RAPID SPEAKER ADAPTATION USING A PRIORI KNOWLEDGE BY EIGENSPACE ANALYSIS OF MLLR PARAMETERS

Nick J.-C. Wang[1,2], *Sammy S.-M. Lee*[1,2], *Frank Seide*[1], *and Lin-Shan Lee*[2]

[1] Philips Research East Asia-Taipei
[2] Graduate Institute of Communication Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
Nick.Wang@philips.com

ABSTRACT

This paper considers the problem of rapid speaker adaptation in speech recognition. In particular, we exploit an approach based on combination of transformations, which utilizes the concepts of both maximum likelihood linear regression (MLLR) and eigenvoice adaptation. We analyze three different possible methods to realize the concept, and formulate a fast algorithm of maximum likelihood coefficient estimation for test speakers. It was found that the best approach can properly utilize the a priori knowledge of speaker-independent models in constructing the eigenspace for speaker characteristics, while using MLLR matrices in representing the specific speakers so as to reduce the on-line memory and computation requirement of the adaptation phase. This best approach leads to identical models as eigenvoice adaptation that is based on MLLR-adapted speaker models. The experimental results and discussions also provide a good analysis towards integration of MLLR and eigenvoice approaches.

1. INTRODUCTION

Speaker variability is one of several important sources of variability in speech recognition. Much work has been done on speaker adaptation techniques based on speaker-independent (SI) models with a small amount of data. Among these approaches, maximum likelihood linear regression (MLLR) and eigenvoice both work very well with sparse adaptation data, and maximum a posteriori (MAP) adaptation performs better with abundant data [1][2]. This paper focuses on sparse data adaptation, in particular the ways to integrate the advantages of both MLLR and eigenvoice.

MLLR adaptation. MLLR adaptation is a well-known systematic linear-regression transformation scheme which is optimized based on the maximum likelihood criterion [3][4]. A set of regression classes is defined. Data within each class are pooled to evaluate a general regression transformation for this class, and the same transformation is applied to a number of model parameters. It offers very good performance when only spase data are available. It uses only a very limited number of parameters to represent the mismatch over speakers. However, it does not utilize any a priori knowledge of the distribution of speaker characteristics beyond the linear-regression-transformation framework.

Eigenvoice adaptation. Eigenvoice adaptation [5][6][7] performs principle component analysis (PCA) on model parameters of many training speakers,

$$\begin{aligned}\mathbf{C_a} &= \mathbf{E_a}\mathbf{\Lambda_a}\mathbf{E_a}^{\mathrm{T}} \\ &\cong \left(\mathbf{E_a}\mathbf{Q_a}\right)\,\mathbf{\Lambda_a}\,\left(\mathbf{E_a}\mathbf{Q_a}\right)^{\mathrm{T}}, \qquad (1)\end{aligned}$$

where $\mathbf{C_a}$ is the covariance matrix of an augmented vector of model parameters $\mathbf{a}$, $\mathbf{\Lambda_a}$ is the eigenvalue matrix, $\mathbf{E_a}$ is the eigenspace matrix formed by eigenvectors, and $\mathbf{Q_a}$ is a diagonal mask matrix with 1-elements in the selected subspaces that correspond to the top n eigenvalues of $\mathbf{\Lambda_a}$ and 0 elsewhere. With PCA result, it is thus possible to reduce the number of necessary adaptation coefficients for a specific speaker by selecting the most significant axes [8]. Consequently, it overcomes to some degree the problem of lack of training data in real applications. It provides a reliable performance in the beginning of adaptation, within first several seconds of speech. However, the adaptation model is very big. Every eigenvoice, or every principle component, has the same size as the SI model. For example, for 50 eigenvoices and an SI model with 60K densities and 25 feature components we need to store 75 million parameters (or 300MB). This is more than the RAM size in a normal PC today.

Combinations of the above approaches. The eigenvocie adaptation has been extended to large-vocabulary continuous-speech recognition (LVCSR) using mixture density HMM with the help of MLLR and MAP in training of speaker models [9]. In this approach, MLLR adaptation solved the alignment of mixture Gaussian densities in LVCSR systems when eigenvoice adaptation is applied.

