IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

# An EM Algorithm for HMMs with Emission Distributions Represented by HMMs


Samy Bengio [1] Hervé Bourlard [2]
Katrin Weber [3]



IDIAP–RR 00-11

May 2000



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
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1 *bengio@idiap.ch*
2 *bourlard@idiap.ch*
3 *weber@idiap.ch*

IDIAP Research Report 00-11

# An EM Algorithm for HMMs with Emission Distributions Represented by HMMs

Samy Bengio Hervé Bourlard Katrin Weber

May 2000

**Abstract.** A novel approach to represent emission distributions of Hidden Markov Models is presented in this paper. Whereas they are usually estimated with Gaussian mixtures or neural networks, we propose to estimate them with another HMM, but in feature space. This representation, referred here as HMM$^2$, could enable the model to more accurately represent feature correlations with fewer parameters than standard HMMs. A full derivation of an EM algorithm is given in order to globally train all the HMM$^2$ parameters. Preliminary experiments on speech data show promising results.

# 1 Introduction

Hidden Markov Models (HMMs) are statistical models for sequential data that have been used successfully in many applications in artificial intelligence, pattern recognition, speech processing and biological sequence modeling [3]. Emission probabilities of HMMs are typically represented using mixtures of Gaussians or neural networks. In this paper, we propose an alternative approach where they are estimated by yet another HMM. By this, we mean that the vector of features at a particular time step is considered as a fixed length sequence, which has supposedly been generated by another HMM for which each state is emitting individual feature components. We call this approach HMM$^2$.

We believe that HMM$^2$ (which includes the classical mixture of Gaussian HMMs as a particular case) has several potential advantages, including: better modeling of the correlation accross features, more flexible modeling capabilities with fewer parameters, and for instance in the case of speech recognition, better modeling of the time/frequency underlying structure.

In this paper, we mainly derive the expectation-maximization (EM) algorithm to train HMM$^2$s. Since an HMM is a special kind of mixture of distributions, an HMM$^2$ is thus a mixture of mixture of distributions, which is also a mixture. It should then be natural that an EM algorithm can be derived when the emission and transition probabilities of the *internal* (feature-based) HMMs are represented by mixtures of Gaussians and multinomials respectively. We propose here one such derivation. After some notation, we derive the main EM equations, and we show how to compute the E-step and the M-step for particular instances of the transition and emission probabilities of the internal HMMs. Finally, we give some preliminary experimental results.

# 2 Notation

Figure 1 gives a graphical illustration of the HMM$^2$ model. We define

- $y_t$ the observed vector at time $t$, and $y_{t,s}$ its $s^{th}$ feature component,
- $p(y_t|q_t{=}i)$ the probability to emit vector $y_t$ in state $i$, and $p(y_{t,s}|r_s{=}l, q_t{=}i)$ the probability to emit feature $y_{t,s}$ in internal state $l$ of the HMM in external state $i$,
- $P(q_t{=}i|q_{t-1}{=}j)$ the probability to go from state $j$ at time $t-1$ to state $i$ at time $t$, and $P(r_s{=}l|r_{s-1}{=}m, q_t{=}i)$ the probability to go from internal state $m$ at feature $s-1$ to internal state $l$ at feature $s$ while in external state $i$ at time $t$,
- $N$ the number of states in the external HMM, and $N_i$ the number of states in the HMM of state $i$.

The likelihood of one sequence of the data given the model is then

$$L(Y|\theta) = p(y_1^T|\theta) \tag{1}$$

where $T$ is the size of the sequence and $y_1^T$ is the sequence $\{y_1, y_2, \cdots, y_T\}$.

