FEATURE DIMENSION REDUCTION USING REDUCED-RANK MAXIMUM LIKELIHOOD ESTIMATION FOR HIDDEN MARKOV MODELS

Don X. Sun
Statistics and Information Analysis Research
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974
Email: dxsun@bell-labs.com

ABSTRACT

This paper presents a new method of feature dimension reduction in hidden Markov modeling (HMM) for speech recognition. The key idea is to apply reduced rank maximum likelihood estimation in the M-step of the usual Baum-Welch algorithm for estimating HMM parameters such that the estimates of the Gaussian distribution parameters are restricted in a sub-space of reduced dimensionality. There are two main advantages of applying this method in HMM: 1) feature dimension reduction is achieved simultaneously with the estimation of HMM parameters, therefore it guarantees that the likelihood function is monotonically increasing; 2) it requires very little extra computation in addition to the standard Baum-Welch algorithm, hence it can be easily incorporated in the existing speech recognition systems using HMMs.

1. INTRODUCTION

In the past few years, the statistical method of linear discriminant analysis (LDA) and principle component analysis (PCA) have been explored for reducing the dimensionality of feature vectors that are used in speech recognition system. LDA is often preferred in comparison with PCA since it gives projections to lower dimensional space that separate different classes the most, while PCA only searches for the directions that have the largest variations.

Although LDA is appealing for dimension reduction in pattern classification, there is one major difficulty in applying LDA to speech recognition problems. Since LDA requires known class membership of each sample in order to compute the within-class and between-class covariance matrices, and most speech training data does not provide precise segmentation information of the speech units that are used in HMMs, such as phonemes, sub-phonemes, microsegmental units of phoneme etc. In the literature, some methods have been proposed to combine segmental k-means algorithm with LDA. One approach is to apply Viterbi segmentation algorithm at the end of each iteration to get boundaries or class labels of all the samples, and then use the standard LDA method to calculate linear projections of the original feature vectors. The problem with this approach is that the feature vectors used in different steps of the Baum-Welch or segmental k-means algorithm are not consistent. Therefore, the desirable property of monotonically increasing likelihood function cannot be guaranteed. Another approach is to use minimum classification error as optimization criterion ([5],[8]) to derive optimal feature transformation. However, such optimization problem can be very difficult and computationally intensive.

In the proposed approach, by drawing the analogy between the ordinary LDA and the reduced-rank maximum likelihood estimation, we can achieve feature dimension reduction without losing the property of the standard EM algorithm for estimating HMM parameters. The idea of the reduced-rank maximum likelihood estimation is based on a well known result in multivariate analysis for estimating single Gaussian distribution parameters under rank constraints ([6],[3]).

The paper is organized as follows. In Section 2 we give a brief review of the ordinary LDA and the relationship between LDA and the constrained maximum likelihood estimation problem. Section 3 establishes some basic notations used in the EM algorithm for estimating parameters in HMM with mixture of Gaussian distributions. In section 4, we present the reduced-rank maximum likelihood estimation algorithm for hidden Markov models.

2. LINEAR DISCRIMINANT ANALYSIS AND REDUCED-RANK MAXIMUM LIKELIHOOD ESTIMATION

Fisher's linear discriminant analysis (LDA) is a popular method for classification based on multi-dimensional predictor variables. The basic idea is to find a low dimensional projection of the raw data such that in the resulting space, the ratio of between-class variation and within-class variation is the largest.

Let p-dimensional vector $(\mathbf{x}_{ji})_{p\times 1}$ represent the observation vector of i-th sample from j-th class, for $j = 1, \cdots, J$ and $i = 1, \cdots, n_j$ where J is the number of classes, and let

$$\bar{\mathbf{x}}_{\cdot\cdot} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{x}_{ji}, \quad \text{and} \quad \bar{\mathbf{x}}_{j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_{ji}$$

be the overall mean and the mean of j-th class, where $n = \sum_{j=1}^{J} n_j$. The total sample covariance matrix

$$\mathbf{S}_T = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{\mathbf{x}}_{ji} - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_{ji} - \bar{\mathbf{x}}_{\cdot\cdot})^T$$