On the other hand, PCA has been performed on the supervector $\mathbf{w}$ formed by structured MLLR parameters for many speakers to generate a basis of linear regression matrices [10]. The expression is analogous to Eq. (1),

$$\begin{aligned}\mathbf{C_w} &= \mathbf{E_w}\mathbf{\Lambda_w}\mathbf{E_w}^{\mathrm{T}} \\ &\cong \left(\mathbf{E_w}\mathbf{Q_w}\right)\,\mathbf{\Lambda_w}\,\left(\mathbf{E_w}\mathbf{Q_w}\right)^{\mathrm{T}}, \qquad (2)\end{aligned}$$

where $\mathbf{C_w}$ is the covariance matrix of the supervector for structured MLLR parameters $\mathbf{w}$, $\mathbf{\Lambda_w}$ is the eigenvalue matrix, $\mathbf{E_w}$ is the eigenspace matrix, $\mathbf{Q_w}$ is the diagonal mask matrix for top n eigenvalues in $\mathbf{\Lambda_w}$. Representing speaker-variability by the eigenspace of MLLR parameters hugely reduces the memory requirement. For example, for 50 eigenvectors of a 10-class structured MLLR and 25 feature components we need to store 325,000 parameters (or 1.3MB). This is feasible in many real applications.

Both these approaches have indicated that the integration of eigenvoice and MLLR adaptation makes good sense.

This paper presents three possible new approaches to integrate the concepts of MLLR and eigenvoice, referred to as Approaches A, B and C, in sections 2, 3 and 4 respectively. The experimental results in section 5 indicated that Approach B gives the best results.

2. APPROACH A: PCA ON MLLR PARAMETERS WITH REDUCED MEMORY AND COMPUTATION REQUIREMENT

Approach A performs PCA on MLLR parameters just as the previously proposed approach [10] as expressed in Eq. (2), but uses a novel fast algorithm for maximum likelihood coefficient estimation. In this way, not only the problem of huge memory requirement in eigenvoices can be solved, but also the computation load can be reduced. There are further advantages of performing PCA on MLLR parameters. To train MLLR matrices for a specific speaker is easier than to train a speaker dependent model with small amounts of data. The size of MLLR matrices remains the same, so is that of their eigenvectors, no matter how many densities are used in an SI model. The a priori knowledge of MLLR matrices could be easily reused in different corpora if the feature extraction is the same.

The approach contains three parts: basis generation (training phase), maximum likelihood estimation of eigen-coefficients, and construction of the adapted model (adaptation phase).

2.1. Basis generation

The first part is the basis generation in the training phase, by performing PCA on MLLR parameters of training speakers. The eigenspace of is generated according to Eq. (2).

2.2. Maximum likelihood coefficient estimation

The coefficient estimation algorithm is derived by taking derivatives over an auxillary function and set them to zeros. For each eigen-dimension i, $i = 1..n$, equations can be written as

$$\sum_{s=1}^{S}\sum_{r=1}^{R_s}\sum_{t=1}^{T}\gamma_{s_r}(t)\Big(\mathbf{o}(t) - \bar{\mathbf{W}}_s\xi_{s_r}\Big)^{\mathrm{T}}\mathbf{\Sigma}_{s_r}^{-1}\mathbf{W}_s^{(i)}\xi_{s_r} \;=\; \sum_{s=1}^{S}\sum_{r=1}^{R_s}\sum_{t=1}^{T}\gamma_{s_r}(t)\Big(\sum_{j=1}^{n}c_j\mathbf{W}_s^{(j)}\xi_{s_r}\Big)^{\mathrm{T}}\mathbf{\Sigma}_{s_r}^{-1}\mathbf{W}_s^{(i)}\xi_{s_r}, \qquad (3)$$

where c_j, $i = 1..n$, are the coefficients to be estimated, $\mathbf{W}_s^{(i)}$ is the s-th class MLLR parameters of the i-th eigenvector in the form of a $D \times (D+1)$ matrix where D the dimension of the acoustic feature, $\bar{\mathbf{W}}_s$ the s-th class speaker mean MLLR matrix over all speakers, $\mathbf{o}(t)$ the observation at time t, $\gamma_{s_r}(t)$ the occupation probability of density s_r at time t, density s_r belongs to the s-th class in structured MLLR, ξ_{s_r} is the extended mean vector of density s_r in the SI model, and $\mathbf{\Sigma}_{s_r}$ the covariance matrix of density s_r.

