BAYESIAN SPEAKER ADAPTATION BASED ON PROBABILISTIC PRINCIPAL COMPONENT ANALYSIS

DONG KOOK KIM, NAM SOO KIM

School of Electrical and Computer Engineering,
Seoul National University, Seoul, Korea.
E-mail: {dkkim11, nkim}@snu.ac.kr

ABSTRACT

In this paper, we propose a Bayesian speaker adaptation technique based on the probabilistic principal component analysis (PPCA). The PPCA is employed to obtain the canonical speaker models which provide the *a priori* knowledge of the training speakers. The proposed approach is conveniently incorporated into the Bayesian adaptation framework where the parameters are adapted to the new speaker's speech according to the maximum *a posteriori* (MAP) criterion. Through a number of continuous digit recognition experiments, we can find the effectiveness of the PPCA-based approach compared to the other adaptation approaches with a small amount of adaptation data.

1. INTRODUCTION

Many adaptation techniques have been studied to reduce the acoustic mismatches between the training and test conditions of an automatic speech recognizer [1]. 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 [2]. To estimate the dependencies between diverse units of speech, a large corpus of training speakers and a variety of correlation modeling approaches are used [2]. In general, the basic adaptation techniques are classified into three categories: the maximum a posteriori (MAP) adaptation [3], parameter transformation based adaptation using maximum likelihood linear regression (MLLR) [4], and speaker clustering based adaptation approaches [5].

The eigenvoice technique which is one of the speaker clustering based adaptation methods was introduced for rapid speaker adaptation in [5]. 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 the *a priori* knowledge of the training speakers, the conventional principal component analysis (PCA) method [6] is applied to a set of supervectors provided by separate speaker dependent (SD) hidden Markov model (HMM) parameters. One of the 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. In order to alleviate this problem, the obtained eigen speaker models are used as the prior information for the MAP adaptation method [5].

In this paper, we propose a Bayesian speaker adaptation technique based on the probabilistic principal component analysis (PPCA) [7] . The PPCA method finds the canonical speaker models based on the expectation maximization (EM) algorithm [8] . The proposed approach provides not only the canonical speaker models but also the *a priori* distribution of the model parameters, which can be directly applied to the MAP adaptation scheme. For that reason, the target speaker model converges to the true SD model when a large amount of adaptation data is available. Performance of the proposed adaptation method is evaluated through a series of speaker independent continuous digit recognition experiments which shows its effectiveness compared to the other adaptation approaches.

2. EIGENVOICES

Let $\{\mathbf{m}_k, k = 1, \cdots, M\}$ be a set of M well-trained SD HMM mean vectors. Here, $\mathbf{m}_k = [\mathbf{m}_{k,0,0}^T, \cdots, \mathbf{m}_{k,i,j}^T, \cdots \mathbf{m}_{k,N,K}^T]^T$ is the supervector of dimension D constructed from the kth speaker model. Specifically, $\mathbf{m}_{k,i,j}$ represents the mean vector of jth Gaussian in the ith state of the kth speaker HMM with N and K being the number of states and mixture components for each state, respectively. The eigenvoice method tries to find the P-dimensional linear subspace (eigenspace) spanned by $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_P$ where $\mathbf{w}_l$ is the lth basis vector called the eigenvoice representing a canonical speaker model. The eigenspace is spanned by the P-dominant eigenvectors of the sample covariance matrix $\mathbf{C} = (1/M) \sum_{k=1}^{M} (\mathbf{m}_k - \boldsymbol{\mu}_\mathbf{m})(\mathbf{m}_k - \boldsymbol{\mu}_\mathbf{m})^T$ such that $\mathbf{CW} = \mathbf{\Lambda W}$, where $\mathbf{\Lambda}$ is a diagonal matrix defined by the largest eigenvalues of

