Rapid Speaker Adaptation Using MLLR and Subspace Regression Classes

Kwok-Man Wong, Brian Mak

Department of Computer Science,
The Hong Kong University of Science and Technology,
Clear Water Bay, Hong Kong
mak@cs.ust.hk

Abstract

In recent years, various adaptation techniques for hidden Markov modeling with mixture Gaussians have been proposed, most notably MAP estimation and MLLR transformation. When the amount of adaptation data is limited, adaptation can be done by grouping similar Gaussians together to form regression classes and then transforming the Gaussians in groups. The grouping of Gaussians is often determined at the full-space level. In this paper, we propose to group the Gaussians at a finer acoustic subspace level. The motivation is that clustering at subspaces of lower dimensions results in lower distortion. Besides, as the dimension of subspace Gaussians reduces, there are fewer parameters to estimate for the subsequent MLLR transformation matrix. This is particular attractive in fast adaptation. Speaker adaptation experiments on the Resource Management task with few seconds of speech show that the use of subspace regression classes is more effective than traditional full-space regression classes.

1. Introduction

The main challenge of rapid speaker adaptation is to maximize the improvement in recognition performance with a very limited amount of adaptation data. Various adaptation techniques for hidden Markov modeling with mixture Gaussians have been proposed in recent years, most notably MAP estimation [1] and MLLR transformation [2]. When the amount of adaptation data is scarce, most of the Gaussians in the HMMs are unobserved. Hence, adaptation is usually done by grouping similar Gaussians together to form regression classes with each class having some observed Gaussians. In MLLR adaptation, Gaussians of the same regression class share the adaptation data to derive a transformation. The transformation is then applied to each member Gaussian in that class. The grouping of Gaussians is often determined at the full-space level. If the distribution of adaptation data over the full-space regression classes (FSRCs) is uneven, subsequent transformation is unreliable.

In this paper, we perform the grouping of Gaussians at a finer acoustic subspace level. That is, full-space Gaussians are projected onto orthogonal and disjoint subspaces, and the resulting subspace Gaussians are clustered to form subspace regression classes (SSRCs). The motivation is that clustering at subspaces of lower dimensions results in lower distortion; or in other words, for the same distortion, subspace clustering results in fewer regression classes. Consequently, the distribution of adaptation data over the fewer SSRCs can be more even than the corresponding distribution over FSRCs. Moreover, as the dimension of subspace Gaussians is reduced, there will be fewer estimating parameters for the subsequent MLLR transformation matrix. Although reduction in estimating parameters lowers the complexity of the transformation, when there are scarce adaptation data, fewer parameters are preferred for robust estimation.

In the next section, MLLR adaptation with subspace regression classes is outlined. This is followed by the comparison between FSRCs and SSRCs in Section 3. The evaluation and the conclusion will be presented in Section 4 and in Section 5 respectively.

2. MLLR adaptation with subspace regression classes

The MLLR adaptation with SSRCs involves three steps. First, full-space Gaussians from the original CDHMMs are projected onto the subspaces to produce subspace Gaussians. In each subspace, clustering is performed on the resultant subspace Gaussians to determine the SSRCs. Second, transformation matrix for each SSRC is derived and the subspace Gaussians are transformed accordingly. Finally, full-space Gaussians are re-constructed from the transformed subspace Gaussians and the resulting CDHMMs are used for recognition.

2.1. Derivation of Subspace Regression Classes

The subspace definitions and the derivation of SSRCs have been introduced in our previous work [3] and [4]. With SSRCs determined in each subspace, MLLR trans-

formation matrix for each SSRC can be estimated.

