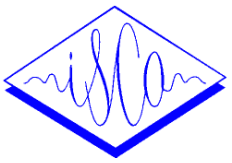

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

FAST SPEAKER ADAPTATION USING EIGENSPACE-BASED MAXIMUM LIKELIHOOD LINEAR REGRESSION

Kuan-ting Chen[1], *Wen-wei Liau*[2], *Hsin-min Wang*[1] *and Lin-shan Lee*[1,2]
[1]Institute of Information Science, Academia Sinica
[2]Graduate Institute of Communication Engineering, Taiwan University
Taipei, Taiwan, China
phone : 886-2-2788-3799 ext 1507, fax : 886-2-2782-4814
email : kenneth@iis.sinica.edu.tw

ABSTRACT

This paper presents an eigenspace-based fast speaker adaptation approach which can improve the modeling accuracy of the conventional maximum likelihood linear regression (MLLR) techniques when only very limited adaptation data is available. The proposed eigenspace-based MLLR approach was developed by introducing *a priori* knowledge analysis on the training speakers via PCA, so as to construct an eigenspace for MLLR full regression matrices as well as to derive a set of bases called eigen-matrices. The full regression matrices for each outside speaker are then constrained to be located in the space spanned by the first *K* eigen-matrices. The proposed eigenspace-based regression matrices, serving as an initial estimate of the speaker-specific MLLR transformation, effectively reduces the number of free parameters, while precise modeling for the inter-dimensional correlation among the model parameters by full matrices was maintained. Experimental results showed that for supervised adaptation using adaptation data with a length of approximately 10 seconds, the proposed approach significantly outperformed the conventional MLLR approaches.

1. INTRODUCTION

Various speaker adaptation techniques have been extensively studied in recent years due to their importance in the practical speech recognition systems. Such techniques were developed to tackle the problem of speaker mismatch between the training set and the testing set. According to [1], the popular model-based speaker adaptation techniques can be classified into three families: the maximum *a posteriori* (MAP) adaptation family, the transformation-based adaptation family including maximum likelihood linear regression (MLLR), and a family related to speaker clustering methods such as cluster adaptive training (CAT) and eigenvoice approach.

Among these techniques, MLLR approach has been widely used for fast adaptation, i.e. adaptation of the model parameters when only very limited amount of adaptation data is available. In MLLR, the speaker independent (SI) model parameters are adjusted according to one or more shared linear transformations. The transformation parameter tying mechanism based on the design of regression class tree can adequately adjust the level of regression matrix parameter sharing according to the amount and content of data and, thus, can effectively improve the robustness of parameter estimation against the sparse data problem [2][3].

In this paper, an eigenspace-based approach for improving the rapid adaptation performance of MLLR using full regression matrices is presented, and significant improvements in supervised adaptation were obtained. The remainder of the paper is organized as follows. In section 2 we present an overview of the proposed adaptation framework. In section 3 the key components of the framework are discussed. Section 4 presents some preliminary experimental results. Finally, a short conclusion is given in section 5.

2. ADAPTATION FRAMEWORK

In this paper, we focus on the adaptation of the mean parameters of Gaussian mixture components in continuous density (CD) HMMs. As in conventional MLLR, the mean vectors are adjusted based on a set of affine transformations. In this section, we first briefly review the general MLLR method, then the proposed framework for finding the transformations is introduced. Finally, since the concept of our approach is similar to that of the eigenspace approach, we briefly discuss the difference between them.