# 3 General EM derivation

Following the general idea of EM, one has to select hidden variables such that the knowledge of these variables would simplify the learning problem. Let us now introduce the hidden variable $Z$ that gives the state $q$ of the external HMM at each time step $t$ such that

$$z_{i,t} = \begin{cases} 1 & \text{if } q_t = i \\ 0 & \text{else} \end{cases} \tag{2}$$

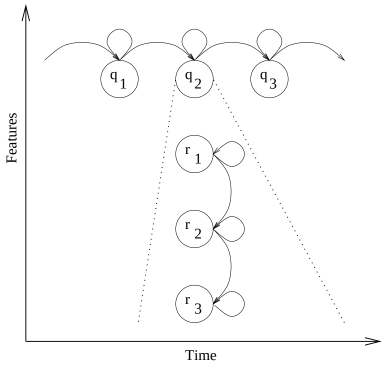


Figure 1: The HMM[2]: emission distributions of an HMM are estimated by an HMM.

and the hidden variable $W$ that gives the state $r$ of the internal HMMs at each feature $s$ such that

$$w_{i,l,s,t} = \begin{cases} 1 & \text{if } q_t{=}i \text{ and } r_s{=}l \\ 0 & \text{else.} \end{cases} \tag{3}$$

As usually done, we consider that the temporal and feature sequences have been generated by a $1^{st}$ order Markov process, yielding:

$$P(q_t|q_1^{t-1}) = P(q_t|q_{t-1}) \tag{4}$$

$$P(r_s|r_1^{s-1}, q_t) = P(r_s|r_{s-1}, q_t). \tag{5}$$

Moreover, we will assume that conditionally on the current state $q_t$ (respectively $r_s$), the probability of the current emission vector $y_t$ ($y_{t,s}$) is independant of the previous vectors $y_1^{t-1}$ ($y_{t,1}^{t,s-1}$):

$$P(y_t|q_t, y_1^{t-1}) = P(y_t|q_t) \tag{6}$$

$$P(y_{t,s}|r_s, y_{t,1}^{t,s-1}, q_t) = P(y_{t,s}|r_s, q_t). \tag{7}$$

Using previous assumptions, one can show that the complete joint likelihood of one sequence of the data (a generalization to many sequences is straightforward) and the hidden variables is as follows:

$$L_c(Y, Z, W) = P(q_0) \prod_{t=1}^{T} \prod_{i=1}^{N} p(y_t|q_t{=}i)^{z_{i,t}} \prod_{j=1}^{N} P(q_t{=}i|q_{t-1}{=}j)^{z_{i,t} \cdot z_{j,t-1}} \tag{8}$$

which has the same form as the complete likelihood in standard HMMs, but where the emission probability is expressed as follows:

$$p(y_t|q_t{=}i) = P(r_0|q_t{=}i) \prod_{s=1}^{S} \prod_{l=1}^{N_i} p(y_{t,s}|r_s{=}l, q_t{=}i)^{w_{i,l,s,t}} \prod_{m=1}^{N_i} P(r_s{=}l|r_{s-1}{=}m, q_t{=}i)^{w_{i,l,s,t} \cdot w_{i,m,s-1,t}} \tag{9}$$

Including equation (9) into equation (8) and taking the log we obtain:

$$\begin{aligned} \log L_c(Y, Z, W) &= \log P(q_0) \\ &+ \sum_{t=1}^{T} \sum_{i=1}^{N} z_{i,t} \left( \begin{array}{l} \log P(r_0|q_t{=}i) + \sum_{s=1}^{S} \sum_{l=1}^{N_i} w_{i,l,s,t} \log p(y_{t,s}|r_s{=}l, q_t{=}i) \\ + \sum_{s=1}^{S} \sum_{l=1}^{N_i} \sum_{m=1}^{N_i} w_{i,l,s,t} \cdot w_{i,m,s-1,t} \log P(r_s{=}l|r_{s-1}{=}m, q_t{=}i) \end{array} \right) \\ &+ \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} z_{i,t} \cdot z_{j,t-1} \log P(q_t{=}i|q_{t-1}{=}j) \end{aligned} \tag{10}$$

where $P(q_0)$ is the initial state probability of the external HMM and $P(r_0|q_t{=}i)$ is the initial state probability of the internal HMM in state $i$. We define an auxiliary function in the same way as in the standard EM approach for HMMs as follows:

$$Q(\theta|\theta^k) = E[\log L_c(Y, Z, W|\theta)|Y, \theta^k] \tag{11}$$

where the expectation is over the hidden variables $Z$ and $W$. Moving the expectation inside the log and expanding the log likelihood as in equation (10) yields

