

Gradient-based MAP for improved HMM adaptation/learning

Darryl W Purnell, Elizabeth C Botha

Department of Electrical, Electronic and Computer Engineering
University of Pretoria, 0002, South Africa.

botha@ee.up.ac.za

Abstract

In this paper, a gradient-based framework for maximum *a posteriori* (MAP) estimation of hidden Markov models (HMM) is presented. The two new algorithm makes no assumption about the form of the prior distribution used in MAP estimation. The algorithms are tested using the South African SUNSpeech database. Improvements of up to 10.2% in relative error rate (over ML training) on the test set is achieved.

1. Introduction

Traditionally, maximum likelihood (ML) estimation has been used to estimate hidden Markov model (HMM) parameters. However, the ML procedure gives inaccurate estimates for sparse training data scenarios. Maximum *a posteriori* (MAP) estimation of continuous density hidden Markov models has gained much attention for such sparse data problems. MAP estimation has been used with much success in various applications, such as parameter smoothing, speaker adaptation, dialect adaptation and cross-language adaptation.

Lee *et al.* [1] introduced a MAP algorithm, where the parameters of multivariate Gaussian state observation densities of HMM models were adapted for speaker adaptation. Gauvain and Lee [2] extended the MAP formulation for HMMs to handle parameters of *mixtures* of Gaussian densities. Gauvain and Lee [3] later presented a theoretical framework for MAP estimation for HMMs with Gaussian mixture state densities, where they proposed using an expectation-maximization (EM) approach to finding the MAP estimate. In their work, the MAP formulation was also extended to include the estimation of the transition probabilities and initial state probabilities.

In this work, we are primarily interested in the usage of the MAP estimation algorithm for adaptation purposes as opposed to applications such as parameter smoothing and corrective training. Model adaptation is a process for adjusting seed models to create more specialized models using a small amount of adaptation data.

The standard *maximum a-posteriori* (MAP) adaptation procedure of Gauvain and Lee [3] makes assumptions about the form of the prior probability distribution used. This can be a problem when the prior is relatively complex. A new *gradient-based* MAP adaptation algorithm is therefore proposed. This adaptation algorithm makes no assumptions about the form of the prior distribution used.

We use the the South African SUNSpeech phonetic database [4] (in a language adaptation experiment) to evaluate the gradient-based MAP algorithms and the MAP algorithm of Gauvain and Lee [3]. We have attempted to keep our notation as close to that of Rabiner and Juang [5] as possible.

The organization of this paper is as follows. Section 2 summarizes the relevant Bayesian theory. The choice of prior distribution

is discussed in Section 3. A new *gradient-based* method of obtaining the MAP estimate is described in Section 4. The two methods are experimentally compared in Section 5. Finally, we summarize results and conclude in Section 6.

2. Bayesian theory

This section summarizes the pertinent Bayesian theory used in this article. For a more complete introductory text on Bayesian statistics, the reader is referred to Box and Tiao [6] and DeGroot [7]. The theory and discussions in this section will be biased towards speech recognition applications of Bayesian adaptation.

2.1. Bayes' theorem

Given a vector $\mathbf{y} = (y_1, \dots, y_n)$ of n observations, with probability distribution $P(\mathbf{y}|\theta)$, which depends on the k parameters $\theta^T = (\theta_1, \dots, \theta_k)$ with probability distribution $P(\theta)$, then given the observed data \mathbf{y} , the conditional distribution of θ is

$$P(\theta|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)P(\theta)}{P(\mathbf{y})}. \quad (1)$$

The denominator in Eq. (1), $P(\mathbf{y})$, is a normalizing factor, which ensures that the integral of $P(\theta|\mathbf{y})$ is equal to one.

Equation (1) is referred to as Bayes' theorem. The distribution $P(\theta)$, is called the *prior* distribution and expresses what is known about the model parameters before any data is observed. The *posterior* distribution $P(\theta|\mathbf{y})$, tells us what is known about the model parameters, given that data has been observed. The distribution $P(\mathbf{y}|\theta)$ is often referred to as the data *likelihood* and can be written $L(\theta|\mathbf{y})$.

