MINIMUM CROSS-ENTROPY ADAPTATION OF HIDDEN MARKOV MODELS

Mohamed Afify and Jean-Paul Haton

LORIA Universite Henri Poincare, Nancy I, B.P. 239 54506 Vandeouvre, France

ABSTRACT

Adaptation techniques that benefit from distribution correlation are important in practical situations having sparse adaptation data. The so called EMAP algorithm provides an optimal, though expensive, solution. In this article we start from EMAP, and propose an approximate optimisation criterion, based on maximising a set of local densities. We then obtain expressions for these local densities based on the principle of minimum cross-entropy (MCE). The solution to the MCE problem is obtained using an analogy with MAP estimation, and avoids the use of complex numerical procedures, thus resulting in a simple adaptation algorithm. The implementation of the proposed method for the adaptation of HMMs with mixture Gaussian densities is discussed, and its efficiency is evaluated on an alphabet recognition task.

1. INTRODUCTION

Model adaptation algorithms are currently a key component in practical speech recognition systems, to overcome the performance degradation resulting from the mismatched testing conditions that are often encountered in real world applications. Although MAP adaptation techniques (e.g. [2]) are theoretically optimal, their convergence is slow because they do not influence unobserved model parameters. A solution to this problem, which is widely employed in practice, is to estimate a set of transformations of the system parameters, as in Maximum likelihood linear regression (MLLR) [5], and to adjust the degree of sharing of these transformations according to the available amount of adaptation data, thus transforming all model parameters even for very limited adaptation speech. Yet, another way of approaching the above mentioned problem is to use parameter correlation to predict unobserved distributions from observed ones. Interestingly, these kinds of predictive techniques can be used separately or even in conjunction with transformation techniques as MLLR [6].

To this end, an optimal way of using correlation is to adopt a joint prior distribution of all the system parameters. This technique was developed in [4] for the means of Gaussian distributions, and referred to as extended MAP (EMAP) estimation. However, applying EMAP to a practical speech recognition system is very expensive in terms of computation and storage requirements, and some interesting approximations for feasible implementation can be found in [7, 9].

In this article we start from EMAP and we propose an approximate optimisation criterion based on iteratively maximising the posterior marginals. Then by making some assumptions we further simplify the posterior marginals into a set of local densities. By viewing the adaptation data as a set of constraints, we formulate the problem in a minimum-cross entropy (MCE) setting, where expressions for these local densities can be easily derived. The solution to the MCE problem is based on an analogy with conventional MAP adaptation, which alleviates the need of using expensive numerical procedures, and results in a very simple adaptation algorithm. Finally, we show the utility of the proposed method in adaptation scenarios having sparse data.

The paper is organised as follows. Section 2 gives the formulation and basic assumptions of the problem, that lead to iterative maximisation of a set of local densities. Expressions for these local densities are obtained by using the principle of MCE in Section 3. Section 4 shows the implementation of the proposed method for the adaptation of mean vectors of hidden Markov models (HMMs) with Gaussian mixture densities, followed by experimental evaluation and conclusion in Sections 5 and 6 respectively.

2. PROBLEM FORMULATION

Consider a system having N_g classes, and refer to the adaptation data as X. The extended MAP criterion for the correlated mean vector $\mu_1^{N_g} \equiv (\mu_1, ..., \mu_{N_g})$ can be written as

$$\hat{\mu}_1^{N_g} = \underset{\mu_1^{N_g}}{\operatorname{argmax}}\, p(\mu_1^{N_g}|X) \tag{1}$$

An approximate solution to (1) can be obtained by iteratively maximising the posterior marginals for each component mean as shown in (2). Compared to the exact criterion, (2) is computationally more efficient, and its use is inspired by coordinate descent methods in optimisation theory.

$$\hat{\mu}_k = \underset{\mu_k}{\operatorname{argmax}}\, p(\mu_k|\mu_l : l \neq k, X) \tag{2}$$

Further, for the distribution on the right hand side of (2) we make the following simplifying assumptions

1. $p(\mu_k|\mu_l : l \neq k, X) \approx p(\mu_k|\mu_l : l \neq k, X_k)$,where X_k is the subset of the adaptation data belonging to the distribution k.
2. $p(\mu_k|\mu_l : l \neq k, X_k) \approx p(\mu_k|\mu_l : l \in \mathcal{N}(k), X_k)$, where $\mathcal{N}(k)$ is the neighbourhood of k .We will show later how to find $\mathcal{N}(k)$.
3. $p(\mu_k|\mu_l : l \in \mathcal{N}(k), X_k) \approx p(\mu_k|f_{kl}(\mu_l, \theta_{kl}) : l \in \mathcal{N}(k), X_k)$, where $f_{kl}(\mu_l, \theta_{kl})$ is a parametric transformation function, and θ_{kl} is a parameter set to be specified later.

