

Discriminative Estimation of Subspace Precision and Mean (SPAM) Models

Vaibhava Goel, Scott Axelrod, Ramesh Gopinath, Peder Olsen, Karthik Visweswariah

IBM T. J. Watson Research Center,
Yorktown Heights, NY 10598

{vgoel, axelrod, gopinath, pederao, kv1}@us.ibm.com

Abstract

The SPAM model was recently proposed as a very general method for modeling Gaussians with constrained means and covariances. It has been shown to yield significant error rate improvements over other methods of constraining covariances such as diagonal, semi-tied covariances, and extended maximum likelihood linear transformations. In this paper we address the problem of discriminative estimation of SPAM model parameters, in an attempt to further improve its performance. We present discriminative estimation under two criteria: maximum mutual information (MMI) and an “error-weighted” training. We show that both these methods individually result in over 20% relative reduction in word error rate on a digit task over maximum likelihood (ML) estimated SPAM model parameters. We also show that a gain of as much as 28% relative can be achieved by combining these two discriminative estimation techniques. The techniques developed in this paper also apply directly to an extension of SPAM called subspace constrained exponential models.

1. Introduction

In state-of-the-art speech recognition systems hidden Markov models (HMMs) are used to estimate the likelihood of an acoustic observation given a word sequence. The probability density function associated with states of these HMMs are typically chosen to be mixtures of Gaussians,

$$P(x|s) = \sum_{g \in \mathcal{M}(s)} \pi_g P(x|g) = \sum_{g \in \mathcal{M}(s)} \pi_g \mathcal{N}(x; \mu_g, P_g), \quad (1)$$

where x is a d -dimensional real-valued feature vector, $\mathcal{M}(s)$ is the set of Gaussians modeling state s , μ_g is the Gaussian mean, and P_g is the Gaussian inverse covariance which is also known as the precision matrix.

There has recently been much work on modeling the Gaussian means and precision matrices which in general belong to \mathbf{R}^d and the space of symmetric positive definite $d \times d$ matrices, respectively. In practice constraints are needed, especially on covariances, to allow for robust estimation, efficient storage, and efficient computations. As mentioned above, the most common constraint is to restrict Σ to the space of diagonal positive definite matrices. Other recently proposed methods constrain inverse covariances or precision matrices to affine subspaces; these include, in the order of least to most general, semi-tied covariances [3] or maximum likelihood linear transformation (MLLT) [4], extended MLLT (EMLLT) [5], and subspace precision and mean modeling (SPAM) [6]. In a recent comparison of these techniques [6], the SPAM model was shown to significantly outperform other models of comparable complexity.

In this paper we address the problem of discriminative estimation of SPAM model parameters. We show that the MMI

estimation procedure [7, 1, 2] can be extended to the SPAM model. It is implemented with the extended Baum-Welch (EBW) method [7], using the derivation for continuous parameter exponential models proposed by Gunawardana et.al. [8].

Furthermore, we compare MMI with another novel discriminative estimation technique which we term *error-weighted* training. This method is similar in spirit to Boosting [9] and also to corrective training [10]; it places higher emphasis on the subset of training data that is recognized incorrectly and does maximum likelihood training with a higher weight placed on the statistics gathered over the sentences with errors.

The rest of this paper is organized as follows. In Section 2 we review the SPAM model and maximum likelihood estimation of its parameters. In Sections 3 and 4 we describe the MMI and error-weighted parameter estimation procedures as applied to the SPAM model. In Section 5 we report on experiments on a digit-string test set.