2.2. MLLR Transformation in Subspace

By maximizing the appropriate auxiliary function with respect to the transformation matrix, the following equation is obtained:

$$\sum_{t=1}^{T} \sum_{g_k \in R_k} L_{gt}\, \Sigma_{g_k}^{-1}\, o_{kt}\, \xi_{g_k}^{'} = \sum_{t=1}^{T} \sum_{g_k \in R_k} L_{gt}\, \Sigma_{g_k}^{-1}\, W_{R_k}\, \xi_{g_k}\, \xi_{g_k}^{'}$$

where

R_k	:	a SSRC constituted by a set of subspace Gaussians on the k-th subspace
g_k	:	the subspace Gaussian of a full-space Gaussian g on the k-th subspace
L_{gt}	:	the occupation likelihood of the original full-space Gaussian g at time t
Σ_{g_k}	:	the covariance of g_k
o_{kt}	:	the projection of the observation vector at time t on the k-th subspace
ξ_{g_k}	:	the extended mean of g_k
W_{R_k}	:	the transformation matrix for R_k

The transformation matrix W_{R_k} can be computed in a similar manner as in the case using FSRCs. Once the transformation matrices for the SSRCs in all subspaces are estimated, subspace Gaussians are transformed accordingly.

2.3. Re-construction of Full-space Gaussians

After the MLLR transformation, subspace Gaussians coming from the same original full-space Gaussian are concatenated together to construct a full-space Gaussian. Here, we make an assumption that all of the full-space Gaussians in the CDHMMs have diagonal or block diagonal covariance.

3. SSRC vs. FSRC

Suppose C regression classes are used in full-space MLLR approach. The number of transformation parameters is $\mathcal{D} \times (\mathcal{D}+1) \times C$, where $\mathcal{D}$ is the dimensionality of a feature vector. On the other hand, in subspace approach, if the original full vector space is split into K subspaces of equal dimensions and C regression classes are derived in each subspace, the number of transformation parameters is $\mathcal{D}/K \times (\mathcal{D}/K+1) \times C \times K$. Hence, the ratio of transformation parameters of FSRC to SSRC is

$$\frac{\mathcal{D} \times (\mathcal{D}+1) \times C}{\mathcal{D}/K \times (\mathcal{D}/K+1) \times C \times K} = \frac{(\mathcal{D}+1)K}{\mathcal{D}+K}$$

The number of transformation parameters can be reduced by approximately K times with the subspace approach when $\mathcal{D} \gg K$.

Reduction in transformation parameters can also be achieved in the full-space approach with the use of block-diagonal transformation matrices. The transformation matrix W_R can be decomposed into

$$W_R = [\ b\ A\ ]$$

where b is a bias vector of $\mathcal{D}$ and A is a $\mathcal{D} \times \mathcal{D}$ matrix. By setting A to be a block-diagonal matrix with K blocks, same reduction of transformation parameters can be achieved as in the subspace approach.

FSRC with block-diagonal transformation matrices can be thought as one kind of SSRC with the regression class membership of subspace Gaussians derived at the full-space level. That is, subspace Gaussians coming from full-space Gaussians which are in the same FSRC will be grouped to the same SSRC.

The advantage of SSRC over FSRC with block-diagonal matrices is that clustering at the subspace level is usually more effective resulting in lower distortion. Nevertheless, both approaches suffer from the problem that the correlation between subspaces is ignored during transformation which offsets part of the benefit from the reduced number of parameters.

4. Evaluation

The Resource Management (RM) [5] task is chosen to evaluate the effectiveness of SSRCs against FSRCs. A speaker-independent (SI) model is trained using the SI section of the database. Training data from test speakers in the speaker-dependent (SD) section are used in the following supervised adaptation experiments. Thirty-nine-dimensional acoustic vectors (consisting of 12 MFCCs and the normalized frame energy plus their first and second time derivatives) are produced from each 20ms of speech at a frame rate of 100Hz. Instead of exploiting a regression class tree, we pre-determine the number of regression classes to be used, so that we can make a direct comparison between the SSRC approach and the FSRC approach under the same number of regression classes. Adapted models are tested on the Feb91-SD test set consisting of 300 utterances from 12 test speakers using the standard RM word-pair grammar (perplexity = 60).

4.1. Training of Baseline SI Model

The baseline SI model is trained from the augmented SI training set of the RM database which consists of 3990 utterances from 109 speakers. All speech units are word-internal triphones which are 3-state left-to-right HMMs, with a maximum of 16 mixture Gaussians per state. From the 2229 distinct word-internal triphones and 41 basis phones present in the training data, 865 triphone HMMs are estimated with 543 tied-states and 5349 Gaussians.