In the posterior marginals in (2) each component mean μ_k is assumed dependent on the whole adaptation data, and on all other means in the system. Assumption 1 restricts the data dependence to the subset of the adaptation data belonging to distribution k. While the dependence on other means is confined to the neighbourhood of distribution k using Assumption 2. Finally Assumption 3 uses a functional form of the correlation between two parameters,

and implicitly enforces a certain form of their joint distribution, e.g. a Gaussian distribution in the case of a linear function.

Thus, the optimisation criterion in (2) further simplifies to iteratively maximising the approximate posterior marginal in (3) for each component mean

$$\hat{\mu}_k = \underset{\mu_k}{\operatorname{argmax}}\, p(\mu_k | f_{kl}(\mu_l, \theta_{kl}) : l \in \mathcal{N}(k), X_k) \qquad (3)$$

where $1 \leq k \leq N_g$. In section 3 we will show how to obtain expressions of the density in (3) using MCE. For simplicity of discussion we will focus on scalar observations.

3. LOCAL DENSITY DERIVATION USING MCE

In this section, we show how to obtain expressions for the densities in (3) by applying the principle of MCE. We first consider the classical MAP case, i.e. no neighbourhood information is used, then we extend the result by introducing the effect of neighbourhoods.

3.1. Conventional MAP

Consider mean μ_k as a random variable having prior distribution $p(\mu_k)$ given by

$$p(\mu_k) = (2\pi\sigma_k^2)^{-1/2} \exp\left(-\frac{(\mu_k - \mu_k^I)^2}{2\sigma_k^2}\right) \qquad (4)$$

where μ_k^I is an initial estimate of μ_k and σ_k^2 the variance of μ_k. Now, consider that new information about μ_k is available through the adaptation data X_k. This information can be represented as the moment constraint on the posterior $q(\mu_k) \equiv p(\mu_k | X_k)$ as follows

$$\int (\mu_k - \bar{x}_k)^2 q(\mu_k)\, d\mu_k = \bar{\sigma}_k^2 \qquad (5)$$

where $\bar{x}_k$ is the sample average of n_k observations belonging to k, and $\bar{\sigma}^2$ will be specified below.
The posterior $q(\mu_k)$ having minimum cross-entropy ($H(q,p) \equiv \int q(\mu_k)\log(q(\mu_k)/p(\mu_k))\, d\mu_k$) with the prior $p(\mu_k)$, and satisfying the constraint in (5) is known to have the form (e.g. [8])

$$q(\mu_k) = p(\mu_k) \exp(-\lambda - \beta_k(\mu_k - \bar{x}_k)^2) \qquad (6)$$

where β_k, and λ are Lagrangian multipliers for the constraint (5), and the normalising constraint $\int q(\mu_k)\, d\mu_k = 1$ respectively.
After some simplification, we can rewrite (6) as

$$q(\mu_k) = (2\pi\sigma_k^{*2})^{-1/2} \exp(-\frac{(\mu_k - \mu_k^*)^2}{2\sigma_k^{*2}}) \qquad (7)$$

where

$$\mu_k^* = \frac{\beta_k \bar{x}_k + \mu_k^I/2\sigma_k^2}{\beta_k + 1/2\sigma_k^2} \qquad (8)$$

$$1/2\sigma_k^{*2} = \beta_k + 1/2\sigma_k^2 \qquad (9)$$

It is well known that the mean and variance of the posterior distribution in conventional MAP adaptation are given by

$$\mu_{k,map} = \frac{n_k \bar{x}_k/\sigma_{x_k}^2 + \mu_k^I/\sigma_k^2}{n_k/\sigma_{x_k}^2 + 1/\sigma_k^2} \qquad (10)$$

$$1/\sigma_{k,map}^2 = n_k/\sigma_{x_k}^2 + 1/\sigma_k^2 \qquad (11)$$

where $\sigma_{x_k}^2$ is the sample variance of observations of distribution k.