$\mathbf{C}$, $\boldsymbol{\mu}_{\mathbf{m}}$ is the mean vector and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_P]$. Let $\hat{\mathbf{m}}$ be the supervector of a new speaker. Then, $\hat{\mathbf{m}}$ can be obtained by a linear combination of P-principal speaker models such that

$$\hat{\mathbf{m}} = \sum_{l=0}^{P} x_l \mathbf{w}_l = \mathbf{W}\mathbf{x} \quad (1)$$

where $\mathbf{x} = [x_1, \cdots, x_P]^T$ is the weight vector. Using the maximum-likelihood eigen-decomposition (MLED) method proposed in [5], the weight vector $\mathbf{x}$ can be found given an adaptation data O as follows:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{argmax}\ \log p(O|\mathbf{W}\mathbf{x})\ . \quad (2)$$

3. PROBABILISTIC PCA

Here we review the concept and formulations for PPCA. Let $\mathbf{y} = [y_1, y_2, \cdots, y_D]^T$ be an observation vector of dimension D. Assume that $\mathbf{y}$ is related to the latent variable $\mathbf{x} = [x_1, x_2, \cdots, x_P]^T$ of dimension $P(\ll D)$ by

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mu_{\mathbf{y}} + \epsilon \quad (3)$$

where $\mathbf{W}$ is the $D \times P$ parameter matrix that represents the principal subspace of the observation data, $\mu_{\mathbf{y}}$ is the mean vector of $\mathbf{y}$ and ϵ is a Gaussian random noise independent of $\mathbf{x}$. Conventionally, the latent variable is defined to be an independent Gaussian of unit variance such that

$$p(\mathbf{x}) = (2\pi)^{-P/2} \exp\{-\frac{1}{2}\mathbf{x}^{\mathbf{T}}\mathbf{x}\}. \quad (4)$$

The noise is also modeled by a Gaussian such that $\epsilon \sim N(0, \sigma^2\mathbf{I})$ where $\mathbf{I}$ is the $D \times D$ identity matrix. Based on the above assumptions, the observation vectors are also normally distributed according to

$$\begin{aligned} p(\mathbf{y}) = {} & (2\pi)^{-D/2}|\Sigma_{\mathbf{y}}|^{-1/2} \\ & \cdot \exp\{-\frac{1}{2}(\mathbf{y}-\mu_{\mathbf{y}})^T {\Sigma_{\mathbf{y}}}^{-1}(\mathbf{y}-\mu_{\mathbf{y}})\} \end{aligned} \quad (5)$$

where $\Sigma_{\mathbf{y}} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T$. We can derive the conditional probability distribution of $\mathbf{y}$ given $\mathbf{x}$ by

$$p(\mathbf{y}|\mathbf{x}) = (2\pi\sigma^2)^{-D/2} \exp\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{W}\mathbf{x} - \mu_{\mathbf{y}}||^2\}. \quad (6)$$

Given an observation sequence $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T\}$, the PPCA estimates the latent variable sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ and finds the optimal model parameters $\hat{\lambda} = \{\hat{\mathbf{W}}, \hat{\mu}_{\mathbf{y}}, \hat{\sigma}^2\}$ according to the maximum likelihood (ML) criterion such that

$$\hat{\lambda} = \underset{\lambda}{argmax}\ [\log p(\mathbf{Y}|\lambda)]\ . \quad (7)$$

Since, however, the latent variables $\{\mathbf{x_t}\}$ are considered to be hidden, it becomes highly difficult to solve (7). For that reason, the EM algorithm which iteratively updates the parameter values is applied. Let $\lambda^{(n)} = \{\mathbf{W}^{(n)}, {\mu_{\mathbf{y}}}^{(n)}, \sigma^{2,(n)}\}$ be the parameter values obtained in the nth iteration. Then, the new parameter values $\lambda^{(n+1)} = \{\mathbf{W}^{(n+1)}, {\mu_{\mathbf{y}}}^{(n+1)}, \sigma^{2,(n+1)}\}$ are obtained by

$$\lambda^{(n+1)} = \underset{\lambda}{argmax}\ Q(\lambda^{(n+1)}, \lambda^{(n)}) \quad (8)$$

where

$$Q(\lambda^{(n+1)}, \lambda^{(n)}) = E\left[\log p(\mathbf{Y}, \mathbf{X}|\lambda^{(n+1)})|\mathbf{Y}, \lambda^{(n)}\right]. \quad (9)$$

