REVIEW OF BAYSIAN SPEAKER ADAPTATION BASED ON PPCA

Kaihua Huang

Department of Electrical and Computer Engineering
Mississippi State University
Mississippi State, MS 39762 USA
email: huang@isip.mstate.edu

ABSTRACT

The Probabilistic Principal Component Analysis (PPCA) is a new approach to find the canonical speaker models based on the expectation maximization (EM). This method provides not only the canonical speaker models but also a prior distribution of the model parameters, which can be directly applied to the maximum a posteriori (MAP) adaptation scheme. Some experiments using this technique showed the effectiveness of the PPCA-based approach compared to the other adaptation approaches with a small amount of data.This paper will focus on analysis the underlying theory of PPCA method. The derivation of the adaptation formula based on PPCA model will be investigated. The effectiveness and drawbacks of this approach are presented. At the same time, the content of using this method is described.

1. INTRODUCTION

Principal component analysis (PCA) is a well-established technique for dimension reduction, and a chapter on the subject may be found in practically every text on multivariate analysis. One limiting disadvantage of PCA is the absence of a probability density model and associated likelihood measure.However, PPCA offer a number of important advantages including:

The definition of a likelihood measure permits comparison with other density - estimation techniques and facilitates statistical testing.

Bayesian inference methods may be applied (e.g. for model comparison) by combining the likelihood with a prior.

If PCA is used to model the class-conditional densities in a classification problem, the posterior probabilities of class membership may be computed.

The probability density function gives a measure of the novelty of a new data point.

The single PCA model maybe extended to a mixture of such models.

Recently, there has been increasing interest in speaker adaptation techniques that require only a small amount of data from the target speaker. Such rapid adaptation schemes have been developed for modeling the dependencies between different speech units for effective use of a small adaptation data. The eigenvoice technique performs speaker adaptation by constructing a new speaker model as a weighted sum of eigen speaker models.To find the eigen speaker models which characterize a priori knowledge of training speaker, the conventional PCA method is applied to a set supervectors given by separate speaker dependent (SD) hidden Markov model (HMM) parameters. One drawbacks of the eigenvoice approach is that the adapted speaker model does not converge to the true SD model even when a large amount of adaptation data is available [1]. In order to alleviate this problem, the obtained eigen speaker models are used as the prior in formation for the maximum a posteriori MAP adaptation.

In this paper, we will review a new approach called PPCA-based Bayesian speaker adaptation which provide both canonical speaker models and a priori distribution of the model parameters, which can be directly applied by the MAP adaptation scheme.

2. PPCA METHOD

2.1. Latent Variable Models

A latent variable model is a basic concept applied in PPCA method and which seeks to relate the set of d-dimensional observed data vectors $\{\mathbf{t}_n\}$ to a corresponding set of q-dimensional latent variables $(\mathbf{x}_n)$:

$$t = y(x;\theta) + \varepsilon \tag{1}$$

where $y(x;\theta)$ is a function of the latent variable x with parameters θ and ε is an x-independent noise process.In standard factor analysis the mapping $y(x;\theta)$ is linear:

$$t = Wx + \mu + \varepsilon \tag{2}$$

where the latent variables **x ~ N(0,I)** have a unit isotropic Gaussian distribution. The error, or noise, model is Gaussian such that $\varepsilon \sim N(0, \Psi)$, with Ψ diagonal, the $(d \times q)$ parameter matrix W contains the factor loading, and μ is a constant whose maximum-likelihood estimator is the mean of data.Given this formulation, the model for t is also normal $N(\mu, C)$, where covariance $C = \Psi + WW^t$.

2.2. A Probability Model of PCA

Principal components emerges when the data assumed to comprise a systematic component, plus an independent error term for each variable with common variance σ^2. This implies that the diagonal elements of the error matrix Ψ in factor analysis above should be identical.As well as assumption the accuracy of the method, by considering the model given by (2) with an isotropic noise structure, such that $\Psi = \sigma^2 I$, an important consequence of this is that PCA may be expressed in terms of a density model, the definition of which now follows.