2.1. MLLR

MLLR finds the optimal affine transformation with respect to the mixture components in SI model by maximizing the likelihood of adaptation data. SI Gaussian mean parameters are clustered into *C* regression classes, and each regression class *c* is associated with an $n \times (n+1)$ regression matrix $\mathbf{W}_c$, where *n* is the dimensionality of the feature vector. Let the mean vector $\boldsymbol{m}_m = [m_m(1), ..., m_m(n)]^T$ of mixture component *m* be one of the M_c mean vectors in the regression class *c*, then the adapted mean vector can be derived as

$$\hat{\boldsymbol{m}}_m = \mathbf{W}_c \tilde{\boldsymbol{m}}_m = \mathbf{A}_c \boldsymbol{m}_m + \boldsymbol{b}_c, \quad m = 1,...,M_c; c = 1,...,C \qquad (1)$$

where $\tilde{\boldsymbol{m}}_m = [1, m_m(1), ..., m_m(n)]^T$ is the $(n+1)$-dimensional augmented mean vector. $\mathbf{A}_c$ and $\boldsymbol{b}_c$ are $n \times n$ matrix and *n*-dimensional vector respectively such that $\mathbf{W}_c = [\boldsymbol{b}_c \ \mathbf{A}_c]$.

In MLLR, each set of regression matrices $\{\mathbf{W}_c\}_{c=1,..,C}$ for a

specific speaker represents a mapping from the SI models to the speaker dependent (SD) models for that speaker and, thus, can be considered as a quantitative description of the specific speaker characteristics. The matrix $\mathbf{A}_c$ can be diagonal, block-diagonal, or full. We refer to the first two cases as diagonal matrix MLLR, and the last case as full matrix MLLR. It is clear that using full regression matrices can model the inter-dimensional correlation among the mean parameters more precisely and, thus, can provide superior description of speaker characteristics [2]. However, the large number of parameters ($C(n^2+n)$ as compared to $2nC$ in diagonal matrix case) makes robust estimation of full regression matrices very difficult when the amount of adaptation data is small.

2.2. Proposed Approach

We now give an overview the proposed adaptation framework. The adaptation involves computing a set of transformations to update the mean vectors of Gaussian mixtures. In addition to the adaptation phase, the overall procedure consists of a training phase to explore useful *a priori* knowledge.

The training phase consists of three steps. We first computed all the C conventional MLLR full regression matrices for each of the T training speakers based on the SI model parameters and the training data they provided, so as to find out all the quantitatively described speaker characteristics in the training corpus. For each training speaker, all C regression matrices are regarded as a single speaker-specific matrix set. Secondly, principal component analysis (PCA) [5] was performed on T speaker-specific regression matrix sets to extract T principal components. These principal components, called eigen-matrices here, thus represent the key information regarding the speaker characteristics as well as the inter-speaker variation for the training speakers and, thus, can be regarded as the bases of the "speaker space". Finally, K ($K \le T$) eigen-matrices were chosen to span an eigenspace, and *a priori* of the full regression matrices for each outside speaker is considered to be located in the eigenspace.

In the adaptation phase, conventional MLLR full regression matrix computation and eigenspace-based regression matrix estimation were carried out individually. A maximum likelihood (ML) estimation process was performed on the adaptation data for each testing speaker to locate the speaker-specific regression matrices in the eigenspace, yielding a set of coordinates. C eigenspace-based full regression matrices for the testing speaker were then derived from linear combination of the K eigen-matrices using these coordinates as coefficients. Since this step involves estimation of only K free parameters instead of n^2+n in the case of conventional full matrix MLLR, robust estimation can be easily made while the modeling for inter-dimensional correlation of model parameters can be maintained. Finally, a smoothing procedure was performed on the conventional MLLR full regression matrices to obtain the transformations for updating mean parameters. The overall procedure of adaptation is depicted in Figure 1.

