INCREMENTAL TRAINING OF CDHMMs USING BAYESIAN LEARNING

Claudio Vair, Massimiliano Mercogliano, Luciano Fissore
CSELT - Centro Studi E Laboratori Telecomunicazioni
Via G. Reiss Romoli, 274 – I-10148 Torino, Italy
[vair, fissore]@cselt.it
http://www.cselt.it

ABSTRACT

The Bayesian Learning approach (MAP - Maximum A Posteriori) can be used for the incremental training of Continuous Density Hidden Markov Models (CDHMM), performed through speech data collected in real applications. The effectiveness of MAP is heavily conditioned by the correct balance between the a-priori knowledge and the field training data.

In this paper we propose and evaluate several optimization methods of the MAP combination function, based either on maximum likelihood (ML) and heuristics criteria. To adjust the relevance of the a-priori knowledge we use the exponential forgetting technique into the MAP framework.

We present several tests that compare the error rate reduction as a function of the selected optimization method and of the size of adaptation data.

1. INTRODUCTION

It is well known that the correct balance between the information derived from real-life data and the knowledge of the acoustic models, initially trained by means of a large amount of read speech, is a key factor to achieve high performance on the application trials. It is also well known that the use of speech data coming from a specific application allows to focus the acoustic models on the running vocabulary, increasing the robustness of the speech recognizer against the most frequent acoustic environments.

In this paper, the parameters of the Gaussian mixture densities are adapted using the MAP estimation [1] with exponential forgetting mechanism [2] and performing the a-priori parameter estimation in a model based outline [3]. The optimal choice of the exponential forgetting factor (FF) into the HMM estimation step is crucial for the incremental training as it allows to balance the importance of the prior information and of the adaptation data. This factor depends on the amount and on the quality of the speech data available for the incremental training and on the mismatch among the prior models and the application environment.

In the paper, we have defined different optimization criteria for the FF computation based either on ML and heuristics methods. We present several tests that compare the error rate reduction as a function of the selected forgetting factor optimization method and of the size of the adaptation data. The experiments were run on a speaker-independent isolated speech recognition task, for the Italian language, with a vocabulary of around 600 city names.

2. INCREMENTAL MAP TRAINING

2.1 MAP training

The incremental training is performed using the Bayesian Learning (MAP): in that framework, the models parameters, estimated from the field data, are constrained to a known probability density function that convey the a-priori information.

We used a fixed prior density function as in [1] and estimated its parameters in a model-based outline [3]. Thus, the a-priori knowledge is readily available since it can be extracted from a laboratory model, trained on a large phonetically balanced, read speech data base.

The incremental training is achieved using the segmental MAP, equivalent to the use of the Viterbi algorithm in the ML estimation. The algorithm performs two iterative steps: first, the training frames X are aligned using the current model θi, getting the optimal state sequence Si; second, the new model parameters are computed on the complete data (X,S). The objective of segmental MAP is the optimization of the function:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} \max_{S} f(X,S|\theta)g(\theta|\Phi)$$

(1)

where f(X,S|θ) is the likelihood on the complete data and g(θ|Φ) is the a-priori density function. The objective function (1) is a good approximation of the MAP estimation with reduced computational load. According to the segmental MAP, the likelihood of the observation X=(x1,…,xT) assigned to a given state by the alignment procedure is:

$$f(X|\theta) = \prod_{t=1}^{T} \sum_{i=1}^{K} \omega_i N(x_t | \mu_i, \Sigma_i)$$

(2)

where N(·) is the normal density function and θ=(ω, μ, Σ) are the model parameters: mixture weights, mean vectors and covariance matrices.

2.2 The forgetting mechanism

The choice of a fixed prior density allows the use of the EM algorithm for MAP estimation by repeating the following steps:

Expectation: $R(\theta, \hat{\theta}) = E[\log f(X,S|\theta) | X, \hat{\theta}] + \log g(\theta|\Phi)$
Maximization: $\hat{\theta} = \arg \max_{\theta} R(\theta, \hat{\theta})$

To balance the relevance of the prior information and of the training data, we use the exponential forgetting mechanism [2] in the estimation step, introducing a coefficient $\rho \in [0,1]$ – the forgetting factor (FF) – as a weight associated to the prior density:

Expectation: $R(\theta, \hat{\theta}) = E[\log f(X|\theta) | X, \hat{\theta}] + \rho \log g(\theta | \Phi)$

When $\rho=1$ we get the usual MAP estimation, without any forgetting; when $\rho=0$ the prior knowledge is ignored and the definition of $R(\theta, \hat{\theta})$ become a classical Maximum Likelihood (ML) estimation. The forgetting mechanism is fundamental to achieve good adaptation performance even with a small field training corpus.