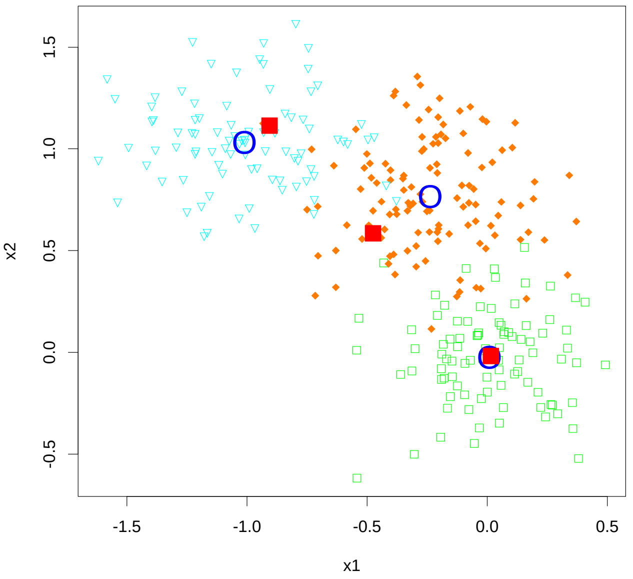

Figure 1. Illustration of reduced-rank estimation of group means: The three circles represent the original group means and the three solid boxes are the constrained estimates of the group means in a one-dimensional space.

can be decomposed into between-class and within-class sample covariance matrices as follows:

$$\begin{aligned}\mathbf{S}_T &= \mathbf{S}_B + \mathbf{S}_W \\ &= \frac{1}{n}\sum_{j=1}^{J} n_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{\cdot\cdot})^T \\ &\quad + \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\mathbf{x}_{ji} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ji} - \bar{\mathbf{x}}_j)^T.\end{aligned}$$

The standard linear discriminant variates are derived through successively maximizing the ratio of between-group to within-group variance of linear combinations of the original variables. The resulting feature transformation matrix is in fact the matrix of leading k eigen vectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

Another way of deriving the linear discriminant projections is through the constrained (reduced-rank) maximum likelihood estimation for single Gaussian distribution models ([3]). Consider the maximization of the Gaussian log-likelihood

$$l(\mu_j, \mathbf{\Sigma}) = -\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\mathbf{x}_{ij} - \mu_j)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_{ij} - \mu_j) - n\log|\mathbf{\Sigma}|$$

subject to the constraints $\mu_j = A\nu_j$, where A is any p by k ($k \leq p$) matrix of full rank, and ν_j a $k \times 1$ vector. There exists an explicit solution to this maximization problem:

$$\begin{aligned}\hat{\mu}_j &= {\mathbf{S}_W}^{1/2} V V^T {\mathbf{S}_W}^{-1/2}\bar{\mathbf{x}}_j \\ \hat{\mathbf{\Sigma}} &= \mathbf{S}_W + {\mathbf{S}_W}^{1/2} U U^T {\mathbf{S}_W}^{-1/2}\mathbf{S}_B{\mathbf{S}_W}^{-1/2} U U^T {\mathbf{S}_W}^{1/2},\end{aligned}$$

where V is the matrix of leading k eigen vectors of ${\mathbf{S}_W}^{-1/2}\mathbf{S}_B{\mathbf{S}_W}^{-1/2}$ and U the matrix of remaining $p-k$ eigen vectors. $\hat{\mu}_j$ is essentially the projection of the jth group mean onto the discriminant subspace of rank k. Figure 1 shows a simple example of reduced-rank estimates of three group means.

To generalize this result to hidden Markov models, we consider the maximization of the joint probability for HMM $P(\mathbf{O}|\theta)$ subject to the reduced-rank constraint of the mean parameters of the Gaussian distributions. Before we present the reduced-rank estimation, some basic notations and framework for HMM with mixture of Gaussian distributions are briefly described in the next section.

3. HIDDEN MARKOV MODELS WITH MIXTURE OF GAUSSIAN DISTRIBUTIONS

Under the assumption for standard HMM, the likelihood function of an observation sequence $\mathbf{O}$ can be expressed as

$$P(\mathbf{O}|\theta) = \sum_{\text{all s}} P(\mathbf{O}, \mathbf{S}|\theta) = \sum_{\text{all s}} \prod_{t=1}^{T}[a_{s_{t-1},s_t} \cdot b_{s_t}(\mathbf{O}_t)] \quad (1)$$

where a_{s_{t-1},s_t} is the transition probability from state s_{t-1} to s_t and $b_{s_t}(\mathbf{O}_t)$ is the probability density function of the observation $\mathbf{O}_t$ at given states s_t. The summation is over all possible state sequences.

The parameters in the HMM can be estimated through the EM algorithm ([2]), also known as Baum-Welch algorithm ([1, 7, 4]). The maximization of the original likelihood function $P(\mathbf{O}|\theta)$ is achieved through iterative maximizations of the following auxiliary function:

$$Q(\theta|\theta_0) = \sum_{\text{all s}} P(\mathbf{O}, \mathbf{S}|\theta_0) \cdot \log P(\mathbf{O}, \mathbf{S}|\theta)$$

where θ_0 is assumed to be constant and the maximization is with respect to θ.

