WEIGHTED PRINCIPAL COMPONENT MLLR FOR SPEAKER ADAPTATION

Sam-Joo Doh and Richard M. Stern

Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
{sjdoh, rms}@cs.cmu.edu

ABSTRACT

We present and describe two new speaker adaptation methods which apply principal component analysis to maximum likelihood linear regression (MLLR). If we apply MLLR after transforming the baseline mean vectors by their eigenvectors, the contributions of these eigenvalues to the variance of the estimates for the MLLR matrix are inversely proportional to their corresponding eigenvalues. In the first new technique, called Principal Component MLLR (PC-MLLR), we reduce estimator variance (at the expense of increased bias) by eliminating the contributions of principal components corresponding to smaller eigenvalues. The second technique, called Weighted Principal Component MLLR (WPC-MLLR) makes use of the contributions of all principal components, but weights them according to the inverse of their putative variance. In experiments using sentences from Spokes 0 and 3 from the 1994 DARPA Wall Street Journal evaluation, the use of WPC-MLLR provided a relative reduction in word error rates of 15.1% for non-native speakers and 6.0% for native speakers compared to conventional MLLR.

1. INTRODUCTION

It is well known that any mismatch between training data and test data results in a degradation in recognition accuracy in an automatic speech recognition system. In this study we seek to reduce this mismatch by developing modified statistical models of vocal productions that better match the characteristics of new speakers.

The new procedures described in this paper are extensions of the widely-used adaptation procedure maximum likelihood linear regression (MLLR) (*e.g.* [6]). In MLLR, the mean of the observed feature vector for a given new speaker (or environmental condition) is assumed to be related to the mean of the unadapted baseline vector by an unknown linear transformation. In our work we assume that the individual components of the feature vectors can be modeled by Gaussian mixture probability density functions.

For example, consider $\hat{\mu}_m$, the mean of the m^{th} mixture density for one of the cepstral features from a given new speaker (or environmental condition). $\hat{\mu}_m$ is assumed to be related to the corresponding baseline mean vector μ_m by the linear transformation.

$$\hat{\mu}_m = A\mu_m + b \tag{1}$$

In the equation above, $\hat{\mu}_m$, μ_m, and b are D x 1 vectors and $\boldsymbol{A}$ is a D x D matrix, where D is the number of components in the feature vector (39 in our case).

In MLLR we estimate the matrix A and the vector b from adaptation data, and then update the mean vectors using Eq. (1). If the amount of adaptation data is small then the estimates of A and b may not be reliable (or have large variances). In this case, even though the estimates are obtained from adaptation data, they may not accurately describe the statistics of the test data, resulting in low recognition accuracy. The goal of this work is to obtain more reliable estimates of $\hat{\mu}_m$ by estimating A and b both by basing the estimations on a smaller number of carefully-chosen feature parameters, and by weighting the contributions of the parameters to place greater emphasis on those that are likely to be more reliable.

Let's consider an arbitrary r^{th} component $\hat{\mu}_{mr}$ of a new mean vector $\hat{\mu}_m$ and the corresponding row vector a_r^{T} of the matrix A. We can see that each $\hat{\mu}_{mr}$ depends on all the components of the corresponding base mean vector μ_m when A is a full matrix,

$$\hat{\mu}_{mr} = a_r^{\mathrm{T}}\mu_m + b_r = \begin{bmatrix} a_{r1} & a_{r2} & \cdots & a_{rD} \end{bmatrix} \begin{bmatrix} \mu_{m1} \\ \mu_{m2} \\ \cdot \\ \mu_{mD} \end{bmatrix} + b_r \tag{2}$$

It is interesting to note that while we typically assume that each component of a mean vector is independent in a speech recognition system, we normally obtain better speech recognition accuracy when A is full (compared to when it is diagonal) [5].

In Eq. (2) some components of a baseline mean vector μ_m may be more important than others in estimating $\hat{\mu}_{mr}$. If we ignore less important terms in the estimation, we can reduce the number of parameters to be estimated, and obtain more reliable estimates when we have a small amount of available adaptation data.

Gales *et al.* [1] constrained the transformation matrix to a block diagonal structure for this purpose, with feature components assumed to have correlation only within a block. Gales *et al.* used three blocks consisting of the static, delta, and delta-delta feature components. However, the block diagonal matrices did not provide better recognition accuracy than the full matrix in their test. This is because they may have enough adaptation data to estimate the full matrix, or because blocks may not be optimal.

