EMAP-BASED SPEAKER ADAPTATION WITH ROBUST CORRELATION ESTIMATION

Eugene Jon, Dong Kook Kim, and Nam Soo Kim

School of Electrical and Computer Engineering, Seoul National University, Seoul, Korea

ABSTRACT

In this paper, we propose a method to enhance the performance of the extended maximum *a posteriori* (EMAP) estimation using the probabilistic principal component analysis (PPCA). PPCA is used to robustly estimate the correlation matrix among separate hidden Markov model (HMM) parameters. The correlation matrix is then applied to the EMAP scheme for speaker adaptation. PPCA is efficient to compute, and shows better performance compared to the method previously used for EMAP. Through various experiments on continuous digit recognition, it is shown that the EMAP approach based on the PPCA gives enhanced performance especially for a small amount of adaptation data.

1. INTRODUCTION

Due to recent advances, speaker independent (SI) continuous speech recognition systems show improved performance. But the performance still degrades for those speakers who are not covered by the training data. Various speaker adaptation methods have been studied to reduce the performance gap between the SI and speaker dependent (SD) systems [1]. One of such methods is the maximum *a posteriori* (MAP) method [2] in which the a priori knowledge concerned with the recognition parameters is used. An important advantage of the MAP approach is that the adapted parameters converge to the SD models when the adaptation data grows larger. However, the MAP approach transforms only the observed parameters, which makes MAP unsuitable for rapid speaker adaptation. For that reason the extended MAP (EMAP) [3] approach was proposed to enhance the performance of the MAP-based method for sparse adaptation data. EMAP uses the correlation among parameters to adapt unobserved parameters.

Principal component analysis (PCA) [4] is a method used in numerous statistical applications to reduce the dimensionality of a data set, while retaining the inherent variation. This property of PCA is useful for rapid speaker adaptation, since it can fully utilize a small amount of adaptation data. With the incorporation of probability density models, PCA becomes the probabilistic PCA (PPCA) [5] where the parameters can be easily obtained based on the given data set.

In this paper, we propose a rapid speaker adaptation method which uses PPCA to compute the correlation matrix for the EMAP approach. The performance of the proposed approach is compared to the conventional EMAP algorithm on various experiments, and gives better results.

2. OVERVIEW OF THE PROPOSED METHOD

The proposed adaptation approach is described as follows: First, we use the data of each individual speaker to obtain the PPCA parameters. This computation is done before the adaptation process. Using the PPCA parameters we can compute the correlation matrix needed for the EMAP adaptation. The correlation matrix and the adaptation speech data are used for EMAP adaptation. Finally, using the adapted model we run the speech recognition tests. The block diagram of the proposed speaker adaptation method is shown in Figure 1.

3. MAP ADAPTATION

The MAP estimation is a scheme that utilizes the prior information of the parameters to be estimated [6]. If $\mathbf{O} = \{\mathbf{o_1}, \ldots, \mathbf{o_L}\}$ is a sequence of observations with a pdf $\mathbf{P}(O)$ where L is the total number of frames and λ is the parameter set defining the distribution, given a sequence of training data $\mathbf{O}$, λ is to be estimated. If λ is assumed to be fixed but unknown, the ML estimate for λ is found by solving

$$\frac{\delta}{\delta\lambda} P(\mathbf{O}|\lambda) = \mathbf{0}. \tag{1}$$

However, if λ is assumed random with a *a priori* distribution function $\mathbf{P}_o(\lambda)$, then the MAP estimate for λ is found by solving

$$\frac{\delta}{\delta\lambda} P(\lambda|\mathbf{O}) = \mathbf{0}. \tag{2}$$

Using Bayes theorem,

$$P(\lambda|\mathbf{O}) = \frac{\mathrm{P}(\mathbf{O}|\lambda)\mathbf{P_o}(\lambda)}{\mathrm{P}(\mathbf{O})}. \tag{3}$$

Compared to ML estimation, the MAP estimation procedure involves a prior distribution function $\mathbf{P}_o(\lambda)$ for the random