2.3 MAP solution

With the model-based a-priori parameter estimation and using the EM algorithm with the forgetting mechanism, we obtain the following recursive equations, solution of the segmental MAP estimation:

$$\tilde{\omega}_k(\rho) = \frac{\rho \tilde{\omega}_k^L + \tilde{\omega}_k^F}{\rho \tilde{\omega}_k^L + \tilde{\omega}_k^F}$$

$$\tilde{\mu}_k(\rho) = (1 - \lambda_k) \mu_k^L + \lambda_k \mu_k^F$$

$$\tilde{\Sigma}_k(\rho) = (1 - \lambda_k) \Sigma_k^L + \lambda_k \Sigma_k^F + \lambda_k (1 - \lambda_k)(\mu_k^L - \mu_k^F)^2$$

where:

$$\gamma_a = \sum_a \tilde{\omega}_k N(x_t | \tilde{\mu}_k, \tilde{\Sigma}_k) ; \quad \xi_k = \sum_a \gamma_a ; \quad \lambda_k = \frac{\xi_k^F}{\xi_k^L + \rho \xi_k^F}$$

$$\mu_k^L = \frac{\sum_a \gamma_a x_t}{\xi_k^L} ; \quad \Sigma_k^L = \frac{\sum_a \gamma_a (x_t - \mu_k^L)(x_t - \mu_k^L)^T}{\xi_k^L}$$

The superscripts $L$ and $F$ refer respectively to Laboratory and Field data while the subscript $k$ is referred to the $k$-th gaussian of given state mixture.

A segmentation step, using the current model parameters $\tilde{\theta} = (\tilde{\omega}, \tilde{\mu}, \tilde{\Sigma})$, is performed before re-estimation (3). Then, the current model is updated with the re-estimated values $\hat{\theta} = (\tilde{\omega}, \tilde{\mu}, \tilde{\Sigma})$ and a new EM step may be iterated.

3. FORGETTING FACTOR OPTIMIZATION

As remarked before, a good choice for the value of the forgetting factor (FF) is a key point for the success of the incremental training step. In fact, the FF allows to balance the relevance of the prior information compared to the adaptation data.

The FF depends on many factors such as the size of the incremental training data set, the generality of such data, the “distance” between the a-priori models and the application environment, the accuracy of the laboratory models. For instance the more field data are available, the more the estimation is reliable. At the same time, less relevance has to be given to the a-priori models ($\rho \rightarrow 0$). On the other hand, if the field data are scarce or not much representative, the weight of the a-priori models is essential to keep an appropriate accuracy of the resulting models ($\rho \rightarrow 1$). It is thus clear the importance to have an optimization criterion for the FF estimation.

The optimization criteria considered in this work belong to two classes:

- heuristics criteria;
- criteria based on estimation and evaluation sets.

The first class includes methods that exploit heuristic functions to set the FF, from the number of available speech samples.

![Figure 1. Forgetting factor optimization with estimation and evaluation sets](image)

The second class includes techniques that estimate the FF using an objective criterion. The algorithm is able to evaluate the usefulness of the field data, choosing the optimal weighting. Figure 1 exemplifies the approach: the field data are split into two sets, the estimation set and the evaluation set. The estimation set is used to produce the field models that are combined with the laboratory models exploiting MAP equations (3) with an initial FF value. The resulting models are then assessed on the evaluation set using the objective function. The procedure can be iterated until getting the optimization of the objective function with respect to the only variable parameter FF. In this work we use a ML objective function.

3.1 ML estimation

The maximum likelihood estimation of the FF is performed using the estimation/evaluation method described before. The objective function is replaced by a maximum likelihood estimator, tested on the evaluation set. We applied this technique obtaining the State dependent FF estimation (SFFE) and the Global FF estimation (GFFE).