When the state-dependent probability $b_{s_t}(\mathbf{O}_T)$ is modeled by a multi-dimensional mixture Gaussian distributions, the "missing" information in the likelihood has two components: 1) the unobserved index of state; and 2) the unobserved index of mixing component for a given state. The formulation of the EM algorithm for single Gaussian can be generalized to the mixture distribution case in a straightforward way ([4]). We give a brief description here for the discussion of the proposed method in Section 4.

Let

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm} b_{jm}(\mathbf{O}) = \sum_{m=1}^{M} c_{jm} N(\mathbf{O}, \mu_{jm}, \mathbf{\Sigma}_{jm})$$

where $N(\mathbf{O}, \mu, \mathbf{\Sigma})$ denotes a multi-variate Gaussian density function with mean μ and covariance matrix $\mathbf{\Sigma}$; M denotes the number of mixture components and c_{jm} is the weight for the mth mixture component satisfying $\sum_{m=1}^{M} c_{jm} = 1$.

If we use $\mathbf{M} = (m_1, \cdots, m_T)$ to denote a mixture-component-index sequence, we can express the joint probability of the observation sequence and the state sequence can be expressed as a summation of the joint probabilities of $(\mathbf{O}, \mathbf{S}, \mathbf{M})$:

$$\begin{aligned} P(\mathbf{O}, \mathbf{S}|\theta) &= \prod_{t=1}^{T}[a_{s_{t-1}s_t} b_{s_t}(\mathbf{O}_t)] \\ &= \sum_{m_1=1}^{K} \cdots \sum_{m_T=1}^{K} \prod_{t=1}^{T}[a_{s_{t-1}s_t} c_{s_t m_t} b_{s_t m_t}(\mathbf{O}_t)] \\ &= \sum_{\text{all } \mathbf{M}} P(\mathbf{O}, \mathbf{S}, \mathbf{M}|\theta) \end{aligned}$$

Therefore, the joint probability of the entire observation sequence is

$$P(\mathbf{O}|\theta) = \sum_{\text{all } \mathbf{S}} \sum_{\text{all } \mathbf{M}} P(\mathbf{O}, \mathbf{S}, \mathbf{M}|\theta). \tag{2}$$

The auxiliary function for the EM algorithm can then be generalized to

$$Q(\theta|\theta_0) = \sum_{\text{all } \mathbf{S}} \sum_{\text{all } \mathbf{M}} P(\mathbf{O}, \mathbf{S}, \mathbf{M}|\theta_0) \log P(\mathbf{O}, \mathbf{S}, \mathbf{M}|\theta) \tag{3}$$

where

$$\begin{aligned} \log P(\mathbf{O}, \mathbf{S}, \mathbf{M}|\theta) &= \sum_{t=1}^{T} \log(a_{s_{t-1}s_t}) + \sum_{t=1}^{T} \log(c_{s_t m_t}) \\ &+ \sum_{t=1}^{T} \log(b_{s_t m_t}(\mathbf{O}_t)). \end{aligned}$$

Since the parameters $a_{s_{t-1}s_t}, c_{s_t m_t}, \mu_{jm}, \mathbf{\Sigma}_{jm}$ are involved in different additive terms, the optimization can be done separately. This paper is only concerned with the estimation of the mean parameters μ_{jm} under reduced-rank constraints. The estimation of the other parameters are the same as the standard Baum-Welch algorithm.

4. REDUCED RANK ESTIMATION

As the dimensionality of the feature vector space increases, estimation of model parameters becomes more difficult due to the sparsity of training data in high-dimensional space. Therefore it is desirable to consider the problem of dimension reduction before estimating model parameters for pattern recognition.

The most common approach of dimension reduction is to use linear discriminant analysis method which derives linear transformations from high dimensional space to low dimensional space. However, the difficulty in applying LDA to speech data is that the class labels of training data is not completely known (in fact, hidden Markov models do not rely too much on the precise segmentation information in the training data). To avoid such difficulty, we propose a rank-constrained maximum likelihood estimation method to achieve dimension reduction simultaneously with the estimation of hidden Markov model parameters.