After some manipulation, we are led to

$${\mu_{\mathbf{y}}}^{(n+1)} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{y}_t - \mathbf{W}^{(n)}\overline{\mathbf{x}_t}\right) \quad (10)$$

$$\mathbf{W}^{(n+1)} = \left[\sum_{t=1}^{T}(\mathbf{y}_t - \mu_{\mathbf{y}}^{(n+1)})\overline{\mathbf{x}}_t\right]\left[\sum_{t=1}^{T}\overline{\mathbf{x}_t\mathbf{x}_t^T}\right]^{-1} \quad (11)$$

$$\begin{aligned} \sigma^{2,(n+1)} = {} & \frac{1}{DT}\sum_{t=1}^{T}\Big\{||\mathbf{y}_t - \mu_{\mathbf{y}}^{(n+1)}||^2 - 2\overline{\mathbf{x}_t^T}\mathbf{W}^{T,(n+1)} \\ & \cdot\left(\mathbf{y}_t - \mu_{\mathbf{y}}^{(n+1)}\right) + tr\left(\overline{\mathbf{x}_t\mathbf{x}_t^T}\mathbf{W}^{T,(n+1)}\mathbf{W}^{(n+1)}\right)\Big\} \end{aligned} \quad (12)$$

where

$$\overline{\mathbf{x}_t} \equiv E\left[\mathbf{x}_t|\mathbf{y}_t, \lambda^{(n)}\right] = \Sigma_{\mathbf{x}}^{-1}\mathbf{W}^T(\mathbf{y}_t - \mu_{\mathbf{y}}) \quad (13)$$

$$\overline{\mathbf{x}_t\mathbf{x}_t^T} \equiv E\left[\mathbf{x}_t\mathbf{x}_t^T|\mathbf{y}_t, \lambda^{(n)}\right] = \sigma^2\Sigma_{\mathbf{x}}^{-1} + \overline{\mathbf{x}_t}\cdot\overline{\mathbf{x}}_t{}^T \quad (14)$$

with $\Sigma_{\mathbf{x}} = \sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W}$ and tr representing the trace of a matrix. The log-likelihood in the right hand side of (7) is maximized when the columns of $\mathbf{W}$ span the principal subspace of the observation data. It is shown in [7] that for the global maximum of the likelihood the ML estimate $\mathbf{W}_{ML}$ contain the principal eigenvectors of the covariance matrix of observation data.

4. PPCA-BASED SPEAKER ADAPTATION

If λ, which is assumed to be a random vector, is the parameter vector to be estimated from the observation O with probability density function (pdf) $f(O|\lambda)$ and its prior pdf is $g(\lambda|\theta)$, where θ is a prior parameter, then the MAP estimate is defined as the posterior mode of λ, i.e.,

$$\lambda_{MAP} = \underset{\lambda}{argmax}\ f(O|\lambda)g(\lambda|\theta)\ . \quad (15)$$

Usually the MAP estimation problem becomes too complicated for incomplete data such as the HMM because of the underlying hidden process . If the prior pdf belongs to the conjugate family of the complete-data density, the EM algorithm can be efficiently applied to MAP estimation.

Let us assume that λ is generated by a model given by (3), which has a hidden variable $\mathbf{x}$ with the prior parameter θ. Then, the complete-data likelihood for λ can be easily defined. In this case we apply the EM algorithm to iteratively increase the posterior likelihood $p(\lambda|O)$. The auxiliary function to be optimized is given as follows:

$$\begin{aligned} R(\lambda, \lambda^{(n)}) &= E\left[\log p(\mathbf{Y}|\lambda) + \log p(\lambda, \mathbf{x}|\theta)|O, \lambda^{(n)}\right] \\ &= E\left[\log p(\mathbf{Y}|\lambda)|O, \lambda^{(n)}\right] + E\left[\log p(\lambda, \mathbf{x}|\theta)|\lambda^{(n)}\right] \end{aligned} \tag{16}$$

where $\mathbf{Y}$ is the complete data for O, and $\{\lambda, \mathbf{x}\}$ means the complete data for λ, respectively. It is shown in [8] that if $R(\lambda, \lambda^{(n)}) \geq R(\lambda^{(n)}, \lambda^{(n)})$, then $p(\lambda|O) \geq p(\lambda^{(n)}|O)$.