2. The SPAM Model and ML Estimation of Parameters

The SPAM model can be described by first writing the Gaussian density function as an exponential density

$$\begin{aligned} P(x|g) &= \mathcal{N}(\mu_g, P_g (= \Sigma_g^{-1})) \\ &= \exp\left(\theta_g^T f(x) + K(\theta_g)\right), \end{aligned} \quad (2)$$

where

$$\begin{aligned} \theta_g &= [\text{vec}(P_g); P_g \mu_g] \\ f(x) &= [\text{vec}(xx^T); x] \\ K(\theta_g) &= -0.5 \mu_g^T P_g \mu_g - \frac{d}{2} \log(2\pi) \\ &\quad + 0.5 \log \det(P_g) \end{aligned}$$

Here $\text{vec}(A)$ denotes a vector constructed from elements of matrix A , with the property that the trace of AB is equal to the inner product of $\text{vec}(A)$ and $\text{vec}(B)$. For a SPAM model, the precision matrices P_g and the linear parameters $P_g \mu_g$ are required to lie in Gaussian independent subspaces,

$$\begin{aligned} P_g &= S_0 + \sum_{k=1}^D \lambda_g^k S_k \\ P_g \mu_g &= M_0 + \sum_{l=1}^L \eta_g^l M_l. \end{aligned} \quad (3)$$

Here $S_0, \{S_k\}$ are real symmetric $d \times d$ matrices and $M_0, \{M_l\}$ are d dimensional vectors. The parameters $S_0, S_k, M_0,$ and M_l

are shared across Gaussians; these are referred to as *tied* model parameters. S_0 and S_k The Gaussian specific parameters, λ_g^k and η_g^l , are referred to as *un-tied* model parameters. The Gaussian priors π_g also form part of the un-tied SPAM model parameters.

Maximum likelihood estimation of SPAM model parameters is carried out using the well known expectation-maximization (EM) algorithm. Given labeled training data (X^*, W^*) and a starting value $\Theta^0 = \{\theta_g^0\}$ of the model parameters, the EM auxiliary function is

$$Q_{em}(\Theta|\Theta^0) = \sum_g \theta_g^T \tilde{f}_g + \tilde{\pi}_g K(\theta_g) + \tilde{\pi}_g \log \pi_g, \quad (4)$$

where,

$$\tilde{f}_g = \sum_t \gamma(t, g) f(x_t)$$

$$\tilde{\pi}_g = \sum_t \gamma(t, g)$$

$$\gamma(t, g) = P(g(t) = g | X^*, W^*; \Theta^0).$$

Here $g(t) = g$ is the indicator for presence of Gaussian g at time t . Thus, $\gamma(t, g)$ is the conditional probability of observing Gaussian g at time t given the training acoustic data and reference word scripts. To maximize $Q_{em}(\Theta|\Theta^0)$ we see that the optimal value of π_g is $\tilde{\pi}_g / \sum_{g' \in \mathcal{M}(s)} \tilde{\pi}_{g'}$. Optimization with respect to the rest of the model parameters is performed using a numerical optimization package; the details of this procedure can be found in Visweswariah et.al. [11].

We note that the SPAM model (3) can be directly generalized by representing the entire vector θ_g in *one* affine subspace. Our treatment presented in this paper carries over to this generalization, however it is discussed elsewhere [11] and in this paper we shall focus on the SPAM model of (3).

We now present two procedures for estimating the SPAM model parameters in a discriminative manner.

3. MMI Estimation of SPAM Parameters

Using (X^*, W^*) to denote a labeled training dataset, the MMI objective function [7, 1, 2] is given as

$$R(\Theta) = \frac{P(X^*|W^*; \Theta)}{P(X^*; \Theta)} \quad (5)$$

For maximizing $R(\Theta)$ we follow the extended Baum-Welch procedure first proposed for discrete distributions by Gopalakrishnan et.al. [7] and subsequently extended to continuous distributions by Normandin [1]. An alternate derivation of Normandin's procedure was recently proposed by Gunawardana et.al. [8], which we now recall.

Using \mathcal{G} to denote the set of all Gaussian sequences that can produce the acoustics X^* , and assuming that the Gaussian priors are fixed (not a part of Θ), we can write

$$\begin{aligned} P(X^*|W^*; \Theta) &= \sum_{G \in \mathcal{G}} P(X^*|G; \Theta) P(G|W^*) \\ P(X^*) &= \sum_{G \in \mathcal{G}} P(X^*|G; \Theta) P(G). \end{aligned}$$

In HMM based systems $P(G)$ includes the Gaussian priors, state-to-state transition probabilities, and a language model probability.