To maximize the likelihood, we compute the derivative of the logarithm of (2) with respect to FF:
\[
\frac{\partial \ln f(X | \theta(\rho))}{\partial \rho} = \sum_{l=25}^{50} \sum_{k=1}^{10} \frac{1}{b(x_i)} \frac{\partial b(x_i)}{\partial \rho}
\]

where
\[
b(x_i) = \omega_x(\rho)N(x_i | \mu_x(\rho), \Sigma_x(\rho))
\]
\[
b(x_i) = \sum_{k=1}^{K} b_k(x_i)
\]

Since the FF is limited to the interval [0,1], we have to deal with a constrained maximization problem. Moreover, the form of the derivative function (5) is:
\[
\frac{\partial \Phi(\alpha)}{\partial \alpha} = \sum_i \varphi_i(\alpha)
\]

To solve that maximization problem we adopt an iterative algorithm: if \( \alpha \) is limited to the interval \([\alpha_{\text{min}}, \alpha_{\text{max}}]\) a re-estimation formula for the parameter \( \alpha \) is:
\[
\alpha^{n+1} = \alpha^n + \sum_i w_i \cdot \Delta \alpha_i
\]

where
\[
\Delta \alpha_i = \begin{cases} 
\alpha_{\text{max}} - \alpha^n & \text{if } \varphi_i(\alpha) > 0 \\
\alpha_{\text{min}} - \alpha^n & \text{if } \varphi_i(\alpha) \leq 0 \end{cases} \\
\Delta \alpha_i \\

w_i = \frac{\varphi_i(\alpha)}{\Delta \alpha_i}
\]

Equation (7) can be viewed as a weighted sum of contribution \( \Delta \alpha_i \), tied to the sign of the derivative component \( \varphi_i(\alpha) \): if the sign is positive, the contribution moves the current parameter value \( \alpha^0 \) towards its upper bound \( \alpha_{\text{max}} \); otherwise the parameter is moved towards the lower bound value \( \alpha_{\text{min}} \). The weights \( w_i \) are set to the normalized positive value of the derivative function, as in equation (8).

The recursive equation (7) can be applied to the State dependent FF estimation (SFFE), extending the summation on the index \( i \) to the indexes \( t \) and \( k \) of the equation (5). The global FF (GFFE) can be obtained by further expanding the summation to the whole set of model states.

### 3.2 Heuristic estimation

The heuristic estimation is carried out defining a combination function \( \lambda(\xi) \) for the MAP estimation. This function takes values in the interval [0,1]. The generalized state-dependent forgetting factor for the heuristic estimation can be obtained as:
\[
\rho = \frac{\xi^F}{\xi^F - (1 - \lambda(\xi^F))} \left( 1 - \frac{\lambda(\xi^F)}{\lambda(\xi^L)} \right)
\]

With these assumptions the reestimation equations for mean vector and covariance matrix become:
\[
\tilde{\mu}_x(\rho) = (1 - \lambda)\mu^F_x + \lambda \mu^L_x
\]
\[
\tilde{\Sigma}_x(\rho) = (1 - \lambda)\Sigma^F_x + \lambda \Sigma^L_x + \lambda(1 - \lambda)(\mu^L_x - \mu^F_x)^2
\]

1. \( \xi^F = \sum_i \xi^F_i \) and \( \xi^L = \sum_i \xi^L_i \) represent the total number of frame assigned to a given state for field and laboratory data respectively.

We used two different definition for the combination function, referred as Lambda I and Lambda II:
\[
\lambda_1(\xi) = 1 - e^{-\frac{\xi}{\xi + 1}}
\]
\[
\lambda_2(\xi) = \frac{\xi}{\xi + 1}
\]

Both definitions (11) are monotonically increasing parametric functions: the first one is deduced from empiric considerations, while the other one exploits the combination weights produced by the ML estimation methods.

![Figure 2. Heuristic functions Lambda I and Lambda II](image)

### 4. TASK, CORPUS AND SYSTEM

For this evaluation, we used the CSELT HMM based speech recognizer for flexible vocabulary. A set of 391 acoustic-phonetic units is modeled by Continuous Density HMMs (CDHMMs); each state of a unit is characterized by a maximum of 32 Gaussian densities [4]. The recognizer is a module of a vocal access train timetable information system, located into several call-centers of the Italian railway provider “Ferrovie dello Stato”.

The front-end consists of a Mel-based spectral analysis followed by a DCT, yielding a vector of 12 Cepstral features at each 10 ms frame. The first and second derivative of the log-energy and of the cepstral coefficients are also computed, providing an observation frame of 39 parameters.