Let $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T\}$ be a sequence of feature vectors generated by an HMM. The observation density $p(\mathbf{o}_t|i)$ for state i is assumed to be a mixture of Gaussians,

$$p(\mathbf{o}_t|i) = \sum_{j=1}^{K} \omega_{i,j} \mathcal{N}(\mathbf{o}_t;\ \boldsymbol{\mu}_{i,j}, \Sigma_{i,j}) \tag{17}$$

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

In this paper, we consider only the adaptation of the mean vectors among the HMM parameters. Let $\boldsymbol{\mu} = [\boldsymbol{\mu}_{1,1}^T, \cdots, \boldsymbol{\mu}_{N,K}^T]^T$ be a supervector that augments all the Gaussian mean vectors. Assume that $\boldsymbol{\mu}$ is generated by a PPCA model with a latent variable $\mathbf{x}$ and parameters $\theta = \{\bar{\boldsymbol{\mu}}, \mathbf{W}, \sigma^2\}$ and $\lambda^{(n)} = \{\boldsymbol{\mu}^{(n)}\}$ be the current estimate and $\lambda = \{\boldsymbol{\mu}\}$ be the new estimate. Then, the auxiliary function for the EM algorithm is defined by

$$\begin{aligned} R(\lambda, \lambda^{(n)}) = &\ E\left[\log p(\mathbf{O}, S, C|\lambda)|\mathbf{O}, \lambda^{(n)}\right] \\ &+ E\left[\log p(\lambda, \mathbf{x}|\theta)|\lambda^{(n)}\right] \end{aligned} \tag{18}$$

where $S = \{s_1, \cdots, s_T\}$ represents the state sequence, $C = \{c_1, \cdots, c_T\}$ is the mixture component sequence. Now, (18) can be rewritten as

$$\begin{aligned} R(\lambda, \lambda^{(n)}) = &\ \sum_S \sum_C p(S, C|\mathbf{O}, \lambda^{(n)}) \log p(\mathbf{O}, S, C|\lambda) \\ &+ E\left[\log p(\lambda|\mathbf{x}, \theta) p(\mathbf{x})|\lambda^{(n)}\right]. \end{aligned} \tag{19}$$

Based upon (6) and (17), it is not difficult to derive

$$\begin{aligned} &R(\lambda, \lambda^{(n)}) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_t(i,j) \left[-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{i,j})^T \Sigma_{i,j}^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_{i,j})\right] \\ &\quad + \sum_{i=1}^{N} \sum_{j=1}^{K} E\left[-\frac{1}{2\sigma^2} || \boldsymbol{\mu}_{i,j} - \mathbf{W}_{i,j}\mathbf{x} - \bar{\boldsymbol{\mu}}_{i,j}||^2 \mid \lambda^{(n)}\right] \end{aligned} \tag{20}$$

where $\gamma_t(i,j) = P(s_t = i, c_t = j|\mathbf{O}, \lambda^{(n)})$ is the posterior probability of being in state i and mixture component j at time t given the observation sequence $\mathbf{O}$, and $\mathbf{W}_{i,j}$ represents the sub-matrix of $\mathbf{W}$ corresponding to the $\boldsymbol{\mu}_{i,j}$ element. After differentiating (20) with respect to $\boldsymbol{\mu}_{i,j}$ and equating to zero, we find the adaptation formula

$$\begin{aligned} \boldsymbol{\mu}_{i,j} = &\ \left[\Sigma_{i,j}^{-1} \sum_{t=1}^{T} \gamma_t(i,j) + \frac{1}{\sigma^2}\mathbf{I}\right]^{-1} \\ &\cdot \left[\Sigma_{i,j}^{-1} \sum_{t=1}^{T} \gamma_t(i,j)\mathbf{o}_t + \frac{1}{\sigma^2}(\mathbf{W}_{i,j} E\left[\mathbf{x}|\lambda^{(n)}\right] + \bar{\boldsymbol{\mu}}_{i,j})\right] \end{aligned} \tag{21}$$