By comparing (8)-(9) to (10)- (11) we see that the two estimates coincide when $\beta_k = n_k/2\sigma_{x_k}^2$. In this case the value of $\bar{\sigma}_k^2$ in (5) will be given by

$$\bar{\sigma}_k^2 = \sigma_k^{*2} + (\mu_k^* - \bar{x}_k)^2 \qquad (12)$$

where μ_k^* and σ_k^{*2} are calculated from (8) and (9), and $\beta_k = n_k/2\sigma_{x_k}^2$.

Thus, the MCE estimate of $q(\mu_k)$ with the constraint (5), and $\bar{\sigma}_k^2$ given by (12) will coincide with the MAP estimate. Note that this result is obtained without making Gaussian assumption on the distribution of observations from k.

3.2. Adding neighbourhood information

The result in the previous section establishes a relationship between both MAP and MCE estimation, the important consequence is that we have a way for calculating the Lagrange multipliers using this analogy without resorting to complex numerical procedures. Here, we will build on this analogy to derive expressions for the local densities in (3).
Now in addition to (5) we consider the constraints imposed by the neighbourhoods of distribution k as:

$$\int (\mu_k - \mu_{k|l})^2 q(\mu_k)\, d\mu_k = \bar{\sigma}_{k|l}^2 \qquad \forall l \in \mathcal{N}(k) \qquad (13)$$

where $\mu_{k|l}$ is short hand for $f_{kl}(\mu_l, \theta_{kl})$, and the parameters $\bar{\sigma}_{k|l}^2$ will be specified below.
Using the same reasoning as in the previous subsection we can write the posterior $q(\mu_k) \equiv p(\mu_k | f_{kl}(\mu_l, \theta_{kl}) : l \in \mathcal{N}(k), X_k)$ as follows:

$$q(\mu_k) = p(\mu_k) \exp(-\lambda - \beta_k(\mu_k - \bar{x}_k)^2 - \sum_{l=1}^{|\mathcal{N}(k)|} \beta_{k|l}(\mu_k - \mu_{k|l})^2) \qquad (14)$$

where β_k, and λ are as the previous subsection, and $\beta_{k|l}$'s are Lagrangian multipliers for the constraints (13).
Inspired by the case of conventional MAP estimation, we take $\beta_k = n_k/2\sigma_{x_k}^2$, and analogously $\beta_{k|l} = 1/2\sigma_{k|l}^2$ (we will show how to calculate its value in the implementation section). After some simplifications, we arrive at

$$q(\mu_k) = (2\pi\sigma_k^{*2})^{-1/2} \exp\left(-\frac{(\mu_k - \mu_k^*)^2}{2\sigma_k^{*2}}\right) \qquad (15)$$

where

$$\mu_k^* = \frac{n_k \bar{x}_k/\sigma_{x_k}^2 + \mu_k^I/\sigma_k^2 + \sum_{l=1}^{|\mathcal{N}(k)|} \mu_{k|l}/\sigma_{k|l}^2}{n_k/\sigma_{x_k}^2 + 1/\sigma_k^2 + \sum_{l=1}^{|\mathcal{N}(k)|} 1/\sigma_{k|l}^2} \qquad (16)$$

$$1/\sigma_k^{*2} = n_k/\sigma_{x_k}^2 + 1/\sigma_k^2 + \sum_{l=1}^{|\mathcal{N}(k)|} 1/\sigma_{k|l}^2 \qquad (17)$$

Also as in the conventional MAP case, for this choice of Lagrange multipliers the values of $\bar{\sigma}_{k|l}^2$'s in (13) must satisfy

$$\bar{\sigma}_{k|l}^2 = \sigma_k^{*2} + (\mu_k^* - \mu_{k|l})^2 \quad \forall l \in \mathcal{N}(k) \qquad (18)$$

Using the density in (15) it is trivial to show that the maximisation in (3) is obtained for $\mu_k = \mu_k^*$, where μ_k^* is given by (16). It should be also noted that by removing the neighbourhood information in (16) it reduces to the conventional MAP case (8).

4. APPLICATION TO HIDDEN MARKOV MODEL ADAPTATION

Having developed an approximate maximisation criterion, and derived expressions for the resulting local densities, we are interested in applying the proposed technique to the adaptation of mean vectors of hidden Markov models with Gaussian mixture densities. In this case, we consider all state component distributions of all models as forming a large pool of size N_g. Moreover, when using diagonal covariance matrices the algorithm can be separately repeated for each vector dimension. We now, following [1], consider some choices for the practical implementation of the algorithm.

