MAP ADAPTATION WITH SUBSPACE REGRESSION CLASSES AND TYING

Kwok-Man Wong and Brian Mak

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

ABSTRACT

In the hidden Markov modeling framework with mixture Gaussians, adaptation is often done by modifying the Gaussian mean vectors using MAP estimation or MLLR transformation. When the amount of adaptation data is scarce or when some speech units are unseen in the data, it is necessary to do adaptation in groups — either with regression classes of Gaussians or via vector field smoothing. In this paper, we propose to derive regression classes of subspace Gaussians for MAP adaptation. The motivation is that clustering at the finer acoustic level of subspace Gaussians of lower dimension is more effective, resulting in lower distortions and relatively fewer regression classes. Experiments in which context-dependent TIMIT HMMs are adapted to the Resource Management task with few minutes of speech show improvement of our subspace regression classes over traditional full-space regression classes.

1. INTRODUCTION

When there is a mismatch between the training and testing conditions, adaptation of speech models can effectively improve recognition in the new environment. In the hidden Markov modeling framework with mixture Gaussians, adaptation is often done by modifying the Gaussian mean vectors using MAP estimation [3] or MLLR transformation [4]. When the amount of adaptation data is adequate, individual models may reliably be modified using either approaches but better with MAP estimation. However, when the amount of adaptation data is scarce or when some speech units are unseen in the data, it is necessary to do adaptation in groups as in MLLR adaptation and vector field smoothing [7]. That is, similar Gaussians are grouped into regression classes so that Gaussians in the same class will share the same adaptation data in deriving the transformation.

In acoustic modeling, it is important to balance model complexity with the amount of training data. The same heuristic also applies to the task of adaptation. The use of regression classes of Gaussians reduces adaptation complexity when the amount of adaptation data is small. In the literature, these regression classes are usually derived by clustering the full-space Gaussians in the models. In this paper, we investigate the derivation of regression classes from subspace Gaussians for adaptation. The motivation is that clustering of subspace Gaussians of lower dimension is more effective, resulting in lower distortions and relatively fewer regression classes.

In the next section, we describe how we derive the subspace regression classes. This is followed by the evaluation in Section 3 and conclusions in Section 4.

2. DERIVATION OF SUBSPACE REGRESSION CLASSES

Recently we have proposed an alternative to conventional continuous-density hidden Markov modeling (CDHMM) which we call *subspace distribution clustering HMM* (SDCHMM) [1]. In SDCHMM, mixture Gaussians with diagonal covariances are first projected into low-dimensional subspaces (usually one- to three-dimensional), and subsequent subspace Gaussians are tied. Because of the efficiency achieved by clustering in low dimensions, SDCHMM attains a high degree of tying — thus very compact — without degradation in performance when compared to its original CDHMM. In our experience, for a common 39-dimensional acoustic vector, 13 to 20 streams with 32 to 256 subspace Gaussian prototypes per stream are adequate to give good performance.

The subspace definitions and the method of subspace Gaussian tying in SDCHMM [5] can be applied to the derivation of subspace regression classes for adap-

This work is supported by the grant HKTIIT 98/99.EG01 from the Cable & Wireless HKT.

0-7803-6293-4/00/$10.00 ©2000 IEEE.

tation purpose. We will briefly describe them here for the sake of completeness.

2.1. Definition of Subspaces

We use the heuristic that correlated features tend to cluster in a similar manner and require each subspace to comprise the most correlated features. For 2-dimensional subspaces, one may use Pearson's moment product correlation coefficient to determine the most-correlated subspaces:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \tag{1}$$

where σ_i and σ_j are the standard deviations of the i-th and j-th feature respectively, and σ_{ij} is the square root of their covariance. For multiple correlation among $k(k > 2)$ features, we use the following measure:

$$R = 1 - \begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2k} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \cdots & 1 \end{vmatrix}. \tag{2}$$

2.2. Creation of Subspace Regression Classes

Full-space observation Gaussians are first projected into each subspace, then regression classes of subspace Gaussians are created by a modified K-means clustering algorithm using the Bhattacharyya distance measure between Gaussians as in Algorithm 1.

3. EVALUATION: TASK ADAPTATION

To evaluate the effectiveness of the subspace regression classes, we adapted TIMIT [8] phoneme models to recognize utterances from the Resource Management (RM) [6] task. All speech models are word-internal triphones which are 3-state left-to-right HMMs, with a maximum of 32 mixture Gaussians per state. The signal processing front-end produces 39-dimensional acoustic vectors (consisting of 12 MFCCs and the normalized frame energy plus their first and second time derivatives) from each 20ms of speech at a frame rate of 100Hz.

3.1. Training of Baseline RM HMMs

For comparison, a baseline RM recognizer was trained from the augmented speaker-independent training set of the RM speech corpora which consists of 230 minutes of speech and 3990 utterances. From the 2229 distinct word-internal triphones and 43 basis phones present in the training data, 869 triphone HMMs are estimated with 535 tied-states and 6188 Gaussians. The models achieve a word accuracy of 91.34% on the Feb'91 test set.

Algorithm 1: Derivation of subspace regression classes

Goal: To derive N regression classes of subspace Gaussians in each of the K subspaces.

Step 1. *Initialization:* First create a Gaussian mixture model with N components from the training data. Project each of the N Gaussian components onto the K subspaces according to a given K-subspace definition. The resultant KN subspace Gaussians will be used as initial subspace Gaussian centroids.

Step 2. Similarly project each Gaussian in the original CDHMMs onto the K subspaces.

Step 3. For each subspace, repeat Step 4 & 5 until some convergence criterion is met.

Step 4. *Membership*: Associate each subspace Gaussian with its nearest centroid as determined by their Bhattacharyya distance. The collection of member subspace Gaussians of the same centroid constitute a regression class.

Step 5. *Update*: New centroid for each regression class is computed by merging all subspace Gaussians in the class.

3.2. Training of TIMIT Triphone HMMs

Context-dependent triphone HMMs were first estimated from the standard TIMIT training dataset which contains 188 minutes of 3696 utterances. There are 6534 distinct word-internal triphones and 47 basis phones out of which, 1260 triphone HMMs are estimated with 578 tied-states and 5186 Gaussians. The models achieve a TIMIT phoneme accuracy of 64.03%, and a word accuracy of 79.20% on the RM Feb'91 test set.

3.3. Adaptation of TIMIT HMMs to RM

Adaptation data sets of various sizes: 1, 2, 5, 10, 15 and 20 minutes were randomly selected from the RM training corpus so that (1) each set is gender-balanced; and, (2) smaller adaptation data sets are subsets of the larger ones. The following adaptation schemes were investigated:

1. CDHMM-MAP: conventional MAP adaptation with no regression classes. Unseen triphones are not modified.

2. MLLR (RC$N-MLLR): HTK's [2] MLLR transformation with $N regression classes.

3. MLLR with MAP (RC$N-MLLR-MAP): HTK's MAP adaptation on MLLR-transformed mean vectors with $N regression classes.

4. Subspace-MAP (S$K-RC$N-MAP): MAP adaptation with $N subspace regression classes for each of the $K subspaces plus tying of the subspace Gaussians in each class.

5. Subspace-ML (S$K-RC$N-ML): Maximum-likelihood re-estimation of subspace Gaussians in each of the $K subspaces, using data shared among those belonging to the same subspace regression class (and there are $N regression classes per subspace). Both mean and variance of the subspace Gaussians are re-estimated.

The following remarks are worth to mention:

- HTK adaptations actually run in two passes: A global MLLR transformation is performed in the first pass, and the adapted models are used to re-align all adaptation data. The required adaptation scheme runs in the second pass using the re-aligned adaptation data. All other adaptation schemes run with one pass.

- To further reduce the amount of adaptation data requirement, all subspace Gaussians within a subspace regression class are tied in the Subspace-MAP scheme. On the other hand, by definition, Subspace-ML estimation renders the subspace Gaussians of a regression class effectively tied.

- All adaptations are done in supervised mode.

- All MAP adaptations use a scaling factor of 15.

3.4. Results and Discussion

The task adaptation results with 128 and 256 regression classes are tabulated in Table 1 and 2. We have the following observations:

- Even without any adaptation and simply by tying the subspace Gaussians in each subspace regression class, the word error rate (WER) is reduced by 11.2% and 14.5% respectively with the use 128 and 256 classes. A possible reason is that the context-dependent triphone models are very sharp and capture the TIMIT phone characteristics too well that they do not match the RM phones. The subspace Gaussian tying generalizes the models effectively by smoothing the Gaussians.

- With less than 5 minutes of adaptation data, MAP adaptation with regression classes of subspace Gaussians outperforms other schemes. The good performance should be due to the more effective clustering of subspace Gaussians of lower dimensions in deriving regression classes.

- However, with more adaptation data, the great reduction of model parameters entailed by Subspace-MAP or Subspace-ML limits model improvement. The limiting effect is smaller with Subspace-ML in which the variances are adapted as well. This is evidenced by (1) their better performance when the number of subspace regression classes was increased from 128 to 256; and (2) the better performance of Subspace-ML over Subspace-MAP with more than 5 minutes of adaptation data.

- As expected, conventional MAP adaptation does not perform well with limited amount of data.

4. CONCLUSIONS

In this paper, we demonstrate the effectiveness of clustering at a finer acoustic level of subspace Gaussians. Subspace MAP adaptation with regression classes derived from these subspace Gaussians achieves better performance than conventional MAP or MLLR when the amount of adaptation data is limited. The additional tying we introduce to the adaptation scheme is both a blessing and a hindrance: on the one hand, it further reduces the adaptation complexity and thus the amount of adaptation data required; on the other hand, it also limits the performance when the amount of adaptation data increases. Anyhow since adaptation with very limited amount of data is usually more important, Subspace-MAP with tying is still preferred and one may remove adaptation tying when data become abundant.

5. REFERENCES

[1] 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*, volume 1, pages 107–110, 1997.

Table 1: Adaptation results with 128 regression classes (* indicates the best performance in the group.)

Method	0 min.	1 min.	2 min.	5 min.	10 min.	15 min.	20 min.
CDHMM-MAP	79.20	80.68	81.12	83.13	84.66	84.86	85.67
RC128-MLLR	79.20	81.48	81.32	83.17	86.11	85.91	87.00*
RC128-MLLR-MAP	79.20	81.20	81.64	83.94*	87.00*	86.71*	86.55
S20-RC128-ML	81.52*	78.66	81.74	82.39	84.25	84.46	84.66
S20-RC128-MAP	81.52*	83.25*	82.49*	83.09	83.41	82.85	83.62

Table 2: Adaptation results with 256 regression classes

Method	0 min.	1 min.	2 min.	5 min.	10 min.	15 min.	20 min.
CDHMM-MAP	79.20	80.68	81.12	83.13	84.66	84.86	85.67
RC256-MLLR	79.20	81.48*	81.32	83.17	86.15	85.71	86.92*
RC256-MLLR-MAP	79.20	81.20	81.64	83.94	87.00*	86.80*	86.80
S20-RC256-ML	81.52*	81.24	79.55	84.22	86.11	86.03	86.35
S20-RC256-MAP	81.52*	81.40	82.37*	84.38*	84.90	85.23	85.27

[2] Steve Young et al. *The HTK Book (for HTK Version 2.2)*. Entropic Ltd., 1999.

[3] 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*, 2(2):291–298, April 1994.

[4] 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*, 9(2):171–185, April 1995.

[5] 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*, pages 339–346, 1997.

[6] 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*, volume 1, pages 651–654, 1988.

[7] M. Tonomura, T. Kosaka, and S. Matsunaga. Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum A Posteriori Probability Estimation. *Journal of Computer Speech and Language*, 10:117–132, 1996.

[8] V Zue, S. Seneff, and J. Glass. Speech Database Development at MIT: TIMIT and Beyond. *Speech Communication*, 9(4):351–356, August 1990.