$$\begin{aligned} Q(\theta|\theta^k) &= \log P(q_0) + \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{N}\hat{\gamma}(i,t)\left(\begin{array}{l} \log P(r_0|q_t{=}i) + \sum_{s=1}^{S}\sum_{l=1}^{N_i}\hat{\gamma}_{i,t}(l,s)\log p(y_{t,s}|r_s{=}l, q_t{=}i) \\ + \sum_{s=1}^{S}\sum_{l=1}^{N_i}\sum_{m=1}^{N_i}\hat{\xi}_{i,t}(l,m,s)\log P(r_s{=}l|r_{s-1}{=}m, q_t{=}i) \end{array}\right) \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\xi}(i,j,t)\log P(q_t{=}i|q_{t-1}{=}j) \end{aligned} \tag{12}$$

with

$$\hat{\gamma}(i,t) \stackrel{\text{def}}{=} E[z_{i,t}|y_1^T;\theta^k] \tag{13}$$

$$\hat{\xi}(i,j,t) \stackrel{\text{def}}{=} E[z_{i,t}, z_{j,t-1}|y_1^T;\theta^k] \tag{14}$$

$$\hat{\gamma}_{i,t}(l,s) \stackrel{\text{def}}{=} E[w_{i,l,s,t}|y_1^T;\theta^k] \tag{15}$$

$$\hat{\xi}_{i,t}(l,m,s) \stackrel{\text{def}}{=} E[w_{i,l,s,t}, w_{i,m,s-1,t}|y_1^T;\theta^k]. \tag{16}$$

The E-step of the EM thus consists in computing the expectations of equations (13) to (16), while the M-step consists in finding the parameters $\theta$ which maximize equation (11). Thus, at the $k^{th}$ iteration, one computes

$$\theta^{k+1} = \arg\max_{\theta} Q(\theta|\theta^k) \tag{17}$$

It can be shown that maximizing $Q()$ also maximizes the likelihood of the data [2]. In the next two sections, we detail the computations involved in the E- and M-step.

# 4 E-step

$\gamma(i,t)$ and $\xi(i,j,t)$ are estimated exactly as in an ordinary HMM, by first introducing the intermediate variables $\alpha(i,t)$ and $\beta(i,t)$ as follows. Let us define $\alpha(i,t)$ as the probability to emit the given sequence up to time $t$ and being in state $i$ at time $t$. This can be computed efficiently and recursively as follows[1]:

$$\begin{aligned} \alpha(i,t) &\stackrel{\text{def}}{=} p(y_1^t, q_t{=}i) \\ &= \sum_{j\in pred(i)} p(y_t|q_t{=}i)P(q_t{=}i|q_{t-1}{=}j)\alpha(j,t-1) \end{aligned} \tag{18}$$

where $pred(i)$ is the set of possible predecessor states of state $i$. In the same way, we can define $\beta(i,t)$ as the probability to emit the given sequence from time $t+1$ to the end of the sequence and being in state $i$ at time $t$. Again, this can be computed recursively as follows:

$$\begin{aligned} \beta(i,t) &\stackrel{\text{def}}{=} p(y_{t+1}^T|q_t{=}i) \\ &= \sum_{j\in succ(i)} p(y_{t+1}|q_{t+1}{=}j)P(q_{t+1}{=}j|q_t{=}i)\beta(j,t+1) \end{aligned} \tag{19}$$

[1]The full derivations of $\alpha$, $\beta$, $\gamma$ and $\xi$ can be found in [1].

where $succ(i)$ is the set of possible successor states of state $i$. Finally, given $\alpha(i,t)$ and $\beta(i,t)$, one can compute $\gamma(i,t)$ and $\xi(i,j,t)$ efficiently:

$$\begin{aligned} \gamma(i,t) &= E[z_{i,t}|y_1^T] \\ &= P(q_t{=}i|y_1^T) \\ &= \frac{\alpha(i,t)\cdot\beta(i,t)}{L} \end{aligned} \tag{20}$$

where $L$ is the likelihood of the sequence and can be computed using the $\alpha(i,T)$:

$$\begin{aligned} L &= \sum_i p(y_1^T, q_T{=}i) \\ &= \sum_i \alpha(i,T) \end{aligned} \tag{21}$$

and

$$\begin{aligned} \xi(i,j,t) &= E[z_{i,t}, z_{j,t-1}|y_1^T] \\ &= P(q_t{=}i, q_{t-1}{=}j|y_1^T) \\ &= \frac{\alpha(j,t-1)P(q_t{=}i|q_{t-1}{=}j)p(y_t|q_t{=}i)\beta(i,t)}{L}. \end{aligned} \tag{22}$$