The neighbourhood $\mathcal{N}(k)$ of distribution k is taken from its mostly correlated distributions (i.e. those having highest values of $|r_{kl}|$). For each distribution l in this neighbourhood, the transformation function is taken to be linear as shown in (19)

$$\mu_{k|l} \equiv f_{kl}(\mu_l, \theta_{kl}) = a_{kl}\mu_l + b_{kl} \tag{19}$$

It can be shown that (see e.g. [1]) the optimum values of a_{kl} and b_{kl} in a MMSE sense are given by

$$a_{kl} = \frac{r_{kl}\sigma_k}{\sigma_l} \tag{20}$$

$$b_{kl} = \mu_k^I - a_{kl}\mu_l^I \tag{21}$$

where $\mu_{k(l)}^I$ is an initial estimate of mean $k(l)$, $\sigma_{k(l)}$ is the corresponding prior standard deviation, and r_{kl} is the correlation coefficient between k and l. In turn, estimates of the correlation coefficients and the variances can be easily obtained using the moment method from a training set consisting of N groups (e.g. speakers). In this case the variance of the estimate $\mu_{k|l}$ can be shown to be equal to

$$\sigma_{k|l}^2 = \sigma_k^2(1 - r_{kl}^2) + a_{kl}^2 Var(\mu_l) \tag{22}$$

where the first term accounts for the prediction error, and the second accounts for the fact that we use an estimate of μ_l not its true value, and $Var(\mu_l)$ can be calculated as in (17).

4.1. Summary of the adaptation algorithm

For completeness, we present a summary of both the training and adaptation phases of the proposed algorithm.

1. Training Phase
   - Start with N instantiations of each mean (e.g. from different speakers), and a set of initial models (e.g. speaker independent models).
   - Assign the training speech to corresponding distributions using the Viterbi algorithm and the initial models.
   - For each distribution estimate the variance, and the correlation with all other distributions using the method of moments. Starting from these values, construct the neighbourhood of each distribution, and calculate the values of its transformation parameters using equations (20) and (21).
2. Adaptation Phase
   - Assign the adaptation data to corresponding distributions using the Viterbi algorithm and the set of initial models.
   - Collect statistics (counts and sample averages) of all distributions.
   - Iterate until convergence
     - For each distribution, calculate the adapted mean using equation (16) (where $\mu_{k|l}$ is calculated using (19), and $\sigma_{k|l}^2$ is calculated using (22)).
     - For each distribution, calculate the variance using (17).

5. EXPERIMENTAL RESULTS

We used two databases to evaluate the performance of the proposed algorithm in an isolated alphabet (26 words) recognition task. The first is the OGI ISOLET alphabet database, which contains 2 repetitions of each letter from 150 speakers (75 male/75 female). The second is the alphabet subset of the TI46 database, which contains 26 repetitions (10 in one session and 16 in 8 sessions i.e. two per session) of each letter from 16 speakers (8 male/8 female). Both databases were down-sampled to 8 kHz. Speech is parametrised using 12 MFCC, and cepstral mean normalisation is carried out at the utterance level as a means of acoustic normalisation. Each letter is represented by 5 state left-to-right HMM with no skipping, and each state is represented by a 4-component Gaussian mixture with diagonal covariances. Feature extraction and initial model training were performed using HTK.

Initial speaker independent models are constructed using the ISOLET database, while TI46 is used for adaptation and testing. The first 10 repetitions from TI46 were used to estimate the neighbourhood structure and transformation functions parameters as outlined in the previous section. From the remaining 8 sessions the first utterance was used for test while the second was reserved for adaptation. Due to some missing files the total number of files used in testing is 3318. The speaker independent recognition rate is 53.5%, which indicates the severe mismatch between the two databases. Experimental results using the same databases but using an online adaptation scenario are reported in [3]

We performed two types of adaptation experiments. In the first one, we used adaptation utterances from all classes and varied the number of repetitions from each class from 1 to 8. We refer to these experiments as smoothing experiments, as in this case the neighbours can be viewed as performing a form of smoothing. In the second set of experiments, we used only 1 adaptation utterance from each class, and in addition we randomly sampled the classes to obtain 13, 8, 6 and 4 classes for use in adaptation. These experiments are referred to as prediction experiments, as neighbourhood information is doing a sort of prediction for the missing classes. The results for both types of experiments are shown in figures 1 and 2 respectively. In the case of using neighbourhood information, three iterations of the adaptation algorithm were used, while in the case of zero neighbourhood (note that this is conventional MAP) no iterations are required.

