shape of a covariance matrix. The original covariance matrix defines a shape in the upper left Figure which is modified by various D matrix (three other plots).

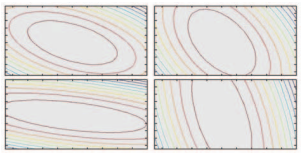

Fig. 1: Parameterization of the covariance matrix (Top left) with various D matrices: $D = diag([1\ 2])$ (Top right), $D = diag([2\ 0.9])$ (Bottom left), $D = diag([0.8\ 3])$ (bottom right).

It is interesting to note that such a model subsumes standard HMM (with means $\bar{\mu}_j$ and covariance matrices $\bar{\Sigma}_j$) by setting W_j^{μ} and W_j^{Σ} to null matrices and by setting $\widetilde{\Sigma_j}$ to null vectors. Also, considering only the mean parameterization yields Parametric HMM as proposed in [2].

2.1.2. Training

To keep notation more compact, we first define the $d \times (c+1)$ matrices $Z_j^{\mu} = \left[W_j^{\mu}\ \ \bar{\mu_j}\right]$, $Z_j^{\Sigma} = \left[W_j^{\Sigma}\ \ \widetilde{\Sigma_j}\right]$ and the column vector $\Omega^k = \left[\theta^k\ \ 1\right]^T$ for sequence k. We consider we get a set of training sequences along with their labels (i.e. classes) and their context variables, $\left\{(\mathbf{x}^k, y^k, \theta^k)\right\}$.

Training consists in modifying the matrices Z_j^{μ} and Z_j^{Σ} so as to maximize the likelihood of the training sequences. It is performed in two steps:

- First, we learn a HMM with parameterized means only, which is equivalent to learning a Parametric HMM [2]. This may be done by using following formulas:

$$Z_j^{\mu} = \left[\sum_{k,t} \gamma_{k,t,j} x_t^k \Omega^{k^T}\right] \left[\sum_{k,t} \gamma_{k,t,j} \Omega^k \Omega^{k^T}\right]^{-1} \quad (1)$$

$$\Sigma_j = \frac{\sum_{k,t} \gamma_{k,t,j} (x_t^k - \hat{\mu}_j(\theta^k))(x_t^k - \hat{\mu}_j(\theta^k))^T}{\sum_{k,t} \gamma_{k,t,j}} \quad (2)$$

  where $\gamma_{k,t,j}$ stands for the usual probability used in standard HMM theory $p(q_t = j|\mathbf{x}^k, y^k)$.

  Then, for every state j, we set $\bar{\Sigma}_j = \Sigma_j$

- Fixing all models parameters, we reestimate the Z_j^{Σ}
  We initialize $Z_j^{\Sigma} = 0$ which allows starting from the covariance matrix obtained in first step: $\hat{\Sigma}_j(\theta) = \Sigma_j$
  Reestimation of Z_j^{Σ} is performed via the Generalized Expectation Maximization algorithm, by computing the derivative of the auxiliary function Q (see [1]) with respect to Z_j^{Σ} and doing a gradient ascent. Omitting details one can show without difficulty that:

$$\frac{\partial Q}{\partial Z_j^{\Sigma}} = \sum_{k,t,i} M_{k,t,j}(i,i) \times \frac{\partial D_{k,j}^{-1}(i,i)}{\partial Z_j^{\Sigma}} \quad (3)$$

  with $M_{k,t,j} =$

$$\gamma_{k,t,j}\left[D_{k,j} - \bar{\Sigma}_j^{-1} D_{k,j}^{-1}(x_t^k - \hat{\mu}_j(\theta^k))(x_t^k - \hat{\mu}_j(\theta^k))^T\right]$$

  where

$$\frac{\partial D_{k,j}^{-1}(i,i)}{\partial Z_j^{\Sigma}(m,n)} = \begin{cases} \frac{-\Omega^k(n)}{D_{k,j}(i,i)} & \text{if } i = m \\ 0 & \text{otherwise.} \end{cases}$$

