



# Clustered Maximum Likelihood Linear Basis for Rapid Speaker Adaptation

Yun Tang and Richard Rose

Department of Electrical and Computer Engineering,  
McGill University, Montréal, Quebec, Canada

yun.tang3@mail.mcgill.ca, rose@ece.mcgill.ca

## Abstract

Speaker space based adaptation methods for automatic speech recognition have been shown to provide significant performance improvements for tasks where only a few seconds of adaptation speech is available. This paper proposes a robust, low complexity technique within this general class that has been shown to reduce word error rate, reduce the large storage requirements associated with speaker space approaches, and eliminate the need for large numbers of utterances per speaker in training. The technique is based on representing speakers as a linear combination of clustered linear basis vectors and a procedure is presented for ML estimation these vectors from training data. Significant word error rate reduction was obtained relative to speaker independent performance for the Resource Management and Wall Street Journal task domains.

**Index Terms:** speech recognition, speaker adaptation

## 1. Introduction

Adaptation of hidden Markov models (HMMs) in automatic speech recognition (ASR) has been performed under a variety of scenarios. HMMs are updated using a variety of parameterizations in order to minimize mismatch between an adapted acoustic model and adaptation utterances based on several different optimization criteria [1]. This paper is concerned with supervised and unsupervised adaptation scenarios where only a very small amount of adaptation speech, on the order of three to five seconds in length, is available for adapting acoustic models.

Techniques that are based on representing the potential sources of variability in adaptation utterances within a low dimensional subspace have been shown to be the most successful at achieving significant reductions in ASR word error rate (WER) when using very short adaptation utterances. Techniques in this general class were originally presented for speaker adaptation under the headings of eigenvoice modeling and cluster adaptive training (CAT) [2][3][4]. Many related techniques that are also based on low dimensional “speaker space” representations have been presented more recently [5]. All of these procedures can be roughly characterized as having two separate stages for training and adaptation. In the training stage, a set of basis vectors in a low dimension subspace is identified from training data either through principle components analysis (PCA) or by maximum likelihood estimation. In the adaptation stage, new HMM parameters are obtained as a weighted linear combination of these basis vectors where the weights are estimated from the adaptation utterances.

This paper presents techniques that both reduce the amount of data required in the training stage of this process and also re-

duce the storage requirements during the adaptation stage. This is important because these adaptation techniques have generally been applied to modelling speaker variability in a very high dimensional parameter space. Typically, adaptation is performed on the supervector  $\vec{\mu} = (\vec{\mu}'_1, \vec{\mu}'_2, \dots, \vec{\mu}'_M)'$  formed by concatenating the mean vectors associated with all Gaussian densities in a continuous Gaussian mixture density HMM. Each of the vectors  $\vec{\mu}_m \in \mathbb{R}^D$  is a mean vector for the  $m$ th Gaussian density and it is not unusual for there to be  $M = 100,000$  densities in a large vocabulary HMM system. This can result in an overall dimensionality,  $MD$ , of over a million. The speaker space procedures estimate a set of basis vectors,  $\vec{e}(k)$ ,  $k = 1, \dots, K$ , from training data where  $K \ll MD$  and typically lies in the range  $10 < K < 100$ . During the adaptation stage, a  $K$  dimensional weight vector,  $\vec{w} = (w_1, \dots, w_K)'$ , is estimated from the adaptation data and the adapted supervector,  $\hat{\vec{\mu}}$ , is computed as

$$\hat{\vec{\mu}} = \vec{\mu} + \sum_{k=1}^K w_k \cdot \vec{e}(k). \quad (1)$$

All of the speaker space methods suffer in some way from two issues. The first issue relates to problems with data sparsity in estimating the basis vectors,  $\vec{e}(k)$ ,  $k = 1, \dots, K$ , and determining the speaker space dimension in the training stage. This problem is particularly severe in the eigenvoice approaches since they require sufficient data from each training speaker to provide statistically robust estimation of speaker specific supervectors for PCA estimation of the basis vectors. The second issue relates to the heavy storage requirements associated with these methods in the adaptation stage. The memory overhead for storage of the  $MD$  dimensional basis functions,  $\vec{e}(k)$ ,  $k = 1, \dots, K$ , is made particularly severe by the high dimensionality of the original feature space which can easily equal  $MD = 4$  million parameters. Even for a  $K = 10$  dimensional speaker space, the storage requirements for the basis vectors in Equation 1 can exceed 160 Mbytes.