From the presented results it can be seen that the proposed method is beneficial compared to conventional MAP when all classes are present but having sparse adaptation data, and also when some of the classes are missing. When sufficient adaptation data exists for all classes the performance slightly degrades, maybe due to the increase of the smoothing effect. Also it can be observed that in the case of prediction experiments it is more desirable to use a larger size of the neighbourhood in contrast to the smoothing case where a smaller size of the neighbourhood seems more adequate.

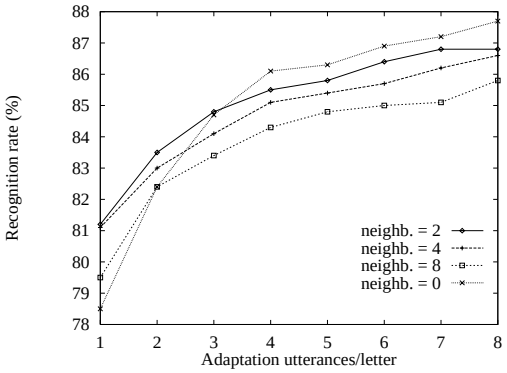

Figure 1: Results for smoothing experiments. In this case all classes are used for adaptation,and the number of tokens from each class is varied from 1 to 8. Speaker independent recognition rate is 53.5%

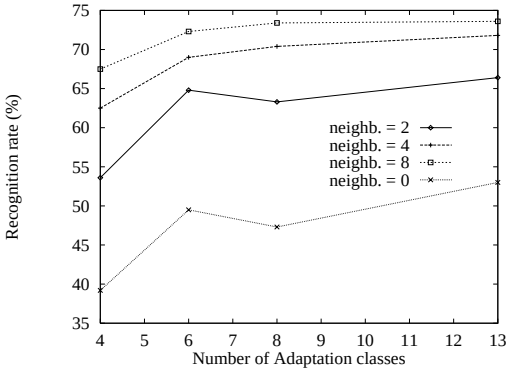

Figure 2: Results for prediction experiments. In this case we randomly sample the classes in order to use 13,8,6,and 4 classes during adaptation, and the number of tokens from each class is fixed to one. Speaker independent recognition rate is 53.5%

6. CONCLUSION

We have presented an adaptation algorithm which approximates the theoretically optimal EMAP algorithm, and requires far less computation. The basic idea is to use a simplified iterative optimisation procedure for a set of local densities. Expressions for these densities were obtained by applying the MCE principle, and using an analogy with classical MAP estimation for direct calculation of the Lagrange multipliers. The resulting algorithm is very simple, and when implemented in the context of HMMs with Gaussian mixture densities, it resulted in significant improvement over classical MAP adaptation for sparse adaptation data, and when some classes were missing during adaptation.

7. REFERENCES

[1] M.Afify, Y.Gong, J.-P.Haton, "Correlation based predictive adaptation of hidden Markov models," Proc. Eurospeech-97, pp.2059-2062, 1997.

[2] J.L.Gauvain and C.H.Lee,"Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, Apr. 1994.

[3] Q.Huo, and C.H.Lee,"On-line adaptive learning of the correlated continuous density hidden Markov model for speech recognition," Proc. ICSLP-96, 1996.

[4] M.Lasry, and R.Stern,"A posteriori estimation of correlated jointly Gaussian Mean vectors," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 6, No. 4, pp.530-535, July 1984.

[5] C.Leggetter, P.Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," Proc. ICSLP-94, pp. 451-454, 1994.

[6] S.C.Scott, P.DeSouza, "Speaker adaptation by correlation (ABC)," Proc. Eurospeech97, pp.2111-2114, 1997.

[7] B.M.Shahshahani,"A Markov random field approach to Bayesian speaker adaptation," Proc. IEEE ICASSP-96, pp.697-700, 1996.

[8] J.E.Shore, R.W.Jhonson, "Properties of cross-entropy minimisation," IEEE Trans. Information theory, Vol. IT-27, pp.472-482, July 1981.

[9] G. Zavaliagkos, R. Schwartz, and J. McDonough, "Maximum a posteriori adaptation for large scale HMM recognisers," Proc. IEEE ICASSP-96, pp.725-728, 1996.