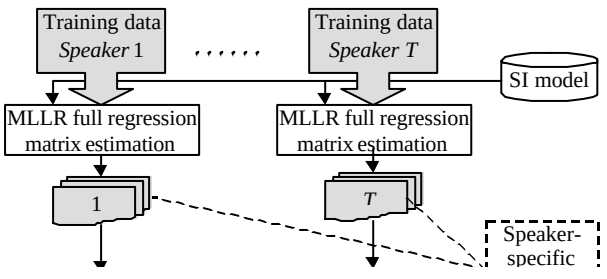

Figure 1: The overall procedure of the proposed approach

2.3. Discussion

The concept of employing an eigenspace constructed with training corpus in the proposed approach is very similar to the eigenvoice approach investigated recently [4]. However, in eigenvoice approach the eigenspace is constructed with respect to SD model parameters of the training speakers, while in our approach the regression matrices are considered instead. Therefore, eigenspace-based MLLR approach provides two advantages over the eigenvoice technique. First, our approach does not require the amount of data for each training speaker to be large enough to train robust SD models. Instead, given speaker-specific training data of size enough to robustly estimate the MLLR full regression matrices, the eigenspace-based MLLR can effectively capture the speaker characteristics and the inter-speaker variation. Second, the number of parameters in each eigenspace basis is merely $C(n^2+n)$, which is significantly smaller than that in eigenvoice approach which equals to the number of all mixture component mean parameters. This implies less on-line memory requirements, which makes the proposed approach more suitable than the eigenvoice for the practical systems.

3. IMPLEMENTATION ISSUES

3.1. Eigenspace Construction via PCA

PCA generates a set of orthonormal bases derived from the eigenvectors of the correlation matrix of the input data, and it guarantees that the mean-square error introduced by truncating the expansion after the first K eigenvectors is minimized [5]. It was used for speaker characteristic extraction and dimensionality reduction in our eigenspace-based MLLR approach.

In the proposed approach, each set of speaker-specific regression matrices were first "vectorized" to a single vector of dimension D, $D = C\ (n^2+n)$. By aligning the transpose of T speaker-specific vectors and then normalizing the elements w.r.t. each column according to sample mean and variance, a $T\times D$ matrix $\mathbf{Z}$ was obtained (usually $T<<D$). PCA requires the eigenvectors $\{\boldsymbol{e}_j\}_{j=1,..,D}$ of the $D\times D$ correlation matrix $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ to be computed, and the bases of eigenspace can then be chosen from the first K eigenvectors. However, when D is very large as the number of regression classes C increases, direct computation of $\{\boldsymbol{e}_j\}$ becomes unrealizable. Therefore, an alternative solution derived from singular value decomposition (SVD) was used

Consider the singular value decomposition of $\mathbf{Z}$

$$\mathbf{Z} = \mathbf{U}\sum\mathbf{V}^{\mathrm{T}} \quad (2)$$

where $\mathbf{U}$ is a $T\times T$ orthogonal matrix, $\mathbf{V}$ is a $D\times D$ orthogonal matrix, and Σ is a $T\times D$ matrix whose off diagonal elements are all 0's and whose diagonal elements $\{s_i\}_{i\ =\ 1,..,T}$ satisfy $s_1 \geq s_2 \geq ..\geq s_T$. In fact, the columns of $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors of $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ and $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ respectively. Since both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, from (2) we have

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{V} = \mathbf{Z}^{\mathrm{T}}\mathbf{U}\sum_{new} \quad (3)$$

where Σ_{new} has the same property as Σ. It can be observed from (3) that the first T columns of $\mathbf{V}$ can be derived from $\mathbf{Z}^{\mathrm{T}}\mathbf{U}$, which means we can compute the eigenvectors $\{\boldsymbol{u}_j\}_{\ j\ =\ 1,..,T}$ of $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$, and then obtain first T eigen-matrices as $\boldsymbol{e}_j = \mathbf{Z}^{\mathrm{T}}\ \boldsymbol{u}_j$. The computational load of $\{\boldsymbol{u}_j\}$ is reasonable for typical value of T, and this method works properly since only K (usually smaller than T) bases are required in the proposed approach.

3.2. Maximum Likelihood Coefficent Estimation

The concept of coefficient estimation process can be explained by (4), where $\boldsymbol{w}$ is the coordinate vector, O represents the observed data, and $L(\bullet)$ indicates the likelihood function. $\mathbf{E}_k^{\ c}$ represents the k-*th* eigen-matrix associated with regression class c containing Gaussian mixture component m.