2.2. Extensions

2.2.1. Time dependent CHMM

Now suppose θ depends on time. We then use the following definition of the pdf in a state j:

$$\hat{\mu}_j(\theta_t) = Z_j^{\mu}\Omega_t$$
$$\hat{\Sigma}_j(\theta_t) = D_j(\theta_t) \times \bar{\Sigma}_j \times D_j(\theta_t)$$

It is straightforward to show that the reestimation formulas (1, 2, 3) apply if one changes systematically θ to θ_t. As a result, Ω^k becomes $\Omega_t^k = \left[\theta_t^k\ \ 1\right]^T$ and $D_{k,j}$ becomes $D_{k,t,j}$. New reestimation formulas are then simple extensions of (1, 2, 3). For instance, the closed form solution for Z_j^{μ} becomes:

$$Z_j^{\mu} = \left[\sum_{k,t} \gamma_{k,t,j} x_t^k \Omega_t^{k^T}\right] \left[\sum_{k,t} \gamma_{k,t,j} \Omega_t^k \Omega_t^{k^T}\right]^{-1}$$

2.2.2. Gaussian mixtures

Extending single Gaussian models to Gaussian mixture modeling may be done easily. The new pdf of l^{th} Gaussian in state j is then defined as:

$$\hat{\mu}_{j,l}(\theta_t) = Z^{\mu}_{j,l}\Omega_t$$
$$\hat{\Sigma}_{j,l}(\theta_t) = D_{j,l}(\theta_t) \times \bar{\Sigma}_{j,l} \times D_{j,l}(\theta_t)$$

There is no difficulty to derive new reestimation formulas by adding component index l in (1, (2) and (3)).

3. RELATED WORKS

A first attempt for conditioning HMM parameters on environment variables seems to be the work from [2] who proposed Parametric HMMs (PHMMs) for gesture recognition, context variables were related to the amplitude of the gestures. As we already said our modeling framework includes PHMM as a special case when ignoring parameterization of covariance matrices. A very similar approach (Multiple Regression HMM, or MR-HMM) has been proposed in [6] for speech recognition, using fundamental frequency as context variable. Basically MR-HMM may be viewed as PHMM with time dependent context variables θ. These models are again embedded in our framework.

A second class of models called Variable Parameter HMMs (VPHMMs) are closely related to our approach. This type of model has been introduced in [3], [4]. It was proposed in the context of speech recognition to improve robustness to noisy conditions. In this approach the means as well as the (diagonal) covariance matrices are expressed as a polynomial function of a static scalar environment variable. Our approach could be easily extended this way by using polynomial expansion of the vector of context variables θ, yielding polynomial dependency. In addition our approach may use time depend multidimensional θ vectors and allows dealing with full covariance matrices. Finally, our approach, alike the one in [2], potentially allows (although we do not report results here) inferring the environment variables at test time.

Finally the work by Dong Yu and Li Deng [5] is probably the most achieved approach of this kind, it has been designed as [4] for noisy speech recognition. It refines VPHMM using piece-wise spline interpolation instead of polynomial regression and handle time dependent environment variable, together with discriminative training. Yet their modeling assume diagonal covariance matrices and allow one-dimensional conditioning variables only. This makes VPHMM not so well adapted to exploit "style environment variables" like emotion, gender, height... that could a priori affect all dimensions of the mean and covariance.

In this context the strength of our approach lies in the conditioning of means and of full covariance matrices with a multidimensional vector of context variables.

4. EXPERIMENTAL RESULTS

We report here comparative results of standard HMMs and of CHMMs on an off-line handwriting dataset [7]. Every sequence is an image of an isolated handwritten letter which is preprocessed and represented at the end as a sequence of 9-dimensional feature vectors. We used 200 sequences for training, 50 sequences for validation and 50 for testing, for each of the 23 classes (3 classes, i.e. letters, have been removed because under represented).