We can use principal component analysis (PCA) [3] to reduce the dimensionality of the data. The original data which consisted of interrelated variables are transformed into a new set of uncorrelated variables by the eigenvectors of the covariance matrix of the original data set. Nouza [8] used PCA for feature selection in a speech recognition system. Kuhn *et al.* [4] introduced "eigenvoices" to represent the prior knowledge of speaker variation. Hu [2] applied PCA to describe the correlation between phoneme classes for speaker normalization. While the general motivation for these approaches was similar to the approach described in this paper, none of them are directly related to MLLR.

In this paper we apply PCA to the MLLR framework to reduce the variance of the estimates of matrix A. We will first review classical PCA and its application to MLLR. We will then describe a refinement to the method which we refer to as weighted principal component MLLR. Finally we compare speaker adaptation performance obtained using the methods described.

2. PRINCIPAL COMPONENT REGRESSION

Principal component regression was developed in classical linear regression theory (*e.g.* [3, 7]). For example, consider n pairs of samples (x_i, y_i), where x_i is a D-dimensional row vector, y_i and ε_i are scalars. ε_i is assumed to have a Gaussian distribution with zero mean and variance σ_ε^2. We assume x_i and y_i are related by a linear regression.

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ c_D \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

$$y = X \cdot \boldsymbol{c} + \varepsilon$$

Letting V_X be the orthonormal matrix whose columns are the eigenvectors of the correlation matrix $X^{\mathrm{T}}X$, and Λ_X be the diagonal matrix consisting of the corresponding eigenvalues λ_{Xj},

$$(X^{\mathrm{T}}X)V_X = V_X\Lambda_X$$

and defining the new variables $Z = XV_X$ and $\gamma = V_X^{\mathrm{T}}\boldsymbol{c}$, we obtain

$$y = X \cdot \boldsymbol{c} + \varepsilon = XV_X \cdot V_X^{\mathrm{T}}\boldsymbol{c} + \varepsilon = Z \cdot \gamma + \varepsilon \tag{3}$$

The estimate for γ is then

$$\hat{\gamma} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y = \Lambda_X^{-1}Z^{\mathrm{T}}y \tag{4}$$

It has been shown that the variance of the individual components of the estimate $\hat{\gamma}$ is inversely proportional to the eigenvalues of $X^{\mathrm{T}}X$ [7].

Therefore, components of $\hat{\gamma}$ which are associated with small eigenvalues have large variances. If we ignore those components and choose $p < D$ principal components we can obtain a substantial reduction in variance. The resulting equation becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ z_{21} & \cdots & z_{2p} \\ \cdot & & \cdot \\ \cdot & \cdots & \cdot \\ z_{n1} & \cdots & z_{np} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \cdot \\ \cdot \\ \gamma_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \tag{5}$$

We now have p parameters to estimate rather than D parameters. The best value for p will depend on the eigenvalues. If the several largest eigenvalues are much larger than the others then p can be small. If smaller eigenvalues have relatively large values then p must be larger.

After we estimate $\hat{\gamma}$ using Eqs. (4) and (5) we can obtain $\hat{\boldsymbol{c}}$ for the original linear regression by letting

$$\hat{\boldsymbol{c}} = V_{(p)} \cdot \hat{\gamma}_{(p)}$$

The subscript (p) denotes that only the principal components corresponding to the p largest eigenvalues were used. The resulting $\hat{\boldsymbol{c}}$ will have smaller variance than that contained using conventional linear regression [7].

3. PRINCIPAL COMPONENT MLLR

The formulation for Principal Component MLLR (PC-MLLR) is very similar to the discussion of principal components above, except that we also consider $P(m, i)$, the probability of the i^{th} observation o_i being the m^{th} Gaussian, as well as the baseline variance σ_m^2 and the shift vector b.

To estimate the r^{th} row vector a_r^{T} of the matrix A and the r^{th} element b_r of the vector b in conventional MLLR, we solve the following equations with μ_m representing the m^{th} baseline mean vector, o_{ir} representing the r^{th} element of the i^{th} observation, and $w_{imr} = P(m, i) \cdot \sigma_{mr}^{-2}$ [5].

$$\sum_i \sum_m w_{imr}\mu_m\mu_m^{\mathrm{T}}a_r + \sum_i \sum_m w_{imr}\mu_m b_r = \sum_i \sum_m w_{imr}o_{ir}\mu_m \tag{6}$$

$$\sum_i \sum_m w_{imr}\mu_m^{\mathrm{T}}a_r + \sum_i \sum_m w_{imr}b_r = \sum_i \sum_m w_{imr}o_{ir} \tag{7}$$