2.2. Sequential nature of Bayes' theorem

If we assume that the observations are independent, then we can write Bayes' theorem as follows

$$P(\theta|\mathbf{y}) \propto P(\theta) \prod_{i=1}^n L(\theta|\mathbf{y}^{(i)}) \quad (2)$$

$$= \left[P(\theta) \prod_{i=1}^k L(\theta|\mathbf{y}^{(i)}) \right] \prod_{i=k+1}^n L(\theta|\mathbf{y}^{(i)}) \quad (3)$$

$$= P(\theta|\mathbf{y}_1, \dots, \mathbf{y}_k) \prod_{i=k+1}^n L(\theta|\mathbf{y}^{(i)}) \quad (k < n). \quad (4)$$

Equation (4) is exactly the same as Eq. (2), except that $P(\theta|\mathbf{y}_1, \dots, \mathbf{y}_k)$, the posterior distribution of θ given $\mathbf{y}_1, \dots, \mathbf{y}_k$, now acts as the prior distribution for

y_{k+1}, \dots, y_n . Bayes' theorem, therefore describes the process of learning as data becomes available. We can therefore, as Eq. (4) suggests, compute the posterior for a given set of data and then use that posterior as a "prior" when more data becomes available. This result is of utmost importance to the MAP algorithm proposed in this article.

2.3. Bayesian learning and prediction

The result of Bayesian learning is a probability distribution (posterior) which expresses our beliefs of how likely individual parameter values are. In a Bayesian approach to HMM parameter estimation and speech recognition, the objective is to find a predictive distribution for an unknown utterance, given the utterance observations, as well as the training observations. Let the observation sequence for the i th example be written as \mathbf{O}_i . For n training examples $\mathbf{O} = (\mathbf{O}_1, \dots, \mathbf{O}_n)$, when we wish to classify an unknown observation, we choose the class which maximizes the following probability:

$$P(\mathbf{O}_{unknown} | \mathbf{O}_1^{(i)}, \dots, \mathbf{O}_n^{(i)}) \\ = \int P(\mathbf{O}_{unknown} | \theta) P(\theta | \mathbf{O}_1^{(i)}, \dots, \mathbf{O}_n^{(i)}) d\theta, \quad (5)$$

where i is the class and $\mathbf{O}_{unknown}$ is the unknown observation sequence.

2.4. Maximum *a-posteriori* probability estimate

Assuming the posterior is sufficiently peaked around the most probable point (θ_{MAP}), we can approximate Eq. (5) as

$$P(\mathbf{O}_{unknown} | \mathbf{O}_1, \dots, \mathbf{O}_n) \approx P(\mathbf{O}_{unknown} | \theta_{MAP}), \quad (6)$$

where θ_{MAP} is the Maximum *a-posteriori* (MAP) estimate of the parameter vector θ .

The MAP point is the set of parameters that maximize Eq. (1), or

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta | \mathbf{O}_1, \dots, \mathbf{O}_n) \\ = \underset{\theta}{\operatorname{argmax}} \left[P(\mathbf{O}_1, \dots, \mathbf{O}_n | \theta) P(\theta) \right]. \quad (7)$$

If we had no prior knowledge about θ , then we would choose a non-informative (improper) prior to be used in Eq. (7), i.e., $P(\theta) = \text{constant}$. Equation (7) then reduces to the normal maximum likelihood (ML) formulation.

3. Prior distributions

In this section, the choice of the prior density family is discussed. The prior distribution is an important part of any Bayesian method, as it expresses our knowledge about the distributions prior to any data being observed. It is especially important when there is little data available.

In the context of HMM's, prior distributions must be chosen for the following parameters:

- transition probabilities a_{ij} ,
- mixture weights C_{jk} ,
- Gaussian means μ_{jk} , and
- Gaussian variances Σ_{jk} .

Conjugate prior distributions have been chosen in this work, not only because of the convenient relationship between prior and posterior, but also due to their usage in the literature.