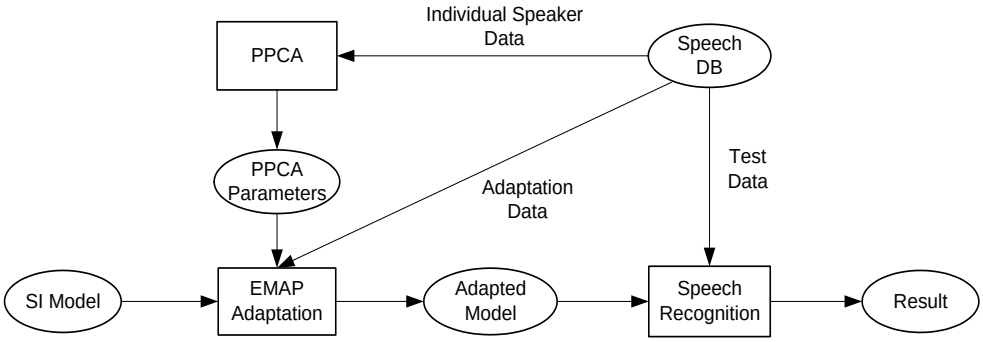

Fig. 1. Speaker Adaptation Procedure

parameter λ. It can also be seen that MAP estimation is more robust for sparse adaptation data for this reason.

For the hidden Markov model (HMM) parameters the estimate of the mean is obtained by the following equation

$$\tilde{\mathbf{m}}_k = \frac{\tau_k \mu_k + \Sigma_{l=1}^{L} c_{kl}\mathbf{o}_l}{\tau_k + \Sigma_{l=1}^{L} c_{kl}}, \tag{4}$$

where k is the index of the Gaussian, μ_k is the prior mean of the kth Gaussian, $\mathbf{o}$ is the observation vector of the lth frame and c_{kl} is the count of kth Gaussian among the lth frame of the observation data. We can observe that the mean is a interpolation of the observation data and the prior data. We can change the importance of the prior data by adjusting τ_k. It can be seen from (4) that only the observed parameters are adapted. For the unobserved parameters c_{kl} is zero and the estimated mean becomes the prior mean.

4. EXTENDED MAP

Speaker adaptation using the MAP approach has the disadvantage of transforming only those parameters that are observed. Typical speech recognition systems based on the HMM use millions of parameters, which makes rapid adaptation through MAP practically impossible. The EMAP approach was developed in order to improve the MAP adaptation method particularly for fast speaker adaptation. The EMAP method assumes that all the Gaussian distributions are correlated, and uses the correlation information to transform the unobserved parameters.

Let $\mathbf{m} = [\mathbf{m}_1^T, \ldots, \mathbf{m}_K^T]^T$ be an augmented column vector, in which $\mathbf{m}_j$ represents the mean vector of the jth Gaussian distribution with K being the total number of Gaussians and T denoting the matrix transposition. We call this augmented vector $\mathbf{m}$, as the supervector since it consists of all the Gaussian mean vectors. It is assumed that the individual mean vectors are correlated such that the *a priori* probability density function (pdf) of $\mathbf{m}$ is given by

$$g(\mathbf{m}) = N(\mathbf{m}_0, \mathbf{S}_0) \tag{5}$$

where $N(a, b)$ represents the normal distribution with mean a and covariance b. In general, $\mathbf{m}_0$ is obtained through HMM training where the relevant parameters are estimated based on a set of training data according to the ML criterion. This process is explained extensively in [3] and only the results will be shown here.

For simplicity, it is assumed that only a single observation sequence $\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_L$ is used for adaptation, where L represents the total number of frames. If the prior pdf for the Gaussian mean vectors are given by (5), it could be shown that according to the MAP criterion, the adapted supervector, $\hat{\mathbf{m}}_0$ is expressed as follows [3]:

$$\hat{\mathbf{m}}_0 = \mathbf{S}(\mathbf{S} + \mathbf{S}_0\mathbf{C})^{-1}\mathbf{m}_0 + \mathbf{S}_0(\mathbf{S} + \mathbf{C}\mathbf{S}_0)^{-1}\mathbf{C}\mathbf{A} \tag{6}$$

where $\mathbf{S} = diag(\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_K)$ with $\mathbf{\Sigma}_j$ being the covariance matrix of the jth Gaussian and $\mathbf{C} = diag(c_1, \cdots, c_K)$ in which $c_k = \sum_{l=1}^{L} c_{kl}$ is the count of the kth Gaussian observed among the adaptation data used. Furthermore, $\mathbf{A} = ((\mathbf{m}_1)_{ML}^T, \cdots, (\mathbf{m}_K)_{ML}^T)^T$ where $(\mathbf{m}_k)_{ML} = \frac{\sum_{l=1}^{L} c_{kl}\mathbf{o}_l}{c_k}$. The right hand side of (6) can be interpreted as a linear interpolation between the prior knowledge and the given data. For effective computation, we express the mean shift $\hat{\mathbf{m}}_0 - \mathbf{m}_0$ as a function of the ML mean shift $\mathbf{A} - \mathbf{m}_0$. Manipulating (6) we can obtain

$$\hat{\mathbf{m}}_0 - \mathbf{m}_0 = \mathbf{S}_0(\mathbf{S} + \mathbf{C}\mathbf{S}_0)^{-1}\mathbf{C}(\mathbf{A} - \mathbf{m}_0). \tag{7}$$