Eq. (7) can be simplified by defining the variables o_r and μ_r according to the equation

$$b_r = \frac{\sum_i \sum_m w_{imr}o_{ir}}{\sum_i \sum_m w_{imr}} - \frac{\sum_i \sum_m w_{imr}\mu_m^{\mathrm{T}}}{\sum_i \sum_m w_{imr}} \cdot a_r \equiv o_r - \mu_r^{\mathrm{T}} \cdot a_r \tag{8}$$

Let $\bar{o}$ be a column vector consisting of the elements o_r and let $\bar{M}$ be the matrix whose rows are all μ_r^{T}. Let us also define the diagonal matrix W and vector o whose elements are $w_{mr} = \sum_i w_{imr}$ and $o_{mr} = \sum_i w_{imr}o_i$ respectively, along with a matrix M with rows equal to the corresponding mean vector μ_m^{T}. Substituting Eq. (8) into Eq. (6), we can rewrite Eq. (6) in matrix form:

$$M^{\mathrm{T}} \cdot W \cdot (M - \bar{M}) \cdot a_r = M^{\mathrm{T}} \cdot W \cdot (o - \bar{o})$$

Letting $M' = M - \bar{M}$ and $o' = o - \bar{o}$, we obtain

$$M^{\mathrm{T}} \cdot W \cdot M' \cdot a_r = M^{\mathrm{T}} \cdot W \cdot o' \tag{9}$$

Defining the matrix V_M and diagonal matrix Λ_M to contain the eigenvectors and eigenvalues of $M'^{\mathrm{T}} \cdot W \cdot M'$

$$(M'^{\mathrm{T}} \cdot W \cdot M')V_M = V_M\Lambda_M \tag{10}$$

and defining the variables $\Omega = M'V_M$ and $\alpha_r = V_M^{\mathrm{T}} a_r$ in a similar fashion as in the previous section, we can write Eq. (9) as

$$V_M \Lambda_M \cdot \alpha_r = V_M \Omega^{\mathrm{T}} W o'$$

The estimate $\hat{\alpha}_r$ will become

$$\hat{\alpha}_r = \Lambda_M^{-1} \Omega^{\mathrm{T}} W o' \tag{11}$$

If we ignore the effect of different probability $P(m, i)$, the variance of the components of the estimate $\hat{\alpha}_r$ is approximately inversely proportional to the eigenvalues λ_{Mj} of $M'^{\mathrm{T}} \cdot W \cdot M'$. As before, we can choose the largest $p < D$ principal components to reduce the variances of the estimates.

It should be remembered that the matrix W consists of the inverse variance σ_{mr}^{-2} as well as the probability $P(m, i)$. The variances are different for each different r^{th} component, so in principle different eigenvectors should be used for the different row vectors a_r of matrix A. Because the probabilities will change for different adaptation data, we should calculate new eigenvectors for each new speaker. In this paper, however, we use the same eigenvectors for each different row of matrix A and for all speakers, for computational simplicity. We pre-calculated the eigenvectors using all of the baseline mean vectors. Let W contain all w_{imr} values for every m and r in order on its diagonal. Gaussian mixture weights in the baseline speech recognition system are used instead of $P(m, i)$. The matrix M' is constructed in a similar fashion. Eigenvectors are calculated from these matrices, and the same eigenvectors are used for our experiments.

4. WEIGHTED PRINCIPAL COMPONENT MLLR

Because we eliminate some of the less important components in PC-MLLR, the estimate $\hat{a}$ becomes biased, which tends to reduce recognition accuracy. In this section we introduce a modification referred to as Weighted Principal Component MLLR (WPC-MLLR) in an attempt to ameliorate this problem. WPC-MLLR applies weights to the MLLR estimates to reduce their mean square error.