The prior distribution for all the parameters of the Gaussian mixtures of the HMMs is chosen to be a normal-Wishart distribution, where the conditional of μ (mean) and R (precision matrix, with $R = \Sigma^{-1}$) are given as follows: The conditional distribution of μ (when $R = r$), is a multivariate normal distribution with mean vector m and precision matrix νr for $m \in \mathbb{R}^D$ and $\nu > 0$. The marginal distribution of R is a Wishart distribution with α degrees of freedom and precision matrix τ . The priors for the transition probabilities and Gaussian mixture weights are chosen to be Dirichlet distributions.

4. Gradient based MAP

The MAP estimation method proposed by Gauvain and Lee [3], assumes that the prior used is of a specific form. This is potentially a limiting factor in the performance of that MAP algorithm. In this section a gradient based MAP estimation algorithm is developed which does not make assumptions about the form of the prior distribution. This algorithm will, so as to prevent confusion, be referred to as the GMAP algorithm.

The above statement is not entirely true, as a prior of fixed form is used. It is, however, a non-informative prior which is used in the calculation of the new prior, which in turn is then used in the adaptation process. Though, if true prior knowledge about the model or system is available, it can be expressed through this parametric prior.

In the proposed algorithm, the prior will not be estimated at all, but will be implicitly included in the update procedure. It will, however, be far more computationally expensive than the regular MAP algorithm. We believe that improved performance will offset the extra computational difficulties for certain tasks. This adaptation algorithm will probably not find a place in rapid adaptation needed in some speaker adaptation tasks. It should, however, be more than useful in tasks such as cross-language adaptation and some speaker adaptation tasks where training time is not critical.

Empirical Bayesian methods [8, 9] or the simpler method of estimating the prior proposed by Gauvain and Lee [3] are typically used to estimate the parametric prior distribution used in the standard MAP approach [3]. Neither of these two methods results in a prior which becomes more peaked as the amount of observed data increases. This is a potential problem, as it does not account for the fact that some HMMs (or states or mixtures) will have been observed more often than others. However, the two methods could potentially help the algorithm when there is a reasonable mismatch between prior data and the task specific data.

The sequential nature of Bayes' theorem can be used to determine the MAP estimate by using the posterior of the prior data as the prior distribution (Eq. (4)). However, there will typically be more prior data than adaptation data, and the prior distribution will therefore tend to dominate in the calculation of the posterior using Eq. (4). Any reasonable mismatch between the prior data and task-specific data will also tend to be a problem. These problems can, however, be addressed by simply weighing the prior distribution (posterior of previously observed data) with a value which is a function of the mismatch (*a-priori* knowledge/belief) and the amount of prior and adaptation data.

Referring to Eq. (3), the first part in square brackets is the posterior of a set of data ($1 \dots k$) and is used as the prior for the

remainder of the data. The posterior of this reference or “prior” set, used as the prior in the adaptation framework will tend to dominate Eq. (3) when there are more training examples than adaptation examples (i.e. when $k \gg n - k$). MAP adaptation is typically used in situations where this will occur. It therefore makes sense to weigh the “prior” in some way so as to ensure that it does not dominate. In our implementation, the weighting is done as follows

$$P(\theta|\mathbf{O}) \propto \left[P(\theta) \prod_{i=1}^k L(\theta|\mathbf{O}_i) \right]^\kappa \prod_{i=k+1}^n L(\theta|\mathbf{O}_i) \quad (k < n). \quad (8)$$

The value κ has the effect of flattening and widening the prior when $0 \leq \kappa < 1$ and making it more peaked around the mode when $\kappa > 1$. Figure 1 presents an example of a Normal distribution with a mean of zero and standard deviation of two, which has been raised to the value of $\kappa = 0.2$ and $\kappa = 2$.