Note that Eq. (3) is identical to the coefficient estimation for eigenvoices given the eigenvoice mean vectors $\mu_{s_r}^{(i)}$ of density s_r, as well as those in the mean model $\bar{\mu}_{s_r}$, by the following on-line transformations:

$$\mu_{s_r}^{(i)} \;=\; \mathbf{W}_s^{(i)}\xi_{s_r}, \qquad (4)$$

$$\text{and} \quad \bar{\mu}_{s_r} \;=\; \bar{\mathbf{W}}_s\xi_{s_r}. \qquad (5)$$

However, Eq. (3) demands much less memory compared to eigenvoice adaptation. On the other hand, the computational effort of the on-line transformations would be intensive. However, we can rearrange this coefficient estimation by new expression of MLLR transformation

$$\mathbf{W}_s^{(i)}\xi_{s_r} \;=\; \mathbf{L}_{s_r}\mathbf{w}_s^{(i)}, \qquad (6)$$

where $\mathbf{w}_s^{(i)}$ is $\mathbf{W}_s^{(i)}$ rearranged in the form of a $(D^2 + D) \times 1$ column vector, and $\mathbf{L}_{s_r}$ is the SI mean μ_{s_r} in the form of a $D \times (D^2 + D)$ matrix.

With this, Eq. (3) can be written as

$$\sum_{s=1}^{S}\sum_{r=1}^{R_s}\sum_{t=1}^{T}\gamma_{s_r}(t)\Big(\mathbf{o}(t) - \bar{\mu}_{s_r}\Big)^{\mathrm{T}}\mathbf{\Sigma}_{s_r}^{-1}\mathbf{L}_{s_r}\mathbf{w}_s^{(i)} \;=\; \sum_{j=1}^{n}\sum_{s=1}^{S}\sum_{r=1}^{R_s}\sum_{t=1}^{T}c_j\mathbf{w}_s^{(j)\,\mathrm{T}}\mathbf{L}_{s_r}^{\mathrm{T}}\gamma_{s_r}(t)\mathbf{\Sigma}_{s_r}^{-1}\mathbf{L}_{s_r}\mathbf{w}_s^{(i)}. \qquad (7)$$

The implementation of Eq. (7) can be sped up by two steps (divide & conquer method). The first step involves no eigenvectors:

$$\mathbf{x}_s \;=\; \sum_{r=1}^{R_s}\sum_{t=1}^{T}\gamma_{s_r}(t)\Big(\mathbf{o}(t) - \bar{\mu}_{s_r}\Big)^{\mathrm{T}}\mathbf{\Sigma}_{s_r}^{-1}\mathbf{L}_{s_r}, \qquad (8)$$

$$\text{and} \quad \mathbf{Z}_s \;=\; \sum_{r=1}^{R_s}\sum_{t=1}^{T}\mathbf{L}_{s_r}^{\mathrm{T}}\gamma_{s_r}(t)\mathbf{\Sigma}_{s_r}^{-1}\mathbf{L}_{s_r}, \qquad (9)$$

where $\mathbf{x}_s$ and $\mathbf{Z}_s$ are auxillary terms of a $1 \times (D^2 + D)$ row vector and of a $(D^2 + D) \times (D^2 + D)$ matrix respectively The second step is the computation of eigenvectors and auxillary terms of the first step.