In the EM-algorithm for standard HMM estimation, the maximization of the original likelihood in (1) and (2) can be achieved by iteratively maximizing the auxiliary functions in (3). Similarly, we can prove that the maximization of the original likelihood subject to the constraint of reduced-rank of the mean parameters can also be achieved by iteratively maximizing the auxiliary functions subject to the same constraints. In fact the constraints only apply to the maximization of the auxiliary function $Q_b(\theta|\theta_0)$ that involves μ_{jm} and $\mathbf{\Sigma}$:

$$\begin{aligned} Q_b(\theta|\theta_0) &= P(\mathbf{O}, \mathbf{S}, \mathbf{M}|\theta_0) \left[\sum_{t=1}^{T} \log(b_{s_t m_t}(\mathbf{O}_t))\right] \\ &= \sum_{j,m} \sum_{t=1}^{T} P(\mathbf{O}, s_t = j, m_t = m|\theta_0) \log(b_{s_t m_t}(\mathbf{O}_t)) \\ &= \sum_{j,m} \sum_{t=1}^{T} \xi_t(j,m) \log(b_{s_t m_t}(\mathbf{O}_t)) \end{aligned}$$

where $\xi_t(j,m) = P(\mathbf{O}, s_t = j, m_t = m|\theta_0)$ is an intermediate probability. It can be calculated as:

$$\xi_t(j,m) = \alpha_{t-1}(i) a_{ij} c_{jm} b_{jm}(\mathbf{O}_t) \beta_t(j)$$

where

$$\begin{aligned} \alpha_t(j) &= P(\mathbf{O}_1, \cdots, \mathbf{O}_t; s_t = j|\theta_0) \\ \beta_t(j) &= P(\mathbf{O}_{t+1}, \cdots, \mathbf{O}_T | s_t = j, \theta_0) \end{aligned}$$

are the usual forward and backward probabilities.

The auxiliary function $Q_b(\theta|\theta_0)$ can be decomposed into separate parts as follows:

$$\begin{aligned} -2Q_b(\theta|\theta_0) &= \sum_{j,m} \sum_{t=1}^{T} \xi_t(j,m) \log(b_{s_t m_t}(\mathbf{O}_t)) \\ &= \sum_{j,m} \sum_{t=1}^{T} \xi_t(j,m)(\mathbf{O}_t - \mu_{jm})^T \mathbf{\Sigma}^{-1} (\mathbf{O}_t - \mu_{jm}) \\ &\quad + \sum_{j,m} \sum_{t=1}^{T} \xi_t(j,m) \log(|2\pi\mathbf{\Sigma}|) \\ &= \sum_{j,m} \sum_{t=1}^{T} \xi_t(j,m)(\mathbf{O}_t - \bar{\mathbf{O}}_{jm})^T \mathbf{\Sigma}^{-1} (\mathbf{O}_t - \bar{\mathbf{O}}_{jm}) \\ &\quad + \sum_{j,m} \xi(j,m)(\bar{\mathbf{O}}_{jm} - \mu_{jm})^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{O}}_{jm} - \mu_{jm}) \\ &\quad + \sum_{j,m} \sum_{t=1}^{T} \xi_t(j,m) \log(|2\pi\mathbf{\Sigma}|) \end{aligned} \tag{4}$$

where

$$\bar{\mathbf{O}}_{jm} = \frac{\sum_{t=1}^{T} \xi_t(j,m) \mathbf{O}_t}{\sum_{t=1}^{T} \xi_t(j,m)}, \quad \text{and} \quad \xi(j,m) = \sum_{t=1}^{T} \xi_t(j,m).$$

The objective is to find μ_{jm} and $\mathbf{\Sigma}$ such that the auxiliary function above is minimized subject to the constraint that the estimates of all the mean parameters stay within a k-dimensional subspace of the original feature space, i.e., $\mu_{jm} = A\nu_{jm}$, where A is any p by k ($k \leq p$) matrix of full rank, and ν_{jm} a $k \times 1$ vector. For a given $\mathbf{\Sigma}$, there is only one part in (4) that involves μ_{jm}. Let

$$\mu^*_{jm} = \mathbf{\Sigma}^{-1/2}\mu_{jm}, \quad \bar{\mathbf{O}}^*_{jm} = \mathbf{\Sigma}^{-1/2}\bar{\mathbf{O}}_{jm}$$

and

$$\mu^*_{jm} = (V|U)\begin{pmatrix} \lambda_{jm} \\ \mathbf{0} \end{pmatrix}$$

where V is a $p \times k$ matrix, λ_{jm} a $k \times 1$ vector and $(V|U)$ a $p \times p$ orthogonal matrix, we then have

$$\begin{aligned} & \sum_{j,m} \xi(j,m)(\bar{\mathbf{O}}_{jm} - \mu_{jm})^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{O}}_{jm} - \mu_{jm}) \\ = & \sum_{j,m} \xi(j,m)(\bar{\mathbf{O}}^*_{jm} - \mu^*_{jm})^T (\bar{\mathbf{O}}^*_{jm} - \mu^*_{jm}) \\ = & \sum_{j,m} \xi(j,m)(V^T\bar{\mathbf{O}}^*_{jm} - \lambda_{jm})^T (V^T\bar{\mathbf{O}}^*_{jm} - \lambda_{jm}) \\ & + \sum_{j,m} \xi(j,m)\bar{\mathbf{O}}^{*T}_{jm} U U^T \bar{\mathbf{O}}^*_{jm}. \end{aligned}$$