Similarly, to estimate $\gamma_{i,t}(l,s)$ and $\xi_{i,t}(l,m,s)$, we first introduce the intermediate variables $\alpha_{i,t}(l,s)$ and $\beta_{i,t}(l,s)$. $\alpha_{i,t}(l,s)$ is the probability to emit the feature vector at time $t$ up to feature $s$ and being in state $i$ of the external HMM and in state $l$ of the internal HMM:

$$\begin{aligned} \alpha_{i,t}(l,s) &\stackrel{\text{def}}{=} p(y_{t,1}^{t,s}, r_s{=}l|q_t{=}i) \\ &= \sum_{m\in pred(l)} p(y_{t,s}|r_s{=}l, q_t{=}i)P(r_s{=}l|r_{s-1}{=}m, q_t{=}i)\alpha_{i,t}(m,s-1). \end{aligned} \tag{23}$$

In the same way, $\beta_{i,t}(l,s)$ is the probability to emit the feature vector at time $t$ from feature $s+1$ to feature $S$, while being in state $i$ of the external HMM and in state $l$ of the internal HMM:

$$\begin{aligned} \beta_{i,t}(l,s) &\stackrel{\text{def}}{=} p(y_{t,s+1}^{t,S}|r_s{=}l, q_t{=}i) \\ &= \sum_{m\in succ(l)} p(y_{t,s+1}|r_{s+1}{=}m, q_t{=}i)P(r_{s+1}{=}m|r_s{=}l, q_t{=}i)\beta_{i,t}(m,s+1). \end{aligned} \tag{24}$$

Using these intermediate variables, we can now estimate $\gamma_{i,t}(l,s)$ and $\xi_{i,t}(l,m,s)$:

$$\gamma_{i,t}(l,s) = \frac{\alpha_{i,t}(l,s)\cdot\beta_{i,t}(l,s)}{L_{i,t}} \tag{25}$$

where $L_{i,t}$ is the likelihood of $y_t$ in state $i$ and can be computed using the $\alpha_{i,t}(l,S)$:

$$\begin{aligned} L_{i,t} &= \sum_l p(y_{t,1}^{t,S}, r_S{=}l) \\ &= \sum_l \alpha_{i,t}(l,S) \end{aligned} \tag{26}$$

and finally we have

$$\xi_{i,t}(l,m,s) = \frac{\alpha_{i,t}(m,s-1)P(r_s{=}l|r_{s-1}{=}m, q_t{=}i)p(y_{t,s}|r_s{=}l, q_t{=}i)\beta_{i,t}(l,s)}{L_{i,t}}. \tag{27}$$

# 5 M-step

In the M-step, we search for the best value of $\theta$ given the estimated variables as well as the current value of $\theta^k$. We are thus looking for a new value of $\theta$ such that

$$\frac{\partial Q(\theta|\theta^k)}{\partial \theta} = 0. \tag{28}$$

This step can be done independently for each probability distribution. The M-step for transition probabilities of the external HMM is exactly the same as for normal HMMs since the proposed modification only alters the emission probabilities, and is thus not explained here.

We now give the equations for emission probabilities represented as HMMs in the case where the internal HMMs are simple: transition probabilities represented as multinomials and emission probabilities represented as diagonal Gaussians. As it will be shown, the equations are very similar to the ones of a standard HMM, simply normalizing them by a posterior over the current state of the external HMM.