$$\sum_{s=1}^{S}\mathbf{x}_s\mathbf{w}_s^{(i)} \;= \sum_{j=1}^{n}\sum_{s=1}^{S}c_j\mathbf{w}_s^{(j)\,\mathrm{T}}\mathbf{Z}_s\mathbf{w}_s^{(i)}. \qquad (10)$$

The computation load is mostly in the first step of accumulation related to density means and variances and occupation probabilities. Since it is independent of the number of eigenvectors and performed once, the computation load is largely reduced. The computation in the eigenvoice coefficient estimation has computational complexity of $\mathcal{O}(n^2DM')$, and the additional on-line transformation of $\mathcal{O}(nD^2M')$. On the other hand, the fist step of this fast implementation takes $\mathcal{O}(\frac{1}{2}D^3M')$, and the second step takes $\mathcal{O}(SnD^3 + \frac{1}{2}Sn^2D^2)$. For example, with 50 eigenvectors and 10-class structured MLLR and 50 feature components and 10-second speech with frame-shift of 10 ms, while the eigenvoice coefficient estimation with on-line density-wise transformation takes 94 mio calculations including 31 mio of on-line transformation, the fast implementation takes only 23 mio including 8 mio of the first step and 15 mio of the second step, in a factor of 4. When 100 second adaptation speech are used, the efficient factor increases to 10. As a result, the eigen-coefficient estimation uses much less memory and computation compared to eigenvoice adaptation.

2.3. Construction of the adapted model

The supervector $\mathbf{w}$ for the structured MLLR matrices for the new speaker can then be obtained from the mean supervector $\bar{\mathbf{w}}$, the eigenspace $\mathbf{E_w Q_w}$, and its coefficient vector $\mathbf{c}$,

$$\mathbf{w} \;\cong\; \bar{\mathbf{w}} + \mathbf{E_w Q_w c}. \qquad (11)$$

Hence, the density mean supervector is

$$\mathbf{a} = \mathbf{L}\;\mathbf{w}, \qquad (12)$$

where $\mathbf{w}$ is obtained in Eq. (11), and $\mathbf{L}$ is the transformation matrix from $\mathbf{w}$ to $\mathbf{a}$ composed of SI model parameters. $\mathbf{L}$ has a size of $\{M \cdot D\} \times \{S(D^2 + D)\}$, where M is the number of densities, and S is the number of classes in structured MLLR adaptation.

3. APPROACH B: INCLUDING SI MODEL INFORMATION IN PCA

Approach A (previous section) is based on combination of transformations, while eigenvoice adaptation is based on combination of density means. We'd like to analyze the relationship between them. It is described by Eq. (12) the relationship between the density mean supervector $\mathbf{a}$ and the MLLR parameter supervector $\mathbf{w}$. As a result, the relationship between their covariance matrices $\mathbf{C_a}$ and $\mathbf{C_w}$ is:

$$\mathbf{C_a} = \mathbf{L}\ \mathbf{C_w}\ \mathbf{L}^{\mathrm{T}}. \tag{13}$$

With this relationship, an improved method to include the SI model information in the framework of Approach A proposed above is developed, referred to as *Approach B*. The method differs from Aprroach A only in the basis generation. The details are given later. Coefficient estimation and adapted-model construction are exactly the same. With this approach, eigenvoice adaptation can be exactly implemented under the constraint of using MLLR-adapted speaker models for basis generation, but with much less memory and computation requirement.

With the linear relationship described in Eq. (13), the PCA performed on $\mathbf{C_a}$ is related to the PCA performed on $\mathbf{C_w}$ as follows:

$$\begin{aligned}
\mathbf{C_a} &= \mathbf{L}\ \mathbf{C_w}\ \mathbf{L}^{\mathrm{T}} \\
&= \mathbf{L}\ \mathbf{E_w}\mathbf{\Lambda_w}\mathbf{E_w}^{\mathrm{T}}\ \mathbf{L}^{\mathrm{T}} \\
&= \left(\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_b^{-1}\right)\ \mathbf{\Omega}_b\mathbf{\Lambda_w}\mathbf{\Omega}_b\ \left(\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_b^{-1}\right)^{\mathrm{T}} \\
&\cong \left(\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_b^{-1}\mathbf{Q}_b\right)\ \mathbf{\Omega}_b^2\mathbf{\Lambda_w}\ \left(\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_b^{-1}\mathbf{Q}_b\right)^{\mathrm{T}},
\end{aligned} \tag{14}$$