It is easy to show that this term is minimized when V is the matrix of leading k eigen vectors of $\sum_{j,m} \xi(j,m)\bar{\mathbf{O}}^*_{jm}\bar{\mathbf{O}}^{*T}_{jm}$ and $\lambda_{jm} = V^T\bar{\mathbf{O}}^*_{jm}$, which leads to

$$\hat{\mu}_{jm} = \mathbf{\Sigma}^{1/2} V V^T \mathbf{\Sigma}^{-1/2} \bar{\mathbf{O}}_{jm}. \tag{5}$$

Furthermore, when $\mathbf{\Sigma}$ is unknown, we can prove that the estimate is obtained similarly by replacing $\mathbf{\Sigma}$ with W in (5), where

$$W = \frac{\sum_{j,m}\sum_{t=1}^{T} \xi_t(j,m)(\mathbf{O}_t - \bar{\mathbf{O}}_{jm})(\mathbf{O}_t - \bar{\mathbf{O}}_{jm})^T}{\sum_{j,m}\sum_{t=1}^{T} \xi_t(j,m)}.$$

In summary, we have the explicit solution to the reduced-rank maximization as follows:

$$\hat{\mu}_{jm} = W^{1/2} V V^T W^{-1/2} \bar{\mathbf{O}}_{jm}$$

where V is the matrix of k leading eigenvectors of

$$W^{-1/2} B W^{-1/2}$$

and

$$B = \frac{\sum_{j,m} \xi(j,m)\bar{\mathbf{O}}_{jm}\bar{\mathbf{O}}^T_{jm}}{\sum_{j,m} \xi(j,m)}.$$

The maximum likelihood estimate for the covariance matrix is:

$$\begin{aligned} \hat{\mathbf{\Sigma}} &= W + \frac{\sum_{j,m}\sum_{t=1}^{T} \xi_t(j,m)(\bar{\mathbf{O}}_{jm} - \hat{\mu}_{jm})(\bar{\mathbf{O}}_{jm} - \hat{\mu}_{jm})^T}{\sum_{j,m}\sum_{t=1}^{T} \xi_t(j,m)} \\ &= W + W^{1/2} U U^T W^{-1/2} B W^{-1/2} U U^T W^{1/2}, \end{aligned}$$

where U is the matrix of the remaining $p - k$ eigenvectors of $W^{-1/2} B W^{-1/2}$.

Since V is a $p \times k$ matrix, it is apparent that all the $\hat{\mu}_{jm}$'s stay in a k dimensional space. If we choose $k = p$, then V is a $p \times p$ orthogonal matrix, and we have:

$$\hat{\mu}_{jm} = \bar{\mathbf{O}}_{jm}, \quad \hat{\mathbf{\Sigma}} = W.$$

which are the ordinary estimates for the standard HMM with mixture of Gaussian distributions.

5. CONCLUSION

This paper proposed a new approach of feature space dimensionality reduction within the framework of hidden Markov models. It avoids the difficulty in applying linear discriminant analysis to speech data which relies on precise segmentation information of HMM states. Also, the idea of using constrained maximum likelihood estimation can be extended to guarantee desirable properties of the estimates in HMMs. One important case could be imposing smoothness constraints on the estimates to avoid unstable estimation when training data is not sufficient.

REFERENCES

[1] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.*, 39:1–38, 1977.

[3] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. Technical report, AT&T Bell Laboratories, 1994.

[4] B. H. Juang. Maximum likelihood estimation for multivariate observations of markov sources. *AT&T Technical Journal*, 64:1235–1249, 1985.

[5] B. H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. Signal Processing*, 40(12):3043–3054, 1992.

[6] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, Harcourt Brace & Co., New York, 1979.

[7] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77:257–285, 1989.

[8] C. Rathinavelu and L. Deng. Use of generalized dynamic feature parameters for speech recognition: maximum likelihood and minimum classification error approaches. In *Proceedings ICASSP*, pages 373–376, 1995.