where

$$E\left[\mathbf{x}|\lambda^{(n)}\right] = (\sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T(\boldsymbol{\mu}^{(n)} - \bar{\boldsymbol{\mu}}) \tag{22}$$

which is equivalent to (13) . Focusing on scalar observations for ease of discussion, the mean adaptation equation (22) can be written as

$$\mu_{i,j} = \alpha \mu_{i,j,ML} + (1-\alpha)\mu_{i,j,PPCA} \tag{23}$$

where $\alpha = (\sigma_{i,j}^{-1} \sum_{t=1}^{T} \gamma_t(i,j))/(\sigma_{i,j}^{-1} \sum_{t=1}^{T} \gamma_t(i,j) + (1/\sigma^2))$. This tells us that the PPCA-based adaptation solution provides a unified framework which simply interpolates the ML estimate of the adaptation data, μ_{ML} with the PPCA prior estimate, μ_{PPCA}. As the amount of adaptation data increases, so does $\sum \gamma_t(i,j)$, α approaches 1, and the PPCA-based solution converges to the 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 estimate.

5. 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, we 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. We extracted a supervector by augmenting all the mean vectors of each SD model. The order in which the mixture Gaussian mean vectors are arranged in the supervector is automatically determined based on the MAP-based adaptation. The speech signal was sampled at 8 kHz and segmented into 30 ms frame at every 10 ms with 20 ms overlap. Each speech frame was parameterized by a 24-dimensional feature vector consisting of 12 mel-frequency cepstral coefficients and their first-order time derivatives. Consequently, each supervector contained $D = 3840 (= 11 \times 7 \times 2 \times 24 + 3 \times 2 \times 24)$ parameters. Using the 105 supervectors constructed from the training data, we obtained the estimates for the PPCA model parameters.

In the recognition experiments, we drew 2, 5 and 10 sentences (2$\sim$20 sec) from each target speaker for adaptation, and performed the recognition test on the remaining sentences. All adaptation procedures were performed in a static supervised manner using the transcripted adaptation data. The SI system gives 88.03 % of word recognition rate. For the purpose of comparison, we carried out two conventional mean adaptation techniques, MLLR and MAP, whose word recognition rate are shown in Table I. Table II shows the recognition results for both the MLED where the MLED is used as a prior for the MAP and the proposed method with varying dimension P. From Table II we can see that both the MLED and PPCA techniques perform much better than the conventional techniques. For two and five adaptation sentences, we can observe that the performance of the PPCA-based adaptation was similar to that of MLED adaptation. It is noteworthy that for very sparse adaptation data the word recognition rate does not increase as the dimension P increases. As for the 10 adaptation sentences, the recognition rate of the PPCA-based approach was slightly better compared to the MLED method.

Table I
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 II
Word recognition rate (%) for MLED and PPCA with various P

Methods	Sent.	P				
		1	2	4	8	16
MLED	2	89.01	88.94	88.85	88.88	88.87
PPCA	2	89.24	88.96	88.90	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

6. CONCLUSIONS

We have proposed a Bayesian speaker adaptation approach for speech recognition based on the PPCA. The PPCA was used for finding the canonical speaker models from the SD HMM's based on the EM algorithm. We derived the adaptation formula based on the PPCA model which combines the prior knowledge with the adaptation data from a new speaker according to the MAP criterion. From the results of a number of continuous digit recognition experiments, we could see that the proposed approach performed well especially when a small amount of adaptation data was available.

7. REFERENCES

[1] P. C. Woodland, "Speaker adaptation: techniques and challenges," *ASRU Workshop*, vol. 1, pp. 85-90, 1999.

[2] V. Digalakis, et al., "Rapid speech recognizer adaptation to new speakers," *Proc. ICASSP*, pp. 765-768, 1999.

[3] J. L. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 291-298, Apr. 1994.

[4] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171-185, 1995.

[5] R. Kuhn, et al., "Eigenvoices for speaker adaptation," *Proc. ICSLP*, pp. 1771-1774, 1998.

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

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

[8] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38, 1977.