We use an 8 states left right model (HMM or CHMM) for each class (letter) with full covariance Gaussian pdfs. CHMMs and HMMs were trained up to convergence with a maximum of 150 EM iterations. The training of CHMM with covariance parameterization had an additional 150 GEM iterations with one gradient step by EM iteration. In both cases, model selection is performed as the set of models, at a given iteration, that performs best on the validation set. Initialization is performed according to a linear alignment of training sequences on the left-right models: every training sequence is divided in a number of segments of equal length, one segment per state. Reestimation formulas are then used with this linear alignment. In case we use Gaussian mixtures, means and covariance matrices of a mixture are initialized by Kmeans on the set of all observations aligned with the state.

We explored a few definitions of external variables θ, all are quantities that are computed from the observation sequence. We used the mean (d dimensional vector noted 'μ') and the variance (d dimensional vector noted 'σ^2') computed on the full sequence. We also tried the instantaneous derivative of the sequence averaged on the whole sequence (noted 'Δ'), and the instantaneous acceleration, also averaged on the whole sequence (noted 'Δ^2'). In addition to these static context vectors we investigated dynamic ones, where these quantities are computed on a sliding window rather than on the whole sequence. This allows extending the approach to sequence labeling tasks where one has to simultaneously segment an input sequence into characters and recognize the characters. This could be done by exploiting θ values that depend on segmentation but this would lead to a costful dynamic programming step. Alternatively one can exploit time varying θ by computing them locally in the sequence, which is what we investigated. A number in parenthesis suffixing a context variable name, like '$\mu(5)$', means that θ is a function of time and is averaged over a window of 5 frames centered at current time (e.g. $\theta_t = mean(x_{t-2}, ..., x_{t+2})$).

We first report in Figure 2 the performance of HMMs wrt the size of Gaussian mixtures. The accuracy in test increases up to a plateau while accuracy still increases on training set, showing the difficulty of learning more complex models.

Next we report in Figure 3 results using parameterization of the means only (μCHMM), and of means and covariance matrices ($\mu\Sigma$CHMM), with static and with dynamic θ's. Here CHMMS are single Gaussian models (with 8 states). In this

Figure, θ is computed as the vector of variances of frame features, computed on the whole sequence or locally on a sliding window of increasing size (absciss). As may be seen all CHMM with either static θ or with dynamic $\theta(t)$ improve over standard single Gaussian HMMs (60.5% accuracy). Although static θ work already well, finding a good set up of dynamic θ (e.g. window size) is harder. Yet equivalent or slightly better results may be obtained with dynamic variables, meaning that the extension of this framework to signal labeling (e.g. continuous speech recognition), where a static θ is less relevant, should not be a problem. Finally, note that the covariance parameterization gives an additional improvement over mean only CHMMs. Figure 4 is similar but this time $\theta = \mu$ or $\theta = \mu(t)$. One sees that as before CHMMs outperform single Gaussian HMM and that, more interestingly, single Gaussian CHMMs outperform the best HMM models whatever the size of Gaussian mixtures.

