IMPROVED COVARIANCE MODELING FOR MAXIMUM LIKELIHOOD MULTIPLE SUBSPACE TRANSFORMATIONS

Xi Zhou[1], *Ye Tian*[2], *Jian-lai Zhou*[2], *Bei-qian Dai*[1]

[1]Dept. of Electronic Science and Technology, University of Science and Technology of China

[2]Microsoft Research Asia, Beijing, China

zhouxi@mail.ustc.edu.cn, { t-yetian ,jlzhou }@microsoft.com

ABSTRACT

Maximum likelihood multiple subspace transformations algorithms, such as Semi-Tied Covariance (STC) and multiple Heteroscedastic Linear Discriminant Analysis (HLDA), have achieved significant improvement. In STC and multiple HLDA, all the Gaussian components are classified as multiple *components sets*. In each set, Gaussian components' full covariance, which is estimated by Maximum Likelihood (ML) criterion, is used to estimate the linear transformation of this set. However, full covariance matrix, which contains large number of free parameters, may not be reliably estimated by ML criterion. Unreliable full covariance will lead to unreliable linear transformation, and will finally lead to poor recognition results. There have been several algorithms proposed to reliably estimate the full covariance, such as mixture of inverse covariance (MIC), SPAM, and Hierarchical Correlation Compensation (HCC). In this paper, we combine HCC with STC and multiple HLDA. Experiments show that standard STC can achieve 12.47% word error rate (WER) reduction on RM database, while our HCC+STC can achieve 19.32% WER reduction.

1. INTRODUCTION

Diagonal covariance matrix implies strong assumption that the feature components are independent. In speech recognition, even Gaussian mixtures with diagonal covariance can model the correlation to some extent; the model precision is still limited. To overcome this problem, feature-space based linear transformation is used for decorrelating the feature. Feature-space based linear transformation includes the Karhunen-Loeve transform [1], linear discriminate analysis (LDA) [2], Maximum Likelihood Linear Transform (MLLT) [3] and heteroscedastic LDA (HLDA) [4]. However, it is hard to find a unique transform which can de-correlate all the features for all the classes in some complex tasks such as LVCSR. It is better to use model-based de-correlation, which allows multiple transforms to be used. Two model-based schemes, semi-tied covariance matrix (STC) [5] and multiple HLDA [6], have been proposed.

STC and multiple HLDA provide a good framework to apply linear transformation tying at any level. In STC and multiple HLDA, all the Gaussian components are classified as multiple components sets. In each set, full covariance of Gaussian components, which is estimated by Maximum Likelihood (ML) criterion, is used to estimate the linear transformation of this set. In this paper, we tie linear transformations at four different levels: global, monophone, monophone-state and tied-triphone state. It is expected that more transforms lead to higher accuracy because the model with more transforms is more precise. However, full covariance matrix, which contains large number of free parameters, may not be reliably estimated by ML criterion. Unreliable full covariance will result in poor estimation of linear transformation, and will finally lead to higher error rate.

Several approaches have been proposed to estimate full precision matrices based on a linear combination of a set of global prototype full precision matrices, such as mixtures of inverse covariances (MIC) [7], SPAM [8], modeling covariance by basis expansion [9]. Lin et al [10] also proposed a Hierarchical Correlation Compensation (HCC) algorithm to reliably estimate the full covariance of all Gaussian components. The experiment showed that all of them could result in robust estimation of the full covariance for Gaussian components. In this paper, we combine Hierarchical Correlation Compensation (HCC) with STC and multiple HLDA. In our experiments, the standard STC and multiple HLDA can achieve 14.2% error rate reduction on RM database, while our HCC+STC/multiple HLDA can achieve 23.9% error rate reduction.

The paper is organized as following: Section 2 describes two model-based linear transformations STC

and multiple HLDA. In section 3, we combined HCC with STC and multiple HLDA to get better transforms, and corresponding experiments results are shown in section 4. Finally, in section 5, the conclusion and the future work are presented.

2. STC AND MULTIPLE HLDA

In this section, we briefly introduce STC and multiple HLDA. In semi-tied covariance, each covariance matrix consists two elements, a component specific diagonal covariance element $\breve{\Sigma}_{diag}^{(m)}$, and a semi-tied class-dependent linear transformation $F^{(\gamma_m)}$. The form of the m 'th Gaussian components' full covariance matrix represented as

$$\breve{\Sigma}^{(m)} = F^{(\gamma_m)}\Sigma_{diag}^{(m)}F^{(\gamma_m)T} \tag{1}$$

$F^{(\gamma_m)}$ can be tied in different level.

Maximum likelihood criterion is used to estimate the STC transform matrix. Firstly, define the mean and covariance of the m 'th Gaussian components as

$$u^{(m)} = \frac{\sum_\tau \gamma_m(\tau) o(\tau)}{\sum_\tau \gamma_m(\tau)} \tag{2}$$

$$W^{(m)} = \frac{\sum_\tau \gamma_m(\tau)\left(o(\tau)-u^{(m)}\right)\left(o(\tau)-u^{(m)}\right)^T}{\sum_\tau \gamma_m(\tau)} \tag{3}$$

,where $\gamma_m(\tau) = p\left(q_m(\tau) \mid M, O_T\right)$ is the posteriori probability of observation $o(\tau)$ belonging to component m .

Linear transformation $F^{(\gamma_m)}$ is estimated by maximize the auxiliary function

$$Q(M,\hat{M}) = \sum_{\gamma, m\in G^{(\gamma)}} \gamma_m^{(\tau)} \log\left(\frac{\left|A^{(\gamma)}\right|^2}{\left|diag(A^{(\gamma)}W^{(m)}A^{(\gamma)T})\right|}\right) \tag{4}$$

where $A^{(\gamma_m)} = F^{(\gamma_m)-1}$, and $G^{(\gamma)}$ is the components set that shares the same linear transformation. The details of estimation formulae for $A^{(\gamma)}$ can be found in [5].

Once the linear transformations is available, the covariance used for decoding is given by

$$\Sigma_{diag}^{(m)} = diag\left(A^{(\gamma)}W^{(m)}A^{(\gamma)T}\right) \tag{5}$$

For multiple HLDA, the feature space is split into two subspaces after transformation; the *useful dimensions* and *nuisance dimensions.* Class-specific covariance matrix is used for useful dimensional, and a simple single Gaussian component nuisance model is used for nuisance dimensions.

ML estimation is used to estimate the multiple HLDA transforms. The auxiliary function for multiple HLDA [6] is

$$Q(M,\hat{M}) = \sum_{\gamma, m\in G^{(\gamma)}} \gamma_m^{(\tau)} \log\left(\frac{|A|^2}{\left|diag(A_p W^{(m)} A_p^T)\right|\left|diag(A_{n-p} T A_{n-p}^T)\right|}\right) \tag{6}$$

, where p is the useful dimension, $A^{(r)}$ is linear transformation,

$$A^{(\gamma)} = \begin{bmatrix} A_{[p]}^{(\gamma)} \\ A_{[n-p]}^{(\gamma)} \end{bmatrix} \tag{7}$$

, T is the global full covariance matrix of the class

$$T = \frac{1}{N}\sum_\tau \left(o(\tau)-\mu^{(g)}\right)\left(o(\tau)-\mu^{(g)}\right)^T \tag{8}$$

, $\mu^{(g)}$ is the global mean of this class. The detail of estimation formulae for A can be found in [6].

For both of STC and multiple HLDA, the component specific full variance matrix $W^{(m)}$ plays an important role in estimating transforms. There is an assumption that if the full covariance matrix can not be reliably estimated, then the linear transformation will not be robust. We proved the assumption by the experiments which will be presented in section 4.

3. HIERARCHICAL COVARIANCE COMPENSATION

In STC and multiple HLDA, the full covariance of Gaussian components, $W^{(m)}$, is estimated by Maximum Likelihood (ML) criterion. However, full covariance matrix, which contains large number of free parameters, may not be reliably estimated by ML criterion. Unreliable full covariance will result in poor estimation of linear transformation, and will finally lead to higher error rate.

Several approaches have been proposed to directly estimate full precision matrices, such as mixtures of inverse covariances (MIC) [7], SPAM [8], modeling covariance by basic expansion [9]. Lin et al [10] also proposed a Hierarchical Correlation Compensation (HCC) algorithm to reliably estimate the full covariance Gaussian components. Experiments showed all of them could reliably estimate the full covariance of Gaussian components.

The general idea in HCC is to build a hierarchical tree in the covariance space, and use each leaf node to represent a Gaussian component in the model set. Since there are no enough data in each leaf node to estimate full covariance, a linear combination is employed to represent the full covariance in leaf nodes by covariance matrix of all its parent nodes.

The outline of HCC is:

1. Train a baseline model set of tri-phone CDHMMs with diagonal covariance matrices. The mean and the