The SI model achieves a word accuracy of 95.61% and 94.42% on the Feb91-SI and Feb91-SD test sets respectively.

4.2. Speaker Adaptation Experiments

As we focus only on fast adaptation, small adaptation data sets ranging from 2s to 60s are used. To account for the variability of small data, three different collections of adaptation data sets are prepared for each test speaker. For each test speaker and for each of his collections, adaptation data sets of four different sizes are prepared: 2s, 10s, 30s and 60s, which are randomly selected from his SD training data; and smaller adaptation data sets are subsets of the larger ones. Thus, there are totally 36 data sets for each size of adaptation data. All experimental results are reported on the average of all 12 test speakers.

In our evaluation, the following adaptation schemes were investigated:

1. FS-RC$*N*: MLLR adaptation using $*N* full-space regression classes
2. FS-B$*K*-RC$*N*: MLLR adaptation using $*N* full-space regression classes with block-diagonal transformation matrices of *K* blocks
3. SS-S$*K*-RC$*N*: MLLR adaptation using $*N* subspace regression classes for each of the $*K* subspaces

The full-space and subspace definitions are as follows:

Full-space Definition:

$$\text{full-space} = 12\text{MFCC} + e + 12\Delta\text{MFCC} + \Delta e + 12\Delta\Delta\text{MFCC} + \Delta\Delta e$$

3-Subspace Definition:

$$\text{Subspace 1} = 12\text{MFCC} + e$$
$$\text{Subspace 2} = 12\Delta\text{MFCC} + \Delta e$$
$$\text{Subspace 3} = 12\Delta\Delta\text{MFCC} + \Delta\Delta e$$

13-Subspace Definition:

$$\text{Subspace 1-12} = 12\text{MFCC}_i + 12\Delta\text{MFCC}_i + 12\Delta\Delta\text{MFCC}_i, \quad 1 \le i \le 12$$
$$\text{Subspace 13} = e + \Delta e + \Delta\Delta e$$

20-Subspace Definition: each subspace holding the most correlated feature pair and the last subspace holding the remaining feature.

4.3. Results and Discussion

The speaker adaptation results using MLLR with FSRCs and SSRCs are shown in Figure 1-2.

As illustrated in Figure 1, when the amount of adaptation data is extremely scarce (∼2s), MLLR with one

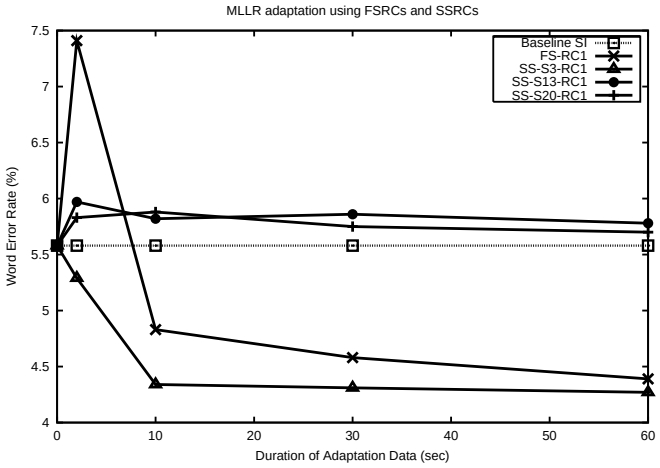

Figure 1: Recognition results of MLLR adaptation using FSRCs and SSRCs with global transformation performed in each subspace

FSRC shows degradation from the baseline SI model and also the worst performance among all the adaptation schemes. This is simply due to the poorly estimated transformation matrix. On the other hand, with the same amount of adaptation data, MLLR using three subspaces and one SSRC per subspace can show improvement over the baseline and the full-space MLLR approach. This indicates that reduction in transformation parameters leads to a more robust estimation of the transformation of the regression classes when adaptation data are scarce. However, MLLR with 13 and 20 subspaces always shows degradation from the baseline. This can be explained by the loss of correlation between subspaces. The larger the number of subspaces, the greater is the adverse effect caused by the loss of correlation.