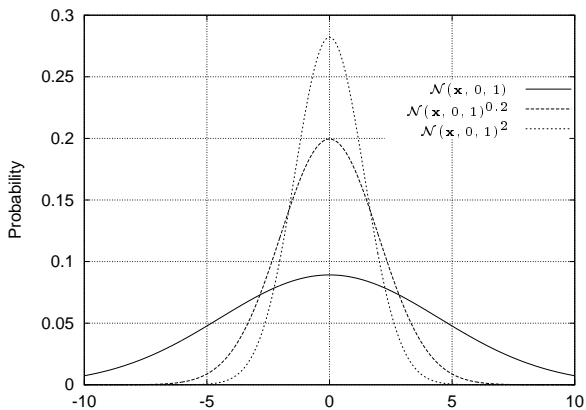


Figure 1: Example of the effects of raising a distribution to a power (assuming it is normalized)

It is convenient at this point to express the posterior in terms of an energy function, which will be optimized to find the MAP estimate. We can write the posterior probability function as follows

$$P(\theta|\mathbf{O}) \propto e^{-E(\theta;\mathbf{O})} \quad (9)$$

where $E(\theta; \mathbf{O})$ is the “potential energy” function. Any probability function can be written in this way by defining $E(\theta; \mathbf{O}) = -\log P(\theta|\mathbf{O}) - \log(Q)$ for a given constant Q . We choose $Q = 1$.

Writing the posterior in Eq. (8) in the above form, and assuming independence of the observations, we get

$$E(\theta; \mathbf{O}) = \kappa \left[-\log[P(\theta)] - \sum_{i=1}^k \log[P(\mathbf{O}_i|\theta)] \right] - \sum_{i=k+1}^n \log[P(\mathbf{O}_i|\theta)], \quad (10)$$

where examples $1, \dots, k$ are the prior (reference) set and examples $(k+1), \dots, n$ are the adaptation set.

The steepest descent algorithm can now be used to iteratively estimate the MAP point on the posterior defined in Eq. (10), i.e.

$$\theta^{(i)}(n+1) = \theta^{(i)}(n) - \epsilon \frac{\partial E(\theta)}{\partial \theta^{(i)}}. \quad (11)$$

The steepest descent algorithm (Eq. (11)) is an unconstrained optimization technique and given that certain constraints must be maintained for HMMs, some modifications are required.

4.1. Parameter adaptation

Instead of using a complicated constrained steepest descent algorithm, we use transformations to maintain the HMM constraints during parameter adaptation. The derivation of the updates for the individual parameters is beyond the scope of this article and can be found in [10].

4.2. Initial prior

Although the GMAP algorithm has been developed such that an initial parametric prior can be used (to incorporate true *a-priori* information), it is unlikely that any such information will exist. This algorithm will therefore often be used with a non-informative initial prior. This results in the parametric prior gradient being zero, with no effect on the resultant update.

5. Experiments

The goal of this section is to experimentally compare the MAP and GMAP algorithms in conditions where limited training data is available, with a reasonable amount of non-task-specific data available for adaptation purposes. The SUNSpeech dataset is used.

The phoneme recognition accuracy of the system is reported, where accuracy is defined as:

$$Accuracy = \frac{Phones - Subs - Dels - Ins}{Phones}, \quad (12)$$

where *Phones* is the number of phones in the correct transcription, *Subs* the number of substitutions, *Dels* the number of deletions and *Ins* the number of insertions. Error rates reported are defined as $100\% - Accuracy$. The speech signal is blocked into frames of length 16ms, with overlap of 6ms. Thirteen Mel-frequency cepstral coefficients (MFCCs), along with their first and second order differentials are used. Unless otherwise stated, each phone is represented by a simple left-to-right, 3 state, 5 mixture HMM. A frame-synchronous modified Viterbi search (trellis search) is used to automatically segment and label the utterances.

Before continuing, it is necessary to explain the convention we have used to describe (or label) the algorithms used:

- MAP - A MAP algorithm labelled as “MAP $T_X (T_Y)$ ” uses the dataset T_X as its adaptation set and the ML model created using T_Y to determine the prior distribution.
- GMAP - A GMAP algorithm labelled as “GMAP $T_X (T_Y)$ ” uses the dataset T_X as its adaptation set and the ML model created using T_Y as a starting point. Note that the prior dataset is not included in the description as it is a constant for the each experiment.

The number of iterations used for each procedure differed. When using MAP, 10 iterations typically proved to be sufficient with the testing set performance converging at or before 10 iterations. GMAP required between 10 and 30 iterations for the testing set performance to converge. The number of iterations required for the GMAP algorithm is dependent on the step size ϵ , the size of the datasets and the weighting factor κ .