where $\mathbf{Q}_b$ is another diagonal mask matrix according to $\mathbf{\Omega}_b^2\mathbf{\Lambda_w}$, and $\mathbf{\Omega}_b$ is a diagonal length matrix of $\mathbf{L}\ \mathbf{E_w}$ for vector length normalization purposes with $\mathbf{\Omega}_b^{(ii)} = \|\mathbf{L}\ \mathbf{E_w}^{(i)}\|$ where $\mathbf{E_w}^{(i)}$ is the i-th eigenvector in eigenspace matrix $\mathbf{E_w}$. The mask matrix $\mathbf{Q}_b$ of this approach is not the same as the mask matrix $\mathbf{Q_w}$ of Approach A, because the former is constructed according to $\mathbf{\Omega}_b^2\mathbf{\Lambda_w}$, while the latter according to $\mathbf{\Lambda_w}$. Hence, the chosen subspaces for eigenspace in Approach B and A are potentially different.

Moreover, by comparison between Eq. (14) and Eq. (1), it is clear that

$$\mathbf{\Lambda_a} = \mathbf{\Omega}_b^2\mathbf{\Lambda_w}, \tag{15}$$

$$\text{and}\quad \mathbf{E_a} = \mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_b^{-1}. \tag{16}$$

Because $\mathbf{\Omega}_b^2\mathbf{\Lambda_w}$ is equal to $\mathbf{\Lambda_a}$ as pointed out in Eq. (15), the construction processes of $\mathbf{Q}_b$ must be equal to those of $\mathbf{Q_a}$. Consequently, the chosen subspaces for the eigenspace in this approach are the same as those of the eigenvoice approach in the sense of the linear relationship of in Eq. (16).

Therefore, the difference between Approach A and B is that the selection of eigenspace in Approach B includes the SI model information, which is missing in Approach A. Aside from this, coefficient estimation and adapted-model construction remain the same as for Approach A. And the adaptation capability provided by this approach are the same as eigenvoice adaptation since the chosen subspaces by this approach are equivalent to those by eigenvoice in the sense of the linear relationship of in Eq. (16), but represented in a memory and computationally efficient way.

4. APPROACH C: PERFORMING LDA INSTEAD OF PCA FOR APPROACH B

The significant components for the vector space can also be generated by linear discriminant analysis (LDA), instead of PCA. *Approach C* is the LDA version of Approach B. In the supervector $\mathbf{a}$ for different speakers, each component represents a different feature of the speaker voice characteristics. The intra-speaker variances of all components can be so different that the discriminating ability of choosing the top n principle axes can be influenced. The difference between the first cepstral coefficient and the twelfth cepstral coefficient can be a factor of ten, for example. LDA takes into account the self-variances when choosing the principle axes.

However, it is difficult to estimate the within-class and total scatter matrix, at least one of which is needed to perform LDA. We assume that the covariances in the SI model can be used to form the total scatter matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{\Sigma}_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathbf{\Sigma}_M \end{bmatrix}, \tag{17}$$

where the $\mathbf{\Sigma}_m$, $m = 1..M$, are diagonal covariance matrices. The optimization is then to minimize the trace

$$tr\left(\mathbf{\Sigma}^{-1}\mathbf{C_a}\right) = tr\left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{C_a}\mathbf{\Sigma}^{-\frac{1}{2}}\right). \tag{18}$$