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \log L\left(O \mid \boldsymbol{m}_m = (\sum_{k=1}^{K} w(k)\mathbf{E}_k^{\ c})\tilde{\boldsymbol{m}}_m\right) \quad (4)$$

It can be observed in (4) that the estimation criterion is also very similar to that of eigenvoice approach. In fact, if we first multiply the SI mean vector of each Gaussian mixture by individual eigen-matrices associated to the corresponding regression class, and concatenate all the mean vectors, then a set of new *eigenvoices* can be obtained on-line. Estimation of coordinate vector $\boldsymbol{w}$ can, then, be carried out according to exactly the same formulae as maximum likelihood eigen-decomposition (MLED) [4].

3.3. Parameter Smoothing

Let N be the total number of feature vectors, m be one of the M_c mixture components in the regression class c, and $g_m(t)$ represent the mixture occupation probability at time t for observation $O=\boldsymbol{o}_1\boldsymbol{o}_2..\boldsymbol{o}_t..\boldsymbol{o}_N$. The full regression matrices $\{\mathbf{W}_c\}_{c=1,..,C}$ derived from conventional MLLR was further smoothed by the eigenspace-based regression matrices $\{\mathbf{W}_c^{\mathrm{EIGEN}}\}_{c=1,..,C}$, as shown in (5), where t is an empirically determined parameter.

$$\hat{\mathbf{W}}_c = \frac{t\cdot\mathbf{W}_c^{\mathrm{EIGEN}} + \sum_{t=1}^{N}\sum_{m=1}^{M_c} g_m(t)\mathbf{W}_c}{t + \sum_{t=1}^{T}\sum_{m=1}^{M_c} g_m(t)} \quad (5)$$

It is worth noting that when very limited data is available, the eigenspace-based regression matrices will dominate the computation and, thus, the resultant transformations can be effectively constrained around an adequate location in the eigenspace. On the other hand, as the amount of data increases, the resultant transformations will be close to the conventional MLLR full regression matrices, which is now robust enough to model the inter-dimensional correlation among the model parameters precisely.

4. EXPERIMENTAL RESULTS

The proposed approach was evaluated on a continuous Mandarin Chinese telephone speech database provided by Telecommunication Laboratories, Taiwan, Republic of China. The database consists of 59 female and 60 male speakers, each produced 120 sentences such that a total of 14,280 sentences (5.84 hrs) are included. The speech was sampled at 8 kHz, then parameterized into 12 MFCCs along with log-energy, and the first and second order time derivatives of these parameters, yielding a 39-dimensional feature vector. Cepstral mean subtraction (CMS) was performed on a per-speaker basis to remove the channel effect of the features.

Baseline gender independent (GI) SI model was trained with the training set which contains 54 female and 55 male speakers. Considering the monosyllabic structure of Chinese language in which each syllable can be decomposed into an INITIAL/FINAL format [6], 112 right-context-dependent INITIALs along with 38 context-independent FINALs were used as the acoustic units for SI model training. Each INITIAL is represented by an HMM with 2 or 3 states, while each FINAL is represented with 4 states. The mixture number per state ranges from 1 to 8,

depending on the amount of training data available. In addition, a 1-state HMM with 32 mixtures was trained to handle silence and short pause. The total number of Gaussian mixtures in the system is approximately 2370. For each testing speaker, the first 60 sentences were taken as adaptation data while the rest were for recognition. Each adaptation sentence is of an average length of 1.37 seconds and consists of 4.6 syllables in average. The recognizer performed only free syllable decoding without any grammar constraints, and the recognition rate was calculated in syllable accuracy. All adaptation experiments were conducted in a supervised manner. The speaker independent recognition accuracy is 55.81%, averaged over 5 female and 5 male testing speakers.