There are three contributions associated with the work described in this paper. The first is a speaker subspace approach that relies on “clustered basis vectors.” The basis vectors are formed by concatenating  $N$  clustered subvectors where  $N$  is much less than  $M$ , the number of HMM Gaussian densities. The effect of these clustered basis functions is improved statistical robustness in training and considerable reduction in complexity during adaptation. The second contribution is an expectation-maximization (EM) algorithm for maximum likelihood estimation of clustered basis vectors from the training data. This representation will be referred to in the paper as a maximum likelihood linear basis (MLLB). Finally, the third contribution is an experimental study for determining whether the success of this general class of techniques is due to its ability to describe interspeaker variability or whether these techniques

This work was performed in collaboration with the DIVINES FP6 project and supported under NSERC Program Number 307188-2004

in practice characterize more general speech variability.

The rest of the paper is organized as follows. A brief introduction to eigenvoice-based speaker adaptation is given in Section 2. In Section 3, the EM algorithm for estimating the MLLB vectors in a clustered feature space will be presented. Finally, the results of an experimental study comparing the methods introduced here with a variant of the eigenvoice modeling approach for supervised and unsupervised speaker adaptation on the Resource Management (RM) task will be presented in Section 4. Experimental results will also be presented for unsupervised adaptation on the Wall Street Journal (WSJ) corpus.

## 2. Eigenvoice Speaker Adaptation

As originally proposed by Kuhn et al, the eigenvoice method for speaker space based speaker adaptation begins by training a set of speaker dependent (SD) continuous mixture density HMMs for each of a relatively large population of speakers [2]. A set of basis vectors are computed using principal components analysis (PCA) applied to the supervectors derived from these SD HMM's as described in Section 1. Adapting the HMM to the observation vectors,  $O^s = \vec{o}_1^s \dots \vec{o}_T^s$ , from speaker  $s$  is performed by estimating the weights,  $\vec{w}^s = (w_1^s, \dots, w_K^s)'$ , and updating the supervector,  $\vec{\mu}$ , as shown in Equation 1. A maximum likelihood estimate of the weight vector,  $\vec{w}^s$ , is obtained using the EM algorithm by maximizing the auxiliary Q function

$$Q(\Lambda, \hat{\Lambda}) = - \sum_{m=1}^M \sum_{s=1}^S \sum_{t=1}^T \lambda_m(t) (\vec{o}_t^s - \hat{\vec{\mu}}_m^s)' \Sigma_m^{-1} (\vec{o}_t^s - \hat{\vec{\mu}}_m^s) \quad (2)$$

$$\hat{\vec{\mu}}_m^s = \vec{\mu}_m + \sum_{k=1}^K w_k^s \vec{e}(k, m) \quad (3)$$

where  $\lambda_m(t)$  is the occupation probability of the  $m$ th Gaussian density for observation vector  $\vec{o}_t$ ,  $\mu_m$  and  $\Sigma_m$  are the mean vector and covariance matrix of Gaussian  $m$ ,  $\vec{e}(k, m)$  is a subvector of basis vector  $\vec{e}(k)$  corresponding to the  $m$ th Gaussian in the HMM, and  $\Lambda$  corresponds to  $\vec{\mu}$ .

By maximizing  $Q(\Lambda, \hat{\Lambda})$  with respect to  $\vec{w}^s$ , it can be shown that the optimum weight vectors can be obtained by solving the following matrix equation

$$\vec{w}^s = A^{-1} \vec{b}, \quad (4)$$

$$b_i = \sum_m \sum_t \lambda_m(t) \vec{e}(i, m)' \Sigma_m^{-1} (\vec{o}_t^s - \vec{\mu}_m)$$

$$a_{i,j} = \sum_m \sum_t \lambda_m(t) \vec{e}(i, m)' \Sigma_m^{-1} \vec{e}(j, m)$$

with  $\vec{b} = (b_1, b_2, \dots, b_K)'$  and  $A = \{a_{i,j}\}$ . A major issue associated with this method concerns the large number of training utterances required from each speaker for PCA estimation of the basis vectors. Another issue is the high dimensionality of these basis vectors as discussed in Section 1.