The experiments were run on a speaker-independent isolated speech recognition task, for the Italian language, with a vocabulary of 664 city names. The initial laboratory models were trained on PSTN read speech, coming from more than 5000 speakers, while the adaptation data pools were collected through the train timetable information system. We used different subsets of the 23156 available tokens for adaptation and 6983 tokens as test set. The channel adaptation is performed by means of the global cepstral mean normalization.

### 5. PERFORMANCE EVALUATION

We tested the different FF optimization techniques varying the size of the adaptation data set. For this purpose we created several lists of tokens with logarithmic growing size. The field speech data of each list were then segmented, for the successive computations, using the startup laboratory model.
Table 1 shows the lower and upper bound performances obtained respectively with the laboratory and field models; the field model was trained using all the 23156 training tokens available.

<table>
<thead>
<tr>
<th></th>
<th>Error Rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory</td>
<td>8.3</td>
</tr>
<tr>
<td>Field (23156 tokens)</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Table 1. Reference recognition performances

Table 2 reports some figures about the performances of three standard training algorithms; the balanced MAP is obtained setting $\rho = \frac{\text{total field frames}}{\text{total laboratory frames}}$. The Forward-Backward and the Segmental K-Means require at least 1000 tokens to reach the laboratory lower bound while the balanced MAP is slightly better but however inadequate with 100 or 200 training tokens.

<table>
<thead>
<tr>
<th>Training tokens</th>
<th>FB</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43.9</td>
<td>26.3</td>
<td>12.7</td>
<td>6.7</td>
<td>4.3</td>
<td>3.4</td>
</tr>
<tr>
<td>SKM</td>
<td>31.4</td>
<td>23.5</td>
<td>11.2</td>
<td>7.2</td>
<td>5.9</td>
<td>3.8</td>
</tr>
<tr>
<td>Balanced MAP</td>
<td>12.1</td>
<td>9.5</td>
<td>7.4</td>
<td>5.0</td>
<td>4.4</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Table 2. Error rate of standard training algorithms: Forward backward (FB), Segmental K-Means (SKM) and Balanced MAP

Table 3 shows the error rate obtained with the maximum likelihood FF optimization methods. The adaptation was carried out using 2/3 of the available filed data as estimation set and the remaining 1/3 as evaluation set. The maximum number of iterations was set to 50. The figures are related to both global FF estimation (GFFE) and state dependent FF estimation (SFFE).

<table>
<thead>
<tr>
<th>Training tokens</th>
<th>GFFE</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.4</td>
<td>6.3</td>
<td>5.8</td>
<td>4.9</td>
<td>4.4</td>
<td>3.6</td>
</tr>
<tr>
<td>SFFE</td>
<td>6.8</td>
<td>6.0</td>
<td>5.4</td>
<td>4.8</td>
<td>4.6</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Table 3. Error rate with ML estimation of global (GFFE) and state dependent FF (SFFE)

The SFFE and GFFE recognition results are quite similar and both techniques can be successfully used for adaptation, without any need of tuning. A 23% of error rate reduction was obtained with only 100 adaptation tokens.

<table>
<thead>
<tr>
<th>Training tokens</th>
<th>Lambda I</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13.7</td>
<td>10.4</td>
<td>7.6</td>
<td>5.3</td>
<td>4.1</td>
<td>3.5</td>
</tr>
<tr>
<td>Lambda II</td>
<td>7.6</td>
<td>6.7</td>
<td>5.7</td>
<td>4.8</td>
<td>4.2</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Table 4. Error rate with heuristic FF estimation

The error rate achieved with the heuristic FF estimations is reported in Table 4. The $\lambda$ variable of both Lambda I and Lambda II parametric functions was set to 50, after some tuning tests.

The function Lambda II provides good recognition results in all the tested conditions while Lambda I is only effective with 1000 or more training tokens.

![Figure 3. Performances comparison](image)

Figure 3 compares the performances of the best FF optimization methods with the Forward-Backward training.

The experimental results show that the ML estimation techniques (GFFE and SFFE) are very effective with respect to other heuristic functions; moreover, no tuning is required.

6. CONCLUSION

A framework for the CDHMMs upgrade, using speech data coming from real applications, is proposed and evaluated. Several solutions to the problem of the optimal weighting between field data and laboratory models are proposed.

The results show that it is possible to improve the recognition performances using the MAP training with the forgetting mechanism. The optimal forgetting factor value can be obtained exploiting the adaptation data with a ML criterion, without any further need of tuning.

7. REFERENCES