Conventionally, the correlation matrix $\mathbf{S}_0$ in (7) is obtained using the SD model parameters as follows:

$$\mathbf{S}_0 = \frac{1}{N_{sp}} \sum_{i=1}^{N_{sp}} (\mathbf{m}_i - \mathbf{m}_{SI})(\mathbf{m}_i - \mathbf{m}_{SI})^T, \tag{8}$$

where $\mathbf{m}_i$ is the supervector of the ith speaker with N_{sp} being the total number of training speakers and $\mathbf{m}_{SI}$ represents the SI supervector consisting of all the SI Gaussian mean vectors.

5. PROBABILISTIC PCA

PCA is a well established technique for dimension reduction [4]. The main idea of PCA is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining most of the variation. PPCA was introduced by incorporating the concept of probability densities in the PCA method. The most significant advantage of PPCA over PCA is that the single PCA model can be extended to a mixture of such models, thus allowing nonlinear projection of the data.

Here we review the concept and formulations for the 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

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mu_{\mathbf{y}} + \epsilon \tag{9}$$

where $\mathbf{W}$ is a $D \times P$ 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

$$p(\mathbf{x}) = (2\pi)^{-P/2} \exp\{-\frac{1}{2}\mathbf{x}^T\mathbf{x}\}. \tag{10}$$

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 vector is also normally distributed according to

$$\begin{aligned} p(\mathbf{y}) = {} & (2\pi)^{-D/2}|\mathbf{\Sigma_y}|^{-1/2} \\ & \cdot \exp\{-\frac{1}{2}(\mathbf{y}-\mu_{\mathbf{y}})^T\mathbf{\Sigma_y}^{-1}(\mathbf{y}-\mu_{\mathbf{y}})\} \end{aligned} \tag{11}$$

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

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) = {} & (2\pi\sigma^2)^{-D/2} \\ & \cdot \exp\{-\frac{1}{2\sigma^2}||\mathbf{y}-\mathbf{W}\mathbf{x}-\mu_{\mathbf{y}}||^2\}. \end{aligned} \tag{12}$$

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

$$\hat{\lambda} = \underset{\lambda}{\text{argmax}} \ [\log p(\mathbf{Y}|\lambda)] \ . \tag{13}$$

However, since the latent variables $\{\mathbf{x}_t\}$ are considered to be hidden, it becomes extremely difficult to solve (13) directly. For that reason, the expectation maximization (EM) algorithm is applied, which iteratively updates the parameter values [5].

6. PROBABILISTIC PCA USED IN EXTENDED MAP

Let the SD supervectors, $\mathbf{m}_1, \cdots, \mathbf{m}_{N_{sp}}$ shown in (8) be considered as an observation sequence in the PPCA problem. Based on the latent variable model given in (9), we can derive the correlation matrix from the parameters, W and σ^2 computed in the previous section such that

$$\mathbf{S}_0 = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T. \tag{14}$$

Substituting $\mathbf{S}_0$ in (7) with the right hand side of (14) will yield

$$\begin{aligned} \hat{\mathbf{m}}_0 - \mathbf{m}_0 = {} & (\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T) \\ & \cdot (\mathbf{S} + \mathbf{C}(\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T))^{-1}\mathbf{C}(\mathbf{A} - \mathbf{m}_0). \end{aligned} \tag{15}$$

We can compute the above equation in a more efficient way using the matrix inversion lemma. First we modify the term inside the inversion bracket in (15).

$$(\mathbf{S} + \mathbf{C}(\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T))^{-1} = (\mathbf{D} + \mathbf{W}_c\mathbf{W}^T)^{-1} \tag{16}$$

where $\mathbf{D} = \mathbf{S} + \sigma^2\mathbf{C}\mathbf{I}$ and $\mathbf{W}_c = \mathbf{C}\mathbf{W}$. Rearranging with the matrix inversion lemma,

$$\begin{aligned} (\mathbf{D} + \mathbf{W}_c\mathbf{W}^T)^{-1} = {} & \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{W}_c \\ & \cdot (\mathbf{I} + \mathbf{W}_c^T\mathbf{D}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{D}^{-1}. \end{aligned} \tag{17}$$

Using this method, we can expect two advantages compared to the conventional EMAP approach. Since PPCA compresses the variation of the data to a smaller dimension, it is suitable for adapting with sparse data. Also, the proposed method can reduce the computational load. When computing the adapted mean, the inversion part compromises most of the computation load. In (17) a matrix of dimension $P \times P$ should be inverted, which is quite smaller compared to the inversion of the matrix of dimension $D \times D$ in (7). Therefore we can expect a substantial computation reduction.