From Eqs. (3) and (4) in Sec. 2,

$$\hat{\gamma} = \Lambda_X^{-1} Z^{\mathrm{T}} y = \Lambda_X^{-1} Z^{\mathrm{T}} (Z\gamma + \varepsilon) = \gamma + \Lambda_X^{-1} Z^{\mathrm{T}} \varepsilon$$

Weighting the each component of $\hat{\gamma}$ by ω_j, we obtain

$$\hat{\gamma}'_j = \omega_j \hat{\gamma}_j = \omega_j \gamma_j + \omega_j \lambda_{Xj}^{-1} Z_j^{\mathrm{T}} \varepsilon, \quad j = 1, 2, \ldots, D$$

In the usual fashion, the mean square error of $\hat{\gamma}'_j$ is

$$MSE(\hat{\gamma}'_j) = E(\hat{\gamma}'_j - \gamma_j)^2 = (\omega_j - 1)^2 \gamma_j^2 + \omega_j^2 \lambda_{Xj}^{-1} \sigma_\varepsilon^2$$

and the value of ω_j that minimizes it can be obtained by solving

$$\frac{\partial}{\partial \omega_j} MSE(\hat{\gamma}'_j) = 2(\omega_j - 1)\gamma_j^2 + 2\omega_j \lambda_{Xj}^{-1} \sigma_\varepsilon^2 = 0$$

Hence, ω_j becomes

$$\omega_j = \frac{\lambda_{Xj} \gamma_j^2}{\lambda_{Xj} \gamma_j^2 + \sigma_\varepsilon^2} = \frac{\lambda_{Xj}}{\lambda_{Xj} + \sigma_\varepsilon^2 / \gamma_j^2} \tag{12}$$

The value of ω_j approaches 1 for large λ_{Xj} and approaches 0 for small λ_{Xj}. This is intuitively appealing because we would want to apply larger weights to components of $\hat{\gamma}_j$ with smaller variance, and smaller weights to components with larger variance. In this method, instead of discarding the less significant components of $\hat{\gamma}_j$, we use all components but with weighting. Unfortunately, we don't know the correct value of the parameter γ_j. We may use the estimated value $\hat{\gamma}_j$, or the average value of $\hat{\gamma}_j$ from prior experiments. The formulation for the MLLR case is similar to Eq. (12). In our experiment we use the weight $\omega_j = \lambda_{Mj} / (\lambda_{Mj} + k)$ and set a proper value k from the experiments.

The steps for the adaptation can be summarized as follows:

(1) Transform baseline mean vectors by their eigenvectors using Eq. (10) and $\Omega = M'V_M$.
(2) Estimate $\hat{\alpha}_r$ and the shift vector b using Eqs. (8) and (11).
(3) Let $\hat{\alpha}'_{rj} = \omega_j \hat{\alpha}_{rj}$ for $r, j = 1, 2, \ldots, D$
(4) Re-calculate the shift vector b_r with $\hat{\alpha}'_r$ using Eq. (8).
(5) Transform $\hat{\alpha}'_r$ to produce the multiplication matrix A using $a_r = V_M \hat{\alpha}'_r$.
(6) Adapt the baseline mean vectors by using $\hat{\mu}_m = A\mu_m + b$.

5. TEST RESULTS

We evaluated the success of PC-MLLR and WPC-MLLR using sentences from the DARPA 1994 Wall Street Journal Spoke 3 (s3-94) and Spoke 0 (s0-94) evaluations. We selected 200 sentences for the recognition test from each data set. 10 non-native speakers read 20 sentences each from the s3-94 database while 20 native speakers read 10 sentences each from the s0-94 database. We also selected 5 adaptation sentences for each speaker which were different from the test sentences, using the correct transcriptions in supervised adaptation fashion. We used one global MLLR class and a small language model weight for the s0-94 data to emphasize the effect of adaptation on the acoustic models. We used SPHINX-III as a baseline speech recognition system, which uses continuous HMMs with 6000 senones, a 39-dimensional feature vector consisting of MFCC cepstra, delta cepstra, and delta-delta cepstra, and a 5,000-word trigram language model. Table 1 sum-

Adaptation Method	s3-94 data (Non-native)	s0-94 data (Native)
Baseline (unadapted)	27.3%	21.9%
Conventional MLLR	23.7% (13.1%)	18.3% (16.4%)
PC-MLLR	20.9% (23.4%)	18.0% (17.8%)
WPC-MLLR	20.1% (26.3%)	17.2% (21.4%)

Table 1. Word error rates for selected data from the 1994 WSJ evaluation after adaptation. (Relative percentage improvement over the baseline is shown in parenthesis).

marizes the word error rates obtained using each adaptation method. WPC-MLLR provides a relative improvement of 15.1% compared to conventional MLLR for the s3-94 data and a 6.0% improvement for the s0-94 data. WPC-MLLR provides corresponding relative improvements of 3.8% and 4.4% compared to PC-MLLR.