## 3. Clustered ML Linear Basis

To make the processes of training and adaptation in speaker space methods more efficient and robust, one can exploit the fact that there is a great deal of redundancy in the supervector representation described in Section 1. Section 3.1 describes an attempt to exploit this redundancy by forming equivalence classes of subvector means where an equivalence class  $\phi(m)$

for subvector mean  $\vec{\mu}_m$  is obtained through a clustering procedure. A ML procedure for estimating the basis vectors,  $\vec{e}(k)$ , is presented in Section 3.2 as a model based alternative to the PCA method for identifying a low dimensional subspace that captures relevant sources of variability.

### 3.1. Identifying Subvector Equivalence Classes

The speaker space methods can be described by a generative model for an observation vector,  $\vec{o}_t^s$ , uttered by speaker  $s$ . It is assumed that at each time  $t$  a mixture component  $m(t)$  for the speaker independent (SI) HMM is generated and the observation vector at time  $t$  corresponds to a speaker dependent variation about the subvector mean  $\vec{\mu}_{m(t)}$ . A SI residual error term,  $\vec{\epsilon}_{m(t)}$ , represents the observation error for the mixture resulting in the generative model,

$$\vec{o}_t^s = \vec{\mu}_{m(t)} + \sum_{k=1}^K w_k^s \vec{e}(k, m(t)) + \vec{\epsilon}_{m(t)}, \quad (5)$$

where  $\vec{\epsilon}_{m(t)} \sim \mathcal{N}(0, \Sigma_{m(t)})$ .

The generative model in Equation 5 can be made more robust by associating subvectors of the basis vectors with equivalence classes of the subvector means,  $\vec{\mu}_m$ , in the following steps. First, all subvectors of the supervector  $\vec{\mu}$  are clustered so that each subvector is associated with class,  $\phi(m)$ . Second, all subvectors  $\vec{e}(k, m)$  of the basis vectors in Equation 5 are associated with this equivalence class rather than with the individual Gaussian. This amounts to tying the basis subvectors as  $\vec{e}(k, \phi(m))$ , so that the generative model can be written as

$$\vec{o}_t^s = \vec{\mu}_{m(t)} + \sum_{k=1}^K w_k^s \vec{e}(k, \phi(m(t))) + \vec{\epsilon}_{m(t)}. \quad (6)$$

The number of physical subvectors,  $N$ , is much smaller than than the number of logical subvectors,  $M$ . It will be shown in Section 4 that a mapping of physical subvector to logical subvector where  $M/N \approx 10$  represents a reasonable trade-off between acoustic sensitivity and statistical robustness.

One mechanism for forming the equivalence classes described above is to cluster the supervector means using a binary splitting algorithm for clustering Gaussians that is similar to that used to form regression class trees [6]. This is a reasonable definition of equivalence class for SI HMM mean vectors since speaker dependent variation about subvectors within a cluster should have similar effect. The  $\vec{e}(k, \phi(m))$  will be referred to as a clustered maximum likelihood linear basis (CMLLB).

### 3.2. EM Algorithm for CMLLB

ML estimates of the basis vectors used in Equation 6 are obtained here using a variant of the EM algorithm. The procedure for estimating  $\vec{e}(k)$  is a generalization of that used in CAT to account for the class tying of the basis subvector means according to the equivalence classes described above [3]. Differentiating the auxiliary Q function given in Equation 2 with respect to  $\vec{e}(k, n)$  results in the expression

$$\frac{\partial Q(\Lambda, \hat{\Lambda})}{\partial \vec{e}(k, n)} = - \sum_{m \in \phi(n)} \sum_{s=1}^S \sum_{t=1}^T \lambda_m(t) (\vec{o}_t^s - \hat{\vec{\mu}}_m^s)' \Sigma_m^{-1} w_k^s \quad (7)$$

where  $n$  is the basis subvector index and  $\tilde{\phi}(n) = \{m : \phi(m) = n\}$ .