## 5.1 Transition probabilities of the internal HMMs as multinomials

Let us represent the transition that goes from state $l$ to state $m$ of the internal HMM in state $i$ of the external HMM as $w_i(l,m)$. All $w_i(l,m)$ must be non-negative and

$$\sum_m w_i(l,m) = 1. \tag{29}$$

To force this constraint, we have to introduce Lagrange multipliers $\lambda_{i,l}$. Thus, instead of maximizing $Q()$, we will maximize

$$Q'(\theta|\theta^k) \stackrel{\text{def}}{=} Q(\theta|\theta^k) + \sum_i \sum_l \left(1 - \sum_m w_i(l,m)\right) \lambda_{i,l}. \tag{30}$$

Solving equation (28) then yields

$$\frac{\partial Q'(\theta|\theta^k)}{\partial w_i(l,m)} = \sum_t \gamma(i,t) \sum_s \frac{\xi_{i,t}(l,m,s)}{w_i(l,m)} - \lambda_{i,l} = 0 \tag{31}$$

and thus, to maximize $Q'()$, we have

$$\begin{aligned} w_i(l,m) &= \frac{\sum_t \gamma(i,t) \sum_s \xi_{i,t}(l,m,s)}{\lambda_{i,l}} \\ &= \frac{\sum_t \gamma(i,t) \sum_s \xi_{i,t}(l,m,s)}{\sum_t \gamma(i,t) \sum_s \gamma_{i,t}(l,s)} \end{aligned} \tag{32}$$

where $\lambda_{i,l}$ is chosen in order to normalize the distribution.

## 5.2 Emission probabilities of the internal HMM as diagonal Gaussians

Let us now represent the emission probability of the feature $y_{t,s}$ when emitted by internal state $l$ of the HMM in state $i$ at time $t$:

$$\log p(y_{t,s}|q_t{=}i, r_s{=}l) = -\left(\log \sigma_l + \frac{(y_{t,s} - \mu_l)^2}{2\sigma_l^2}\right) - \frac{\log 2\pi}{2}. \tag{33}$$

Solving again equation (28), we derive update equations for means $\mu_l$ and standard deviations $\sigma_l$ and obtain:

$$\mu_l^{k+1} = \frac{\sum_t \gamma(i,t) \sum_s y_{t,s} \cdot \gamma_{i,t}(l,s)}{\sum_t \gamma(i,t) \sum_s \gamma_{i,t}(l,s)} \tag{34}$$

and

$$\sigma_l^{k+1} = \frac{\sum_t \gamma(i,t) \sum_s \gamma_{i,t}(l,s)(y_{t,s} - \mu_l)^2}{\sum_t \gamma(i,t) \sum_s \gamma_{i,t}(l,s)}. \tag{35}$$

# 6 Preliminary results

We have tested this EM algorithm on a speech database (Numbers95, a telephone speech, free-format numbers, speeker independant database), comparing a standard HMM with emission probabilities modeled as mixtures of Gaussians to an HMM$^2$. The training set had 486537 frames, while the test set had 180348 frames. In order to take advantage of frequency correlations, each frame was coded into the spectral domain (and not into a cepstral domain as it is often done in the speech community). The total number of features per frame was 24. In both models, the external HMM architecture was the classical left-to-right 3-states-per-phone. The emission probabilities of the standard HMM was a mixture of 4 diagonal Gaussians (in 24 dimensions). The internal HMM of the HMM$^2$ had 3 states fully connected, where each emission probability of the internal HMM was modeled by a mixture of 4 Gaussians in 3 dimensions[2]. The number of parameters of the HMM$^2$ was thus less than half the number of parameters of the standard HMM.

The average negative log likelihood of the standard HMM for the 27 phonemes over the test set was 26.5, while it was only 11.1 for the HMM$^2$, even with half the number of parameters. These results show that an HMM$^2$ can take advantage of the correlation structure of the spectral features more accurately and with fewer parameters than a standard HMM. It is clear that these preliminary results need to be confirmed on a word recognition task (which will be reported in the final version of the paper), but they are already very promising.

# 7 Conclusion

In this paper, we have presented a new model, HMM$^2$, which is an extension of Hidden Markov Models for sequences where the emission probability distributions are themselves represented as Hidden Markov Models in the feature domain. We have provided an exact EM algorithm to train such models. Preliminary results comparing such models with standard HMMs with Gaussian mixtures showed promising performances in negative log likelihood minimization.

# References

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statistic.*, 41:164–171, 1970.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

[3] Laurence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[2] The vector of 24 features was thus segmented in a sequence of 8 3-dimensional vectors.