5.1. SUNSpeech

This section compares the MAP and GMAP adaptation algorithms within a *language adaptation* framework. The SUNSpeech database [4] was compiled by the Department of Electrical and Electronic Engineering of the University of Stellenbosch (South Africa) to contain phonetically labelled speech in both English and Afrikaans. A total of 59 phonetic categories, including both a *silence* and *unknown* category, were used to segment both the Afrikaans and the English speech. In this work, the Afrikaans subset of the SUNSpeech database has been used as adaptation data, with the English subset being used as the prior dataset. Table 1 summarizes the SUNSpeech training and testing sets as used.

Table 1: Details of SUNSpeech training and testing sets used

Description	Label	Speakers	Duration (minutes)
English dataset			
Training	E	76	135.5
Afrikaans dataset			
Training set	A	39	22
Training subset	A_S	8	5.5
SI test set		15	13

Table 2 summarizes the results obtained using the different algorithms for the SUNSpeech dataset. Considerable improvements are realized from the usage of the two algorithms when the small Afrikaans training set is used as adaptation data, with the GMAP algorithm performing best under these conditions (8.0% and 10.2% relative improvement in error rate for the 5 and 10 mixture models respectively). Smaller, but finite, improvements are attained when the full Afrikaans training set is available, where the GMAP algorithm once again performed better than the standard MAP algorithm.

Table 2: Summary of the best results obtained for the SUNSpeech dataset. Relative improvement in error rate over ML baseline is given in braces.

Description	Mixtures	
	5	10
Baseline ML A_S	42.5 (0.0%)	41.2 (0.0%)
MAP $A_S(A_S + E)$	45.4 (5.0%)	44.8 (6.1%)
GMAP $A_S(MAP)$	47.1 (8.0%)	47.2 (10.2%)
Baseline ML A	48.6 (0.0%)	51.5 (0.0%)
MAP $A(A + E)$	51.1 (4.9%)	52.4 (1.9%)
GMAP $A(MAP)$	51.3 (5.3%)	52.5 (2.1%)

The baseline recognition results appear relatively low, as compared to that obtained for continuous recognition on other

continuous phoneme recognition tasks, such as TIMIT. This is due to the fact that there are 59 phonetic categories and is therefore a more difficult task.

6. Conclusion

This article extends and elaborates on Bayesian adaptation (MAP) and its usage within a continuous speech recognition framework. A new gradient-based MAP algorithm which makes no assumption about the form of the prior was proposed and the implementation thereof was discussed.

The two algorithms (MAP and GMAP) were experimentally evaluated in Section 5, using the South African SUNSpeech database for language adaptation. The GMAP algorithm performed best, with the MAP algorithm, in general, resulted in worse performance than the GMAP algorithm.

7. Acknowledgments

The authors would like to thank the Mellon Foundation and the National Research Foundation of South Africa for their support of this work. The authors would also like to thank the University of Stellenbosch for the use of their SUNSpeech database.

8. References

- [1] C-H. Lee, C-H Lin, and B-H Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 806–814, Apr. 1991.
- [2] J.-L. Gauvain and C.H. Lee, "Bayesian learning for hidden Markov models with Gaussian mixture state observation densities," *Speech Communication*, pp. 205–213, June 1992.
- [3] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
- [4] T. Waardenburg, J.A. du Preez, and M.W. Coetzer, "The automatic recognition of stop consonants using hidden Markov models," in *Proc. ICASSP '92*, San Francisco, CA, Mar. 1992.
- [5] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- [6] G.E.P. Box and G.C. Tiao, *Bayesian Inference in Statistical Analysis*, John Wiley and Sons, New York, 1973,1992.
- [7] M.H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.
- [8] J. S. Maritz and T. Lwin, *Empirical Bayes Methods*, Chapman and Hall, London, 1989.
- [9] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Annals of Mathematical Statistics*, vol. 35, pp. 1–20, 1964.
- [10] D.W. Purnell and E.C. Botha, "Gradient-based MAP and MAPMCE for improved HMM adaptation/learning," *IEEE Transactions on Speech and Audio Processing*, submitted.