We can write

$$\begin{aligned}
&\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{C_a}\mathbf{\Sigma}^{-\frac{1}{2}} \\
=\ &\mathbf{\Sigma}^{-\frac{1}{2}}\ \left(\mathbf{L}\ \mathbf{E_w}\mathbf{\Lambda_w}\mathbf{E_w}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\right)\ \mathbf{\Sigma}^{-\frac{1}{2}} \\
=\ &\left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_c^{-1}\right)\ \mathbf{\Omega}_c\mathbf{\Lambda_w}\mathbf{\Omega}_c\ \left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_c^{-1}\right)^{\mathrm{T}} \\
\cong\ &\left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_c^{-1}\mathbf{Q}_c\right)\mathbf{\Omega}_c^2\mathbf{\Lambda_w}\left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{L}\ \mathbf{E_w}\mathbf{\Omega}_c^{-1}\mathbf{Q}_c\right)^{\mathrm{T}},
\end{aligned} \tag{19}$$

where $\mathbf{Q}_c$ is another diagonal mask matrix constructed according to $\mathbf{\Omega}_c^2\mathbf{\Lambda_w}$, and $\mathbf{\Omega}_c$ is a diagonal length matrix of $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{L}\ \mathbf{E_w}$ for vector length normalization purposes with $\mathbf{\Omega}_c^{(ii)} = \|\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{L}\ \mathbf{E_w}^{(i)}\|$.

The result is very similar to that in the previous section. Coefficient estimation and adapted-model construction are exactly the same as for Approach B.

5. EXPERIMENTAL RESULTS

5.1. Experimental setup

In the experiments, the Philips research speech recognition system [11] was used, which is a HMM-based large-vocabulary continuous-speech recognizer. Standard MFCC features with first-order derivatives, sentence-level cepstral mean subtraction (CMS), and Gaussian mixture densities with density-specific diagonal covariance matrices were applied.

	Train	Adapt	Test
#Speakers	241	20	
#Utterances	27606	1000	1000
#Syl./Utt.	30.1	35.0	35.3

***Table 1**: Corpus characteristics.*

The experiments were conducted on a PC dictation database of Mandarin Chinese recorded in Taiwan. Training data of 241 speakers were used to train an SI model, to estimate 241 MLLR full matrices, and to generate eigenvectors and so on. Supervised adaptation was performed by adaptation data of another 20 testing speakers. Test data of the same 20 testing speakers were used for free syllable decoding, without bias from the prior knowledge of word occurrences or connections, for performance evaluation. The syllable error rate (SER)[1]was taken as the performance measure. Table 1 summarizes the corpus.

5.2. Comparison between MLLR adaptation and Approach A

Approach A can be considered as a PCA version of MLLR adaptation, in which the estimated coefficients are located in a subspace of the MLLR parameter space. As a result, it is more reliable if only a small amount of data is used, but less accurate when adequate amount of data become available.

Adapt. data	SI	MLLR	Approach A	
				#coef.
0 sec	**28.3**	-	28.6	0
3 sec	28.3	30.4	**28.2**	20
7 sec	28.3	28.4	**27.8**	81
15 sec	28.3	27.5	**27.2**	81
27 sec	28.3	**26.9**	**26.9**	121
43 sec	28.3	**26.6**	**26.6**	121
60 sec	28.3	**26.5**	**26.5**	121
120 sec	28.3	**26.2**	26.4	121
240 sec	28.3	**26.2**	26.3	201

Table 2: *Performance (SER %) of Approach A compared with SI and MLLR.*

The dimension of this subspace can be dynamically chosen according to the amount of available adaptation data. The recognition performance of Approach A in Table 2 were obtained by choosing the best number of dimensions, optimized on the test set, among 1, 13, 20, 41, 81, 121, 161, 201 and 240. The minimum dimension was chosen if more than one cases gave the same results.

The results in Table 2 show that Approach A is more reliable than MLLR when the amount of adaptation data is small, e.g. 3-15 seconds, and the two approaches are nearly the same when there is adequate data for estimation, e.g. 27-60 seconds. MLLR adaptation performs slightly better when there are more data, e.g. 120-240 seconds, probably because the degree of freedom for Approach A in this case is at most 240 (constrained by the number of training speakers) which is less than that of MLLR, 506 here.