As baseline adaptation experiments, both of conventional MLLR approaches using full and diagonal regression matrices were conducted with respect to different number of adaptation sentences. The full matrix MLLR was based on a global regression class, while the diagonal matrix MLLR utilized a regression class tree for dynamic regression class generating. The results are summarized in Table 1. As can be seen, full matrix MLLR performed poorly when only 10 adaptation sentences were applied due to the large number of parameters. However, as the data increases, full matrix MLLR gave better adaptation result for its superior modeling of the speaker characteristics.

In the training phase of the eigenspace-based approach, 120 sentences for each of 109 training speakers were used to compute speaker-specific regression matrices and, then, PCA was performed to extract 109 bases for eigenspace. The proposed approach was first tested with the eigenspace-based regression matrices taken directly as the transformations for mean adaptation. We applied different numbers of eigen-matrices for 10 adaptation sentences. The results are summarized in Table 2. We see that for all numbers of eigen-matrices used, the eigenspace-based regression matrices gave great improvement in adaptation over both conventional MLLR approaches, which means the eigen-matrices can effectively capture the inter-speaker variation as we expected.

Finally, with the number of eigen-matrices fixed at 50, the overall adaptation procedure of our approach was conducted with respect to different number of adaptation sentences. The adaptation results are shown in Table 1. Ws can see that the proposed eigenspace-based MLLR approach (Eigen) significantly outperformed both types of conventional MLLR approaches when the amount of adaptation data is small (less than 30 sentences). Moreover, as adaptation sentences increased, the proposed approach gave slightly better results over full matrix MLLR. This means that with the transformation computation dominated by full matrix MLLR, the proposed approach can take advantage of precise modeling of full regression matrices, as well as robust initial estimate due to the use of eigen-matrices.

Num. Sen.	10	20	30	40	50	60
Diag	57.12	58.00	59.17	59.38	59.89	59.98
Full	47.97	56.80	58.58	60.12	60.88	61.48
Eigen	59.27	60.00	60.53	60.72	61.02	61.50

Table 1: Adaptation performance of conventional diagonal matrix MLLR (Diag), full matrix MLLR (Full), and eigenspace-based MLLR (Eigen) with respect to different number of adaptation sentences, in syllable accuracy (%). SI accuracy is 55.81%.

K	10	20	30	40	50
Syl. Acc.(%)	57.63	58.64	58.55	59.24	59.20

Table 2: Adaptation performance of eigenspace-based MLLR using only eigenspace-based regression matrices, with respect to different number of eigen-matrices (*K*).

5. CONCLUSION

In this paper, an eigenspace-based approach for improving the modeling accuracy of conventional MLLR was proposed. We introduced *a priori* knowledge analysis on training speakers so as to construct an eigenspace in which the speaker-specific regression matrices are located. Eigenspace-based matrices were then used as an initial estimate of MLLR regression matrices for each new speaker. Improvements in supervised adaptation showed the effectiveness of the proposed approach. We will further study this technique on the issues of combining several heterogeneous databases, as well as on the application of environmental adaptation.

REFERENCES

[1] P. C. Woodland, "Speaker Adaptation: Techniques and Challenges", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.85-90, 2000.

[2] C. J. Leggetter & P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, vol. 9, pp.171-185, 1995.

[3] M. J. F. Gales, "The Generation and Use of Regression Class Trees for MLLR Adaptation", *Technical Report*, CUED/F-INFENG/TR263, Cambridge University, 1996.

[4] R. Kuhn, *et. al.*, "Eigenvoices for Speaker Adaptation", *Proc. ICSLP' 98*, pp.1771-1774, 1998.

[5] I. T. Jolliffe, "Pricipal Component Analysis", Springer-Verlag, 1986.

[6] B. Chen, H. M. Wang & L. S. Lee, "Retrieval of Broadcast News Speech In Mandarin Chinese Collected In Taiwan Using Syllable-Level Statistical Characteristics", *Proc. ICASSP2000*, pp.2985-2988, 2000.