For the isotropic, noise model $\varepsilon = N(0, \sigma^2 I)$, equation (2) implies a probability distribution over t-space for a given by

$$p(t|x) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left\{-\frac{1}{2\sigma^2}\|t - Wx - \mu\|^2\right\} \tag{3}$$

Which a Gaussian prior over the latent variables defined by

$$p(x) = (2\pi)^{-2/q} \exp\left\{-\frac{1}{2}x^T x\right\} \tag{4}$$

we obtain the marginal distribution of t in the form

$$p(t) = \int p(t|x)p(x) \tag{5}$$

from this we have

$$p(t) = (2\pi)^{-d/2}|C|^{-1/2} \exp\left\{-\frac{1}{2}(t-\mu)^T C^{-1}(t-\mu)\right\} \tag{6}$$

where the model covariance is

$$C = \sigma^2 I + WW^T \tag{7}$$

Using Bayes's rule, the posterior distribution of the latent variables x given the observed t may be calculated:

$$\begin{aligned} p(x|t) = {} & (2\pi)^{-q/2}\left|\sigma^{-2}M\right|^{-1/2} \times \\ & \exp\left\{-\frac{1}{2}\{x - M^{-1}W^T(t-\mu)\}^T\right. \\ & \left.(\sigma^{-2}M)\{x - M^{-1}W^T(t-\mu)\}\right\} \end{aligned} \tag{8}$$

where the posterior covariance matrix is given by

$$\sigma^2 M^{-1} = \sigma^2(\sigma^2 I + W^T W)^{-1} \tag{9}$$

Note that M is $q \times q$ while C is $d \times d$.

The log-likelihood of observing the data under this model is:

$$L = \sum_{n=1}^{N} \ln\{p(t_n)\} \quad (10)$$
$$= -\frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln|C| - \frac{N}{2}tr[C^{-1}S]$$

where

$$S = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mu)(t_n - \mu)^T \quad (11)$$

the sample covariance matrix of the observed $\{t_n\}$. The parameters, say $\lambda(W, \mu, \sigma^2)$, for this model can thus be estimated by maximizing the log-likelihood L, and an EM algorithm to can use to achieve this[1].

2.3. PPCA-Based Speaker Adaptation

Here we review the PPCA-Based Adaptation proposed in [1].This approach only discuss the adaptation of mean vectors among the HMM parameters. Let O = $\{o_1, o_2, ..., o_t\}$ be a sequence of feature vectors generated by HMM. The observation density $p(o_t| i)$ fro state i is assumed to be a mixture of Gaussians,

$$p(o_t|i) = \sum_{j=1}^{K} \omega_{i,j} N(o_t; \mu_{i,j}, \Sigma_{i,j}) \quad (12)$$

where K is the number of mixtures, $\omega_{i,j}$ is the probability of mixture component j in state i, and N represents the conventional d-dimensional normal distribution.

Let $\mu = [\mu_{1,1}^T, ..., \mu_{N,K}^T]$ be a supervector that arguments all the Gaussian mean vectors. Assume μ is generated by a PPCA model with a latent variable x and parameters $\theta = \{\bar{\mu}, W, \sigma^2\}$ and $\lambda^{(n)} = \{\mu^{(n)}\}$ be the current estimate and $\lambda = \{\mu\}$ be the new estimate. Then the auxiliary function for the EM algorithm is defined by

$$R(\lambda, \lambda^{(n)}) = E[\ln p(O, S, C|\lambda)|(O, \lambda^{(n)})] \quad (13)$$
$$+ E[\ln p(\lambda, x|\theta)|\lambda^{(n)}]$$

where $S = \{s_1, ..., s_T\}$ represents the state sequence, $C = \{c_1, ..., c_T\}$ is the mixture component sequence. Now, (13) can be rewritten as

$$R(\lambda, \lambda^{(n)}) = \sum_S \sum_C p(S, D|O, \lambda^{(n)}) \ln p(O, S, C|\lambda) \quad (14)$$
$$+ E[\ln p(\lambda|x, \theta) p(x) \lambda^{(n)}]$$

Here we expand the first part of (13).Based on (3) and (17), we can obtain:

$$R(\lambda, \lambda^{(n)}) = \quad (15)$$
$$\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{K} \phi(i,j)\left[-\frac{1}{2}(o_t - \mu_{i,j})^T \Sigma_{i,j}^{-1}(o_t - \mu_{i,j})\right]$$
$$+ \sum_{i=1}^{N}\sum_{j=1}^{K} E\left[-\frac{1}{2\sigma^2}\|\mu_{i,j} - W_{i,j}x - \bar{\mu}_{i,j}\| \,|\lambda^{(n)}\right]$$

where $\phi(i,j) = P(s_t = i, c_t = j|O, \lambda^{(n)})$ is the posterior probability of being in state i and mixture component j at time t given the observation sequence O, and $W_{i,j}$ represents the sub-matrix of W corresponding to the $\mu_{i,j}$ element After differentiating (15) with respect to $\mu_{i,j}$ and equating to zero, we find the adaptation formula

$$\mu_{i,j} = \left[\Sigma_{i,j}^{-1}\sum_{t=1}^{T}\phi(i,j) + \frac{1}{\sigma^2}I\right]^{-1} \quad (16)$$
$$\cdot\left[\Sigma_{i,j}^{-1}\sum_{t=1}^{T}\phi(i,j)o_t + \frac{1}{\sigma^2}(W_{i,j}E[x|\lambda^{(n)}] + \bar{\mu}_{i,j})\right]$$