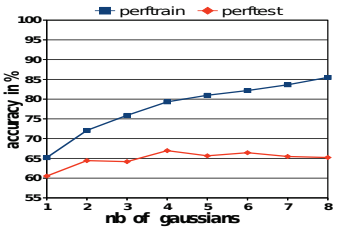

Fig. 2: 8 states Gaussian mixtures HMMs

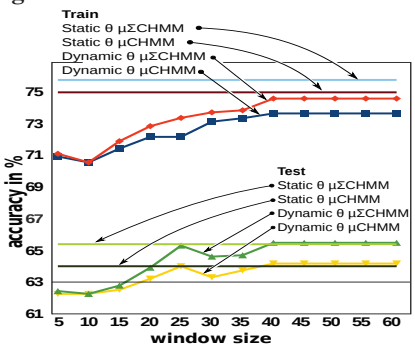

Fig. 3: 8 states CHMM with $\theta = \sigma^2$ or $\theta_t = \sigma^2(t)$

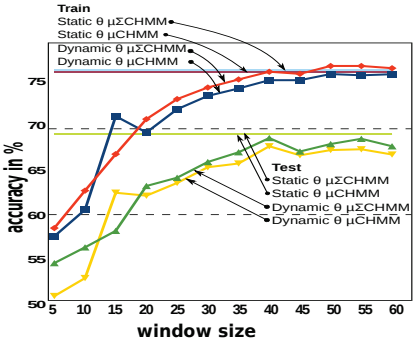

Fig. 4: 8 states CHMM with $\theta = \mu$ or $\theta_t = \mu(t)$

Lastly, Table 1 shows a number of richer modeling using dynamic θ mixing μ, σ, Δ, and Δ^2 context variables, for learning 8 states left-right CHMMs with Gaussian mixtures of size 1 to 4. It appears clearly here that one can get

nb gauss	θ	Train μCHMM	Train $\mu\Sigma$CHMM	Test μCHMM	Test $\mu\Sigma$CHMM
1	$\mu(60)\ \sigma^2(55)$	80.7	81.2	67.3	69.0
1	$\mu(60)\ \sigma^2(55)\ \Delta(15)$	82.6	82.6	68.4	68.4
1	$\mu(60)\ \sigma^2(55)\ \Delta(15)\ \Delta^2(15)$	84.9	85.3	70.3	70.8
2	$\mu(60)\ \sigma^2(55)\ \Delta(15)\ \Delta^2(15)$	97.8	92.3	71.48	**73.2**
3	$\mu(60)\ \sigma^2(55)\ \Delta(15)\ \Delta^2(15)$	95.3	95.3	72.22	72.2
4	$\mu(60)\ \sigma^2(55)\ \Delta(15)\ \Delta^2(15)$	97.3	97.3	71.9	71.9

Table 1: 8 states CHMM using mixed $\theta(t)$

significant improvement over single Gaussian HMMs, and over previous single Gaussian CHMM, by combining context variables. Furthermore combining CHMM modeling with increasing the size of Gaussian mixtures interestingly still yields improvements. This may mean that these are two complementary ways of modeling and capturing variability. Overall, the best model uses mixed context variables and mixtures of two Gaussian and outperforms best standard HMMs in Figure 2 by more than 6% accuracy.

5. CONCLUSION

We proposed a complete framework for learning contextual Hidden Markov Models where means and full covariance matrices are defined as function of external, i.e. context, variables. We show that this type of modeling global variability significantly improve over standard HMMs. Moreover such a modeling may be combined to more traditional ways of handling variability such as increasing Gaussian mixture size.

6. REFERENCES

[1] Juang B-H. Rabiner, L., *Fundamentals of speech recognition*, Prentice Hall, Prentice Hall, 1993.

[2] Bobick A.F. Wilson, A.D., "Parametric HMMs for gesture recognition," *IEEE Trans. on PAMI*, vol. 21, no. 9, pp. 884 –900, sep 1999.

[3] Y. Gong X. Cui, "Variable parameter Gaussian mixture Hidden Markov Modeling for speech recognition," in *ICASSP '03*, april.

[4] Y. Gong X. Cui, "A study of variable-parameter Gaussian mixture Hidden Markov Modeling for noisy speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1366 –1376, may 2007.

[5] Y. Gong A. Acero D. Yu, L. Deng, "A novel framework and training algorithm for variable-parameter HMMs," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1348 –1360, sept. 2009.

[6] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression HMM," in *ICASSP '01*.

[7] Bunke H. Marti, U.-V., "A full english sentence database for off-line handwriting recognition," in *International Conference on Document Analysis and Recognition, 1999*.