5.3. Comparison between eigenvoice adaptation and Approaches A, B, & C

Adapt. data	#coef.	eigenvoice	Approach		
			A	B	C
3 sec	13	28.4	28.7	**28.3**	28.5
3 sec	20	28.5	**28.2**	28.3	28.4
7 sec	13	28.0	28.3	**27.9**	28.2
7 sec	20	28.0	**27.9**	**27.9**	28.0
15 sec	13	**27.7**	28.0	**27.7**	27.8
15 sec	20	27.7	27.7	**27.6**	27.7
27 sec	13	27.8	27.9	**27.6**	27.9
27 sec	20	**27.5**	27.7	**27.5**	27.7

Table 3: *Comparison between the three proposed approaches and the eigenvoice adaptation (SER %).*

Table 3 shows the results obtained by Approaches A, B, & C, and the eigenvoice adaptation. The eigenvoice experiments were carried out with a special training of the speaker models, which includes the normal SI model training procedure plus MLLR adaptation performed on the SI model using the training speakers' training data. From the data in Table 3, Approach B is shown to be better than Approach A especially when the number of eigen-coefficients is small. Apparently this is because in Approach B the speaker independent characteristics has been included, while this is not the case in Approach A. Secondly, the performance of Approach B is shown to be almost identical to that of eigenvoice, even if Approach B requires much smaller memory size. The minor differences in performance are probably due to different programming or truncation errors. At last, Approach C in general yielded no improvement. It may be due to the relatively strong assumption of the diagonal total scatter matrix.

6. CONCLUSIONS

In this paper three new approaches of rapid speaker adaptation have been developed, which perform PCA/LDA on MLLR matrices and maximum likelihood cofficient estimation with significantly reduced memory and computation requirement. With PCA, we can extract a set of ordered principle components and use them flexibly with respect to the amount of available adaptation data. It has been shown that this approach is better than MLLR adaptation in accuracy with a small amount of adaptation data. One variant (Approach B) is equivalent to that of eigenvoice adaptation based on MLLR-adapted speaker models, but requires far less online memory and computation during adaptation.

7. REFERENCES

[1] J.-L. Gauvain & C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov Chains. In *IEEE Trans. on Speech and Audio Processing*, v. 2, pp. 291-298, April, 1994.

[2] E. Thelen, X. Aubert, & P. Beyerlein. Estimation of the a priori speaker variability for very fast adaptation. In *Proc. ICASSP*, v. 2, pp. 1035–1038, Munich, 1997.

[3] C.J. Leggetter & P.C. Woodland. Speaker adaptation of HMMs using linear regression. *Tech. Report TR. 181, Engineering Department, Cambridge Univ.*, June 1994.

[4] C.J. Leggetter & P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. In *Computer Speech & Language*, v. 9, pp. 171–185, 1995.

[5] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, & M. Contolini. eigenvoices for speaker adaptation. In *Proc. ICSLP*, v. 5, pp. 1771–1774, Sydney, 1998.

[6] P. Nguyen, C. Wellekens, & J.-C. Junqua. Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments. In *Proc. EUROSPEECH*, v. 6, pp. 2519–2522, Budapest, 1999.

[7] R. Westwood. Speaker adaptation using eigenvoices. *MPhil Thesis, Univ. of Cambridge*, August 1999.

[8] I.T. Jolliffe. Principle component analysis. Springer, 1986.

[9] H. Botterweck. Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices. In *Proc. ICSLP*, v. 4, pp. 354-357, Beijing, 2000.

[10] K.-T. Chen, W.-W. Liau, H.-H. Wang, & L.-S. Lee. Fast speaker adaptation using eigenspace-based maximum likelihood linear regression. In *Proc. ICSLP*, v. 3, pp. 742-745, Beijing, 2000.

[11] F. Seide and N. Wang. Phonetic modelling in the Philips Chinese continuous-speech recognition system. In *Proc. ISCSLP*, pp. 54–59, Singapore, 1998.

[1] In Mandarin, it is common to assess the accuracy of an acoustic model by free syllable recognition, in analogy to phone recognition in Western languages.