As MLLR with three subspaces achieved the highest accuracy in the previous experiment, we conducted another experiment to compare the performance of MLLR using SSRCs with 3 subspaces against MLLR using FSRCs with 3-block diagonal transformation matrices. Adaptation using 2, 4 and 8 regression classes was attempted in each scheme. The best result among adaptation with different numbers of regression classes was reported for each scheme in Figure 2.

The plot illustrates that MLLR using SSRCs still shows the best performance over the other two FSRC approaches for various sizes of adaptation data. The experimental results of adaptation using SSRCs with three subspaces and FSRCs with 3-block diagonal transformation matrices are tabulated in Table 1.

It should be noticed that SS-S3-RC1 and FS-B3-RC1 should have the same performance as there is only one regression class in each subspace. The groupings of the subspace Gaussians in these two cases have to be the same. The comparison shows that MLLR us-

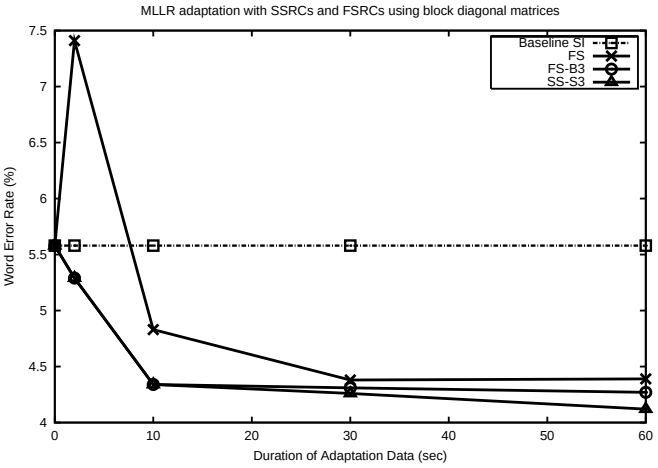

Figure 2: Recognition results of MLLR adaptation using SSRCs and FSRCs with block diagonal matrices

Method	2 sec.	10 sec.	30 sec.	60 sec.
SS-S3-RC1	5.29	4.34	4.31	4.27
SS-S3-RC2	6.03	4.56	4.26	4.22
SS-S3-RC4	8.13	4.76	4.32	4.26
SS-S3-RC8	24.98	4.94	4.28	4.12
FS-B3-RC1	5.29	4.34	4.31	4.27
FS-B3-RC2	6.06	4.50	4.50	4.41
FS-B3-RC4	9.18	4.82	4.50	4.47
FS-B3-RC8	26.82	5.48	4.40	4.35

Table 1: Adaptation results of SSRCs with three subspaces and FSRCs with 3-block diagonal transformation matrices

ing SSRCs always outperforms MLLR using FSRCs with block diagonal matrices for a given number of regression classes and a given amount of adaptation data. Thus, the effectiveness of SSRCs over FSRCs is proven.

5. Conclusion

In this paper, we propose to perform MLLR adaptation with subspace regression classes. Experimental results show that MLLR with SSRCs achieve better performance than with conventional full space MLLR approach when the amount of adaptation data is limited. The introduction of SSRCs reduces the number of transformation parameters in the adaptation process so that the transformation matrices can be estimated more robustly given a few adaptation utterances. Derivation of regression classes in the subspace level is also shown to be more effective than in the full-space level. However, the use of SSRCs will lead to the loss of correlation between subspace features during subsequent transformation. In practice, one has to balance between the two effects — parameter reduction and loss of correlation. Empirically we find that for the common 39-dimensional feature vector, using three streams gives good adaptation performance.

6. Acknowledgements

This work is supported by the Hong Kong RGC under the grant number CA97/98.EG02, and by the grant HKTIIT 98/99.EG01 from the Cable & Wireless HKT.

7. References

[1] Jean-Luc Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[2] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Journal of Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.

[3] E. Bocchieri and B. Mak, "Subspace Distribution Clustering for Continuous Observation Density Hidden Markov Models," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997, vol. 1, pp. 107–110.

[4] B. Mak, E. Bocchieri, and E. Barnard, "Stream Derivation and Clustering Schemes for Subspace Distribution Clustering HMM," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 1997, pp. 339–346.

[5] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988, vol. 1, pp. 651–654.