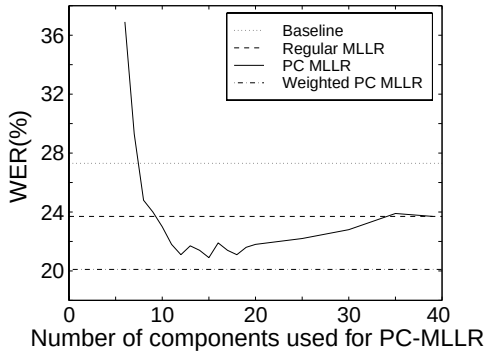

Figure 1. Word error rate for each adaptation method for s3-94 data as a function of the number of principal components used for PC-MLLR

Figure 1 depicts the word error rates (WER) for each adaptation method for s3-94 data, plotted as a function of the number of components used for PC-MLLR. As expected, the WER for PC-MLLR decreases and then increases as the number of components p increases. If p is too small, the estimates become highly biased, producing high WER. As the number of components increases to 39 (*i.e.* $p \rightarrow D$), the WER obtained with PC-MLLR increases, asymptoting to that obtained with conventional MLLR.

Figure 2 plots the ratio of sum of the p largest eigenvalues to the sum of all 39 eigenvalues. The optimum value of p will depend on the eigenvalues spread. In this experiment, we get the best recognition accuracy for PC-MLLR when p equals 15, with the ratio equal to 0.983. The ratio drops rapidly when p becomes smaller than 10, as does recognition accuracy.

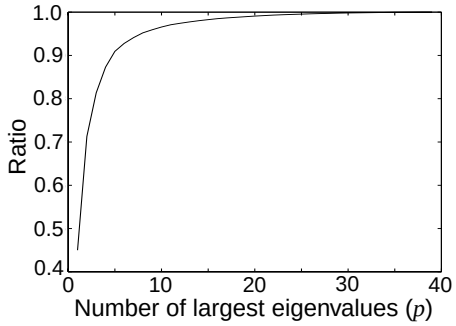

Figure 2. Ratio of the sum of the p largest eigenvalues to the sum of all 39 eigenvalues.

As noted in Sec. 3, we pre-calculated eigenvectors and use the same eigenvectors for different speakers. In other experiments using different eigenvectors based on the observation probability *P(m,i)* we observed similar recognition accuracy. This may be because the adaptation data are insufficient to estimate proper eigenvectors or because the variance of $\hat{\alpha}$ is only approximately inversely proportional to their corresponding eigenvalues, and ignores the effects of different *P(m,i)*. Even though WPC-MLLR provides less relative improvement for native speakers than it does for non-native speakers, it still is consistently better than PC-MLLR.

6. SUMMARY

In this paper, we applied principal component analysis to the MLLR framework for speaker adaptation (PC-MLLR). By eliminating highly variable components and choosing the p principal components corresponding to the largest eigenvalues we can reduce the variance of the estimates and improve speech recognition accuracy. The best value for p depends on the eigenvalues. Choosing fewer principal components increases the bias of the estimates which can reduce recognition accuracy. To compensate for this problem, we developed Weighted Principal Component MLLR (WPC-MLLR). We applied weights to the MLLR estimates so that they minimize the mean square error. In our experiments, WPC-MLLR provides relative improvements in recognition accuracy compared to conventional MLLR of 15.1% for non-native speakers (s3-94 data) and 6.0% for native speakers (s0-94 data).

ACKNOWLEGEMENT

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred. Sam-Joo Doh is partially supported by Korea Telecom Overseas Education Program.

7. REFERENCES

[1] M.J.F. Gales, D. Pye and P. C. Woodland, "Variance Compensation Within the MLLR Framework for Robust Speech Recognition and Speaker Adaptation," *Proc. of ICSLP*, p.1832-1835, 1996.

[2] Z. Hu, *Understanding and adapting to speaker variability using correlation-based principal analysis*, Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, 1999.

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

[4] R. Kuhn *et al*., "Eigenvoices for Speaker Adaptation," *Proc. of ICSLP*, p.1771-1774, 1998.

[5] C. J. Leggetter, *Improved Acoustic Modeling for HMMs Using Linear Trandformations*, Ph.D. Thesis, Cambridge Unversity, 1995.

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

[7] R. H. Myers, *Classical And Modern Regression With Applications*, PWS-KENT Publishing Company, 1990, Boston.

[8] J. Nouza, "Feature Selection Methods for Hidden Markov Model-Based Speech Recognition," *Proc. of the 13th International Conf. on Pattern Recognition*, Vol.2, p.186-190, 1996.