For the unclustered MLLB, the expression for the  $d$ th component of the  $m$ th basis subvector,  $\vec{e}(:, m, d) = (e(1, m, d), \dots$

$\cdot, e(K, m, d))'$  where  $e(k, m, d)$  is the  $d$ th component of  $\vec{e}(k, m)$ , can be obtained from Equation 7 as

$$\vec{e}(:, m, d)' = \sum_s \sum_t \lambda_m(t) (o_t^s(d) - \mu_m(d)) (\vec{w}^s)' C_m^{-1}, \quad (8)$$

$$C_m = \sum_s \sum_t \lambda_m(t) \vec{w}^s (\vec{w}^s)'. \quad (9)$$

Equation 7 can also be solved for the CMLLB case, where the equivalence classes  $\phi()$  are non-trivial. If a diagonal subvector covariance matrix,  $\Sigma_m$ , is assumed where  $\sigma_m(d)$  is the  $d$ th diagonal covariance component, then

$$\vec{e}(:, n, d)' = \sum_{m \in \tilde{\phi}(n)} \sum_s \sum_t \lambda_m(t) (o_t^s(d) - \mu_m(d)) \sigma_m^{-1}(d) (\vec{w}^s)' C_n^{-1}(d), \quad (10)$$

$$C_n(d) = \sum_{m \in \tilde{\phi}(n)} \sum_s \sum_t \lambda_m(t) \vec{w}^s \sigma_m^{-1}(d) (\vec{w}^s)'. \quad (11)$$

This expression can be used as part of an iterative procedure for obtaining the basis vectors that includes the following steps. First, the basis vectors are randomly initialized. Second, given the estimated basis vectors, the speaker dependent weight vectors are estimated for all speakers in the training set using Equation 4. Third, given estimates of the weight vectors, the basis vectors are computed using Equation 10. Steps two and three are repeated until there is no significant change in likelihood.

## 4. Experimental Study

This section presents an experimental study to evaluate the performance of the MLLB and CMLLB based adaptation procedures described in Section 3 on the RM and WSJ tasks. The adaptation scenarios of primary interest are those that require small amounts of adaptation speech which, in this study, will be assumed to be a single 2 to 5 second length utterance.

### 4.1. Adaptation and Training Scenarios

Performance comparisons were made with respect to speaker independent ASR on both the RM and WSJ tasks. The SI HMM training scenarios, baseline system configurations, and WERs are described below for each task. Comparisons were also made on the RM task to eigenvoice based speaker adaptation. The eigenvoice implementation, referred to here as  $E_{mltr}$ , follows the method proposed by Botterweck for speaker space adaptation of large vocabulary HMMs [5]. In the  $E_{mltr}$  training procedure, eigenvectors were computed by performing PCA analysis on SD models which were obtained from block diagonal MLLR adaptation of the SI model using utterances for each of the speakers in the SI-109 training set.

A single utterance based unsupervised adaptation scenario was investigated for both the RM and WSJ tasks. This involved a two-pass decoding strategy for each utterance. A hypothesized transcription was obtained using the SI model during the first pass. A weight vector,  $\vec{w}^s$ , was estimated from Equation 10 and used to adapt the SI model as described in Equation 1. The final result was obtained in a second decoding pass using the adapted model. A supervised adaptation scenario was also investigated for the RM task. From 1 to 3 transcribed speech utterances per speaker were used for estimating weight vectors and adapting SI models.

One of the main advantages of speaker space based adaptation algorithms is generally thought to be the fact that relevant interspeaker variability is captured in the basis vectors of

a low dimensional speaker space during training. These basis vectors are estimated from speaker specific supervectors using either PCA or ML estimation. However, it is also reasonable to assume that this low dimensional subspace can capture any of the many sources of intraspeaker variability as well. This issue is investigated here by modifying the algorithm described in Section 3 so that instead of estimating the basis vectors from speaker specific weight vectors as implied by Equation 10, the basis vectors are estimated from weight vectors obtained from each utterance. This corresponds to removing all speaker specific information from the training procedure, and implies that the technique is not simply a speaker space based representation. This will be referred to as an utterance based subspace representation.