this mean adaptation equation (16) can also be written as

$$\mu_{i,j} = \alpha\mu_{i,j,ML} + (1-\alpha)\mu_{i,j,PPCA} \quad (17)$$

where $\alpha = (\sigma_{i,j}^{-1}\Sigma_{t=1}^{T}\phi(i,j))/(\sigma_{i,j}^{-1}\Sigma_{t=1}^{T}\phi(i,j) + 1/\sigma^2)$. This tell us that the PPCA-based adaptation solution provides a unified framework which simply interpolates the ML estimate of the data, μ_{ML} with the PPCA prior estimate, μ_{PPCA} .As the amount of adaptation data increases, so does $\Sigma\phi(i,j)$, α approaches 1, and the PPCA-based solution converges to ML solution. On the other hand, for a small amount of adaptation data α becomes smaller, and the adapted mean depends more on the PPCA prior estimates.But ,in this method, considerable care must be taken in the choice of the number of factors q. An inappropriate choice can easily give misleading results.A major problem is that if the observations

can be explained sufficiently by , say, two factors, a model which attempts to indentify only a single factor may often fail to find either of the sufficient two, but may instead find a third alternative. This ultimately a result of mis-specification of q being compensated for the factor loading **W**, an effect which does not occur in the case of the proposed model for PCA. In this latter case , the use of the isotropic noise model implies that the first two principal axes will clearly include the first alone

3. EXPERIMENTS

Performance of the proposed method was evaluated with speaker-independent continuous Korean digit recognition experiments. Utterances from 105 speakers constructed the training data and those from the other 35 speakers were used for evaluation. Each speaker contributed 30~40 sentences consisting of 3~7 digits.Each digit was modeled by a seven-state left-to-right HMM without skips and two mixture components for each state and the silence was modeled by an one-state HMM. To obtain the SD HMM models, they trained first a set of speaker-independent (SI) models on the speech from all the 105 speakers and then carried out the MAP-based adaptation for each training speaker. The supervector was extracted by augmenting all the mean vectors of each SD model.Each speech frame was parameterized by a 24-dimensional feature vector consisting of 12 mel-frequency cepstral coefficients and their first-order time derivatives. The test result are showing as follow:

Table 1: Word recognition rate(%) for MAP and MLLR

No. of Sentence	2	5	10
MAP	88.61	88.77	89.74
MLLR	88.15	88.18	88.60

Table 2: Word recognition rate(%) for MLED and PPCApaper we have shown how principal component

Methods	Sent.	P(dimensions)				
		1	2	4	8	16
MLED	2	89.01	88.94	88.85	88.88	88.97
PPCA	2	89.24	88.96	89.90	88.70	88.70
MLED	5	89.55	89.55	89.35	89.28	89.26
PPCA	5	89.51	89.36	89.32	89.26	89.28
MLED	10	90.07	89.88	89.90	89.98	89.96
PPCA	10	90.09	90.12	90.20	90.21	90.21

From these tables we can see that the PPCA method performs better than the conventional techniques such as MLLR.But under the condition of only two or five sentences, PPCA method is not better than MLED method, which make it questionable that PPCA method should be more effective than other adaptation method as [1] claimed.

4. SUMMARY

This paper review the PPCA-based adaptation approach, which is applicable to adapt the mean vectors among HMM parameters. Theoretically, this method is especially useful when only a small amount of sample data presents. At the same time, the derived mean adaptation equation can be useful when we evaluate the mean vectors. Also, the defects of this method was shortly discussed.The experiment on the approach was performed and the result is acceptable, but which didn't show a great effectiveness of this new method.

REFERENCE

[1] D.K. Kim and N.S. Kim, "Bayesian Speaker Adaptation Based on Probabilistic Principal Component Analysis", Proc. ICSLP, vol. 3, pp. 734-737, Beijing, China, October 2000.

[2] I.T.Jolliffe, Principal Component Analysis. Springer-Verlag, 1986

[3] M.Tipping and C.Bishop, "Mixtures of probabilistic principal component analyzers," Neural Computation, vol. 11, pp. 435-474, 1999

[4] J.L.Gauvain and C.H. Lee, "Maximun a posteriori estimation for multivariate Gaussian mixture observation of Markov chains, " IEEE Trans. Speech Aucdio Proc. ICSLP, pp. 1771-1774, 1998

[5] R. O. Duda, P.E.Hart and D.G.Stork, Pattern Classification, Wiley-Interscience, 2001

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data viathe EM algorithm. Journal of the Royal Statistical Society series B, 39:1¨C38, 1977