7. 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.

The speech signal was sampled at 8 kHz and segmented into 30 ms frames at 10 ms intervals with 20 ms overlaps. Each speech frame was parameterized by a 24-dimensional feature vector consisting of 12 mel-frequency cepstral coefficients and their first-order time derivatives. 11 digits were characterized by 7-state HMM's, 3 silence HMM's with a

Table 1. Word Recognition Rate (%) for SI, MAP, EMAP and EMAP based on PPCA(P).

Method	SI	MAP	EMAP	PPCA(3)	PPCA(5)	PPCA(10)	PPCA(20)	PPCA(25)	PPCA(30)
2 sent		88.6	89.0	89.7	90.0	90.0	90.1	90.3	90.2
5 sent	87.5	88.8	89.8	90.5	90.6	90.6	90.9	91.0	90.9
10 sent		89.7	91.6	91.0	90.9	91.5	91.8	91.8	91.9

single state were used, and there were 2 Gaussian mixture components for each state. In order to implement the EMAP approach, the Gaussian means were augmented into a single supervector. The supervectors were separately constructed for the cepstrum and delta-cepstrum. Therefore, the dimension of a supervector was $(11 \times 7 + 3) \times 2 \times 12 = 1920$.

In the recognition experiments, we drew 2, 5 and 10 sentences from each target speaker for adaptation data, and 30 sentences were used for recognition tests. The adaptation modes were supervised and static(batch). Dimensions for the latent variables of the PPCA were set at 3, 5, 10, 20, 25, 30. The speech data used to compute the PPCA parameters was the individual speech data of the 105 speakers which were used as training data to compute the SI model.

Results were compared with those of the SI, MAP adapted and EMAP adapted systems. These results are shown in Table 1 where P represents the dimension of the latent variable used for PPCA. Figure 2 shows that the proposed method gives enhanced performance compared to the conventional EMAP method especially for small amounts of adaptation data. For 2 adaptation sentences the proposed method shows 6.4% ($P = 3$) to 11.8% ($P = 25$) of reduction in word error rate (WER) compared to the original EMAP scheme. As P grows larger the computation load increases accordingly. However, the performance deteriorates after P reaches an optimum value because the proposed system converges to the original EMAP method. Therefore, a value of P that can optimize the performance of the system needs to be found. In this experiment $P = 25$ is sufficient for small amounts of adaptation data.

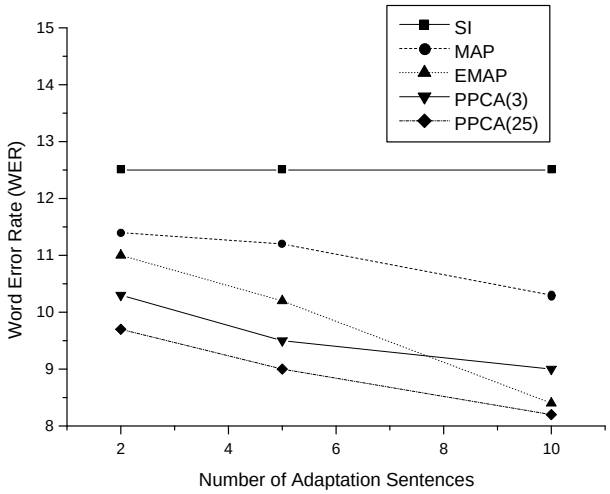

Fig. 2. Word Error Rate for Various Methods

8. CONCLUSIONS

We have proposed a novel approach to enhance the EMAP-based speaker adaptation method by means of the PPCA technique. Using the PPCA, we obtain a robust correlation matrix which will be applied to the EMAP approach. The proposed method leads to enhanced speaker adaptation performance, especially for small amounts of adaptation data and the computational load is somewhat reduced. For these reasons, we can conclude that the proposed method is suitable for rapid speaker adaptation.

9. REFERENCES

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

[2] 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.

[3] G. Zavaliagkos, "Maximum a Posteriori Adaptation Techniques for Speech Recognition," PH.D thesis,Northeastern University, Oct 1995.

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

[5] M. E. Tipping and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analysers," *Technical Report* NCRG/97/003, July 1998.

[6] S. M. Ahadi-Sarkani, "Bayesian and Predictive Techniques for Speaker Adaptation", *Ph.D. thesis*, University of Cambridge, Jan. 1996