The rest of this section will discuss the evaluation of four different variants of the MLLB based adaption algorithms. Both the unclustered (MLLB) case and clustered (CMLLB) case, will be considered. For CMLLB applied to the RM task the ratio of logical basis subvectors to physical basis subvectors was approximately 10, and for the WSJ task this ratio was approximately 20. Basis vector training scenarios that incorporate speaker dependent information in training (S) and that rely strictly on utterance dependent information (U) are considered. All systems are evaluated for subspace dimensionalities of  $K = 10$  and 20.

### 4.2. Adaptation on the RM task

Unsupervised and supervised adaptation was performed on the RM corpus under the following scenario. Acoustic SI HMMs were trained using 3990 utterances from 109 speakers taken from the standard RM SI-109 training set. The basis vectors for the  $E_{mltr}$  and the MLLB techniques were also trained from this 109 speaker training set. ASR WER was evaluated using 1200 utterances from 12 speakers taken from the RM speaker dependent evaluation (SDE) set. For supervised adaptation, adaptation utterances were randomly chosen for each of the 12 test speakers from the speaker dependent development (SDD) set. To ensure statistical robustness of the measured WER, six different adaptation sets were randomly chosen for each speaker and the WER obtained after adaptation to these data sets were averaged.

The baseline SI HMMs contained left-to-right 3-state state clustered triphones with 6 mixtures per state for a total of 10,224 Gaussians. The standard RM word-pair grammar language model was used. Feature analysis included 12 mel frequency cepstrum coefficients (MFCCs), normalized log energy, and their first and second difference coefficients for a 39-dimension feature vector. The baseline WER of the SI models on the 1200 utterance test set was 4.91%.

#### 4.2.1. Unsupervised adaptation experiments

The ASR WER for the unsupervised adaptation scenario evaluated on the RM test set are displayed in Table 1 for subspace dimensionalities equal to 10 and 20. The table shows the performance for baseline SI training, eigenvoice ( $E_{mltr}$ ) based adaptation, MLLB adaptation, and CMLLB adaptation using 1000 shared subvectors. For both MLLB and CMLLB, WER for speaker dependent (S) and utterance dependent (U) basis vector training as described in Section 4.1 are shown.

There are several observations that can be made from Table 1. The first observation is that there is a significant reduction in WER for CMLLB adaptation relative to the unclustered MLLB adaptation, and that the MLLB provides no improve-

ment over  $E_{mltr}$  adaptation. This is partially a result of the tendency of the ML based procedure that is used in MLLB for estimating the basis functions to “overfit” the training data. This effect is diminished in CMLLB through the use of clustered basis vectors. Second, there is no significant difference in WER for any of the techniques when the subspace dimensionality is increased from 10 to 20. Third, there is no significant difference in WER between the MLLB/CMLLB approaches that occurs as a result of using speaker dependent (S) or utterance dependent (U) basis vector training. This suggests that there is no need to collect large numbers of utterances for each speaker for the purpose of identifying basis vectors during training. Finally, The CMLLB approach results in as much as a 14.9% reduction in WER relative the SI system and a 7% reduction with respect to  $E_{mltr}$ . The statistical significance of the differences

Table 1: WER for unsupervised adaptation on the RM test set.

SubSpace Dimen.	SI	$E_{mltr}$	MLLB		CMLLB	
			S	U	S	U
10	4.91	4.53	4.62	4.54	4.21	4.30
20	4.91	4.48	4.53	4.51	<b>4.18</b>	4.23

in measured WER for the systems displayed in Table 1 was investigated using the suite of significance tests implemented by NIST [7]. It was found that the WER differences between the SI,  $E_{mltr}$ , and CMLLB systems were all statistically significant at a 5% level of significance according to the matched pair sentence segment test.