Define

$$\begin{aligned} F(\Theta|\Theta^0) &= \sum_{G \in \mathcal{G}} P(X^*|G; \Theta) P(G|W^*) \\ &\quad - R(\Theta^0) \sum_{G \in \mathcal{G}} P(X^*|G; \Theta) P(G) \\ &\quad + \int_X \sum_{G \in \mathcal{G}} D_G P(X|G; \Theta) P(G) \\ &= \sum_{G \in \mathcal{G}} \int_X q(X, G, \Theta^0) P(X|G; \Theta) \end{aligned}$$

where

$$\begin{aligned} q(X, G, \Theta^0) &= \delta(X - X^*) P(G|W^*) + D_G P(G) \\ &\quad - \delta(X - X^*) R(\Theta^0) P(G) \end{aligned} \quad (6)$$

where $\delta(X - X^*)$ is a delta function centered at X^* . It is easily verified that $F(\Theta^0|\Theta^0) = \sum_{G \in \mathcal{G}} D_G P(G)$, and if a Θ^1 can be found such that $F(\Theta^1|\Theta^0) > F(\Theta^0|\Theta^0)$, then $R(\Theta^1) > R(\Theta^0)$.

Now define the following auxiliary function

$$\begin{aligned} Q(\Theta|\Theta^0) &= \sum_{G \in \mathcal{G}} \int_X q(X, G, \Theta^0) P(X|G; \Theta) \log P(X|G; \Theta) \end{aligned} \quad (7)$$

It can be reasoned [8] that if, for each G , a D_G is chosen large enough such that $q(X, G, \Theta^0)$ is positive for all X , then $Q(\Theta|\Theta^0)$ is a valid auxiliary function for $F(\Theta|\Theta^0)$, i.e., increasing $Q(\Theta|\Theta^0)$ would lead to an increase in $F(\Theta|\Theta^0)$, and consequently an increase in $R(\Theta)$. However, we note that due to the presence of $\delta(X - X^*)$ functions, such a D_G will not generally exist. In practice these values are chosen based on some other heuristics, as will be discussed later in this paper.

Having chosen suitable D_G , we can simplify $Q(\Theta|\Theta^0)$ by substituting $q(X, G, \Theta^0)$ back and dividing through by $P(X^*|W^*; \Theta^0)$

$$\begin{aligned} Q(\Theta|\Theta^0) &= \sum_{G \in \mathcal{G}} P(G|X^*, W^*; \Theta^0) \log P(X|G; \Theta) \\ &\quad - \sum_{G \in \mathcal{G}} P(G|X^*; \Theta^0) \log P(X|G; \Theta) \\ &\quad + \sum_{G \in \mathcal{G}} \int_X \tilde{D}_G P(G) P(X|G; \Theta^0) \log P(X|G; \Theta) \end{aligned}$$

where $\tilde{D}_G = D_G / P(X^*|W^*; \Theta^0)$. The first term is the EM auxiliary function which simplifies to (4) without the $\tilde{\pi}_g \log \pi_g$ term since we did not consider Gaussian priors a part of Θ . The second term is similar except it is with the statistics \tilde{f}_g^d and $\tilde{\pi}_g^d$ that are gathered with denominator occupancy counts

$$\gamma^d(t, g) = P(g(t) = g | X^*; \Theta^0). \quad (8)$$

Lastly, the third term can be rearranged as

$$\sum_g D_g \left(\theta_g^T E_{\theta_g^0}(f(x)) + K(\theta_g) \right) \quad (9)$$

where $E_{\theta_g^0}(f(x))$ is the expected value of $f(x)$ under θ_g^0 , and

$$D_g = \sum_t \sum_{G: G(t)=g} \tilde{D}_G P(G) \quad (10)$$

The resulting Q function is

$$Q(\Theta|\Theta^0) = \sum_g \theta_g^T \left(\tilde{f}_g - \tilde{f}_g^d + D_g E_{\theta_g^0}(f(x)) \right) + \left(\tilde{\pi}_g - \tilde{\pi}_g^d + D_g \right) K(\theta_g) \quad (11)$$

Since the form of this auxiliary function is identical to $Q_{em}(\Theta|\Theta^0)$ (4), it can be maximized using the same numerical optimization procedure that was used for (4).

3.1. Selecting Gaussian Dependent D_g

We first note that the formulation presented here naturally allows for a Gaussian dependent D_g value according to (10); something that has empirically been found to be of value in MMI estimation [2]. However, deriving D_g according to (10) where \tilde{D}_G values are selected to ensure positivity of $q(X, G, \Theta^0)$ of (6) is impractical, if not impossible. We instead follow a D_g selection procedure that is analogous to a method described by Woodland et.al. [2]

$$D_g = \max \left(C_1 \tilde{\pi}_g^d, C_2 D_g^* \right) \quad (12)$$

where D_g^* is the smallest value such that when a full covariance matrix (diagonal in [2]) is estimated from the MMI stats, it comes out to be positive definite. The MMI update of a full covariance matrix is given as

$$\hat{\Sigma}_g = \frac{\tilde{f}_g^{(F)} - \tilde{f}_g^{d(F)} + D_g(\Sigma_g^0 + \mu_g^0 \mu_g^{0T})}{\tilde{\pi}_g - \tilde{\pi}_g^d + D_g} - \hat{\mu}_g \hat{\mu}_g^T \quad (13)$$

where superscript (F) denotes the xx^T portion of the stats, and

$$\hat{\mu}_g = \frac{\tilde{f}_g^{(L)} - \tilde{f}_g^{d(L)} + D_g \mu_g^0}{\tilde{\pi}_g - \tilde{\pi}_g^d + D_g} \quad (14)$$

with superscript (L) denoting the x portion of the \tilde{f} statistics. Substituting $\hat{\mu}_g$ from (14) into (13) results in the following quadratic eigenvalue problem

$$\begin{aligned} 0 &= (A_0 + D_g A_1 + D_g^2 A_2) y \\ A_0 &= \tilde{c}_g \tilde{\Sigma} - \tilde{\mu} \tilde{\mu}^T \\ A_1 &= \tilde{c}_g (\Sigma_g^0 + \mu_g^0 \mu_g^{0T}) - \mu_g^0 \tilde{\mu}_g^T - \tilde{\mu}_g \mu_g^{0T} + \tilde{\Sigma}_g \\ A_2 &= \Sigma_g^0 \\ \tilde{c}_g &= \tilde{\pi}_g - \tilde{\pi}_g^d \\ \tilde{\mu}_g &= \tilde{f}_g^{(L)} - \tilde{f}_g^{d(L)} \\ \tilde{\Sigma}_g &= \tilde{f}_g^{(F)} - \tilde{f}_g^{d(F)} \end{aligned}$$

The largest positive real D_g for which there exists a y solving the quadratic eigenvalue problem is the desired D_g .

3.2. Handling Priors

Considering Gaussian priors to be part of the model parameters to be estimated discriminatively leads to adding a $(\tilde{\pi}_g - \tilde{\pi}_g^d + D_g) \log \pi_g$ term to the MMI auxiliary function of (11). An alternative prior update that has been reported to result in larger objective function increase [2] is obtained by maximizing

$$\sum_{g \in \mathcal{M}(s)} \tilde{\pi}_g \log \pi_g - \frac{\tilde{\pi}_g^d}{\pi_g^0} \pi_g \quad (15)$$

subject to the constraint $\sum_{g \in \mathcal{M}(s)} \pi_g = 1$. In this paper we experiment with both these methods, as well as with updating the priors with the numerator counts $\tilde{\pi}_g$.

4. Error Weighted Training

Our second training approach is motivated by the idea of paying specific attention to the training cases where the current model is making a mistake. This is the basic idea behind a number of training procedures such as boosting [9] and error corrective training [10].

We follow a simple approach whereby we first determine the training sentences that are in error under the current model. We then gather the statistics mentioned in (4) on these sentences, say $\tilde{f}_g^e, \tilde{\pi}_g^e$, in addition to the statistics on all the sentences, $\tilde{f}_g, \tilde{\pi}_g$. These two sets are combined to create new stats as

$$\begin{aligned} \tilde{f}_g^n &= \tilde{f}_g + \alpha \tilde{f}_g^e \\ \tilde{\pi}_g^n &= \tilde{\pi}_g + \alpha \tilde{\pi}_g^e \end{aligned} \quad (16)$$

The SPAM model parameter estimation is then carried out according to the description of Section 2.

Our error-weighted procedure described above has the advantage that it is exceedingly simple to implement. On the other hand, it is expected to work best when the training sentence error rate is already fairly small. We view it as a "cheap" version of boosting, which would provide a more complex hierarchical structure for the weighting of error data.

5. Experimental Results

All experiments reported in this paper were conducted on a testset containing digit strings of constrained length (seven and ten). These strings were recorded in a car under three conditions: idling, moving at about 30 miles per hour, and moving at about 60 miles per hour. There are altogether 10298 sentences and 71084 words in the test set.

The acoustic feature vectors were obtained by first computing 13 Mel-cepstral coefficients (including energy) for each time slice under a 250 msec. window with a 15 msec. shift. Nine such vectors were concatenated and projected to a 52 dimensional space using LDA. The acoustic models were built on these features with a digit specific phone set containing 37 phones. Each phone was modeled with a three state left to right HMM. This, in addition to six silence states results in 117 context independent states. Each digit phone state was modeled with a mixture of 15 Gaussians, and the each silence state was modeled with a mixture of 100 Gaussians.

The training data set contains about 462K sentences which are a mix of digit and non-digit word strings. The digit specific subset was collected in a car under the three conditions described above; however the majority of the digit data is under the idling condition. This digit specific subset has 66608 sentences.

The seed for our baseline SPAM model (D=13, L=13 in (3)) was built using the procedure specified in [11], using full covariance statistics collected on all of the training data. This model had a word/sentence error rate of 2.14/12.64. Maximum likelihood training of this model was carried out using the procedure described in [11] on the digit portion of the training data. The resulting model had a word/sentence error rates of 1.78/10.41; this was used as the baseline model for all the discriminative training experiments described in this paper.

5.1. Error-Weighted Training

A digit training data decode with the baseline SPAM model resulted in a word/sentence error rate of 36/10.19. The 10.19%

sentences with errors were then used to collect statistics, in addition to statistics collected over the entire training data, to carry out error-weighted training.

The error rate performance of this training method is presented in Table 1. We note that there is a striking reduction in the error rate, even when optimizing just the un-tied parameters, with increasing weight on the statistics gathered from the sentences decoded incorrectly. Optimization of tied parameters in conjunction with un-tied parameters reduces the error rates further, resulting in the largest sentence error rate improvement of about 22% relative over the baseline.

| baseline word/sentence error rate : 1.78/10.41 | | | | | | | |
|--|-----|-----------------|------|------|------|------|------|
| optimized parameters | | α values | | | | | |
| | | 0 | 8 | 64 | 128 | 256 | 1024 |
| un-tied | wer | 1.79 | 1.55 | 1.46 | 1.45 | 1.45 | 1.45 |
| | ser | 10.46 | 9.37 | 8.97 | 8.97 | 9.00 | 9.03 |
| all | wer | 1.78 | 1.54 | 1.40 | 1.39 | 1.39 | 1.38 |
| | ser | 10.40 | 9.16 | 8.56 | 8.50 | 8.49 | 8.44 |

Table 1: Error weighted training with different values of α

5.2. MMI Estimation

| baseline word/sentence error rate : 1.78/10.41 | | | | | | | |
|--|-----|----------------------------|------|------|------|------|-------|
| optimized parameters | | C_1 values, $C_2 = 2.0$ | | | | | |
| | | 1.0 | 0.5 | 0.25 | 0.1 | 0.05 | 0.01 |
| un-tied | wer | 1.69 | 1.62 | 1.56 | 1.44 | 1.41 | 1.43 |
| | ser | 10.0 | 9.68 | 9.37 | 8.76 | 8.59 | 8.67 |
| all | wer | 1.68 | 1.60 | 1.55 | 1.40 | 1.39 | 1.41 |
| | ser | 9.97 | 9.58 | 9.37 | 8.54 | 8.41 | 8.59 |
| optimized parameters | | C_2 values, $C_1 = 0.05$ | | | | | |
| | | 2.0 | 1.5 | 1.1 | 1.05 | 1.01 | 1.005 |
| un-tied | wer | 1.41 | 1.36 | 1.37 | 1.36 | 1.36 | 1.36 |
| | ser | 8.59 | 8.31 | 8.46 | 8.41 | 8.41 | 8.41 |
| all | wer | 1.39 | 1.36 | 1.36 | 1.36 | 1.36 | 1.36 |
| | ser | 8.41 | 8.25 | 8.31 | 8.33 | 8.35 | 8.35 |

Table 2: MMIE with different C_1 and C_2 values

The MMI numerator statistics ($\tilde{f}_g, \tilde{\pi}_g$) and denominator statistics ($\tilde{f}_g^d, \tilde{\pi}_g^d$) were collected under an acoustic scaling [2] of 1.0. These statistics were combined to form the auxiliary function of (11) with a D_g value selected according to (12). The three different prior update methods discussed in Section 3.2 were within 0.5% relative of each other, hence the third method (updating priors based on $\tilde{\pi}_g$) was used in all the experiments described in the following.

The values of constants C_1 and C_2 in (12) were searched over sequentially, starting from their recommended values of 1.0 and 2.0, respectively [2]. We first found the optimal value of C_1 keeping C_2 fixed at 2.0, and then C_2 was searched over with C_1 fixed at this optimal value. These results are presented in Table 2.

We note that the optimal MMI performance of 1.36/8.25% is only marginally better than the optimal error-weighted performance of 1.38/8.44%.

5.3. Combination of Error-Weighted Training and MMIE

Comparing the error rate performance of un-tied parameter estimation under the two discriminative criteria (Tables 1 and 2), it appears that MMIE is significantly better than error-weighted training at estimating these parameters. However when both tied and un-tied parameters are estimated the error-weighted training is quite close in performance to MMIE, suggesting that error-weighted training may be better at estimating the tied model parameters. To confirm this hypothesis, we combined the two estimation procedures by taking the tied model from error-weighted training ($\alpha = 1024$) and updated the un-tied parameters using MMIE. This resulted in our best model with a word/sentence error rate of 1.27/8.13% which is a word error rate improvement of over 28% relative over the baseline.

6. Acknowledgements

Authors would like to thank Stanley Chen for marvelous code-base infrastructure that enabled the experiments described in this paper. Also many thanks to George Saon for useful discussions.

7. References

- [1] Normandin, Y., "Hidden Markov models, maximum mutual information estimation and the speech recognition problem," Ph.D. Thesis, McGill University, Montreal, 1991.
- [2] Woodland, P. and Povey D., "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, 16:25–47, 2002.
- [3] Gales, M., "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. on Speech and Audio Proc.*, 1999.
- [4] Gopinath, R., "Maximum likelihood modeling with Gaussian distributions for classification," *ICASSP*, 1998.
- [5] Olsen, P. and Gopinath, R., "Modeling inverse covariance matrices by basis expansion," *ICASSP*, 2002.
- [6] Axelrod, S., Gopinath, R., Olsen, P., and Visweswariah, K., "Dimensional reduction, covariance modeling and computational complexity in ASR systems," *ICASSP* 2003.
- [7] Gopalakrishnan, P., Kanevsky, D., Nadas, A., and Nahamoo, D., "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Info. Theory*, 37:107–113, 1991.
- [8] Gunawardana, A. and Byrne, W., "Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression," *ICASSP*, 2002.
- [9] Freund Y. and Schapire, R., "Decision theoretic generalization of on-line learning and an application to boosting," *Second European Conference on Computational Learning Theory*, 1995.
- [10] Bahl, L., Brown, P., de Souza, P., and Mercer, R., "Estimating hidden Markov model parameters so as to maximize speech recognition accuracy," *IEEE Trans. on Speech and Audio Proc.*, 1(1):77–83, 1993.
- [11] Visweswariah, K., Axelrod, S., and Gopinath, R., "Acoustic modeling with mixtures of subspace constrained exponential models," submitted to *Eurospeech* 2003.