#### 4.2.2. Supervised adaptation experiments

Table 2 displays the WER obtained using supervised adaptation with from 1 to 3 adaptation utterances. A subspace dimensionality of  $K = 20$  was used for all systems. It is clear that CMLLB outperformed all other systems in Table 2 and that the differences in WER are most apparent when a single adaptation utterance was used. However, there was no significant reduction in WER for CMLLB as additional adaptation utterances were used. It is also clear that there was no significant difference for the CMLLB system between the case where speaker dependent information was used for training basis vectors (S) and when it was not used (U). Table 2 also displays the WER for block diagonal maximum likelihood linear regression (MLLR) based adaptation. While it was found that MLLR outperformed all of the subspace based techniques when greater than 5 adaptation utterances were used, MLLR obtained higher WER for all of the scenarios shown in the table.

Table 2: WER for supervised adaptation on the RM test set.

Adapt. Utt.	MLLR	$E_{mltr}$	MLLB		CMLLB	
			S	U	S	U
1	4.91	4.39	4.67	4.59	4.15	4.21
2	4.86	4.39	4.58	4.50	4.10	4.12
3	4.44	4.32	4.55	4.44	4.11	4.19

#### 4.3. WSJ evaluation

Unsupervised adaptation was performed on the WSJ task where the baseline SI system was configured as follows. The SI HMM’s were trained using 107,937 utterances from 988 speakers contained in the WSJ0, WSJ1, and TIMIT training sets. The basis vectors for the MLLB techniques were trained using the same training sets as the acoustic model. The ASR WER for the single utterance based unsupervised adaptation scenario described in Section 4.1 was evaluated using the Nov92 test set,

which contains 330 utterances. The baseline SI HMMs contained left-to-right 3-state state clustered triphone models with 16 mixtures per state for a total of 214,256 Gaussian densities. Recognition was performed using a 5k word vocabulary and the WSJ 5k bigram language model.

The ASR WER using unsupervised adaptation for the MLLB adaptation and CMLLB adaptation where each basis vector contains 10,000 shared subvectors are displayed in Table 3. Both speaker dependent (S) and utterance dependent (U) basis vector training as described in Section 4.1 are shown. A subspace dimensionality of  $K = 10$  was used for all of the adaptation experiments. It is clear from Table 3 that the unclustered MLLB based adaptation did not provide a significant reduction in WER relative to the SI system. However, CMLLB adaptation resulted in a WER reduction of 12.6%.

Table 3: WER for unsupervised adaptation on the WSJ test set.

SubSpace Dimen.	SI	MLLB		CMLLB	
		S	U	S	U
10	5.17	5.12	5.08	4.63	<b>4.52</b>

## 5. Conclusions

A subspace approach to speaker adaptation relying on clustered basis vectors, CMLLB, has been presented. The approach was applied to single utterance based unsupervised speaker adaptation on the RM and WSJ task domains. WER reductions of 14.9% and 12.6% relative to SI training were obtained for RM and WSJ respectively. Furthermore, it was found that the memory required for storing linear basis was reduced by over an order of magnitude with respect to unclustered subspace adaptation methods resulting in a savings of well over 100 Mbytes for the WSJ task. It was also shown that a variant of CMLLB requiring no speaker specific information for basis vector training performed as well as standard scenarios where a large number of utterances are required from each training speaker.

## 6. Acknowledgements

The authors would like to thank Parya Momayyez training baseline HMM models for the WSJ task.

## 7. References

- [1] P. C. Woodland, “Speaker adaptation for continuous density HMMs: A review”, ITRW on Adaptation Methods for Speech Recognition, 2001, 11-19.
- [2] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” IEEE Trans. Speech and Audio Proc., 8(4):695-707, 2000.
- [3] M. J. F. Gales, “Cluster adaptive training of hidden Markov models,” IEEE Trans. Speech and Audio Proc., 8(4):417-428, 2000.
- [4] P. Nguyen, C. Wellekens and J. C. Junqua, “Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments”, EUROSPEECH, 1999, pp. 2519-2522.
- [5] H. Botterweck, “Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices”, ICSLP, 2000, pp. 354-357.
- [6] M. J. F. Gales and P.C. Woodland, “Mean and variance adaptation within the MLLR framework”, Computer Speech and Language, 10(4):249-264, 1996.
- [7] “http://www.nist.gov/speech/tests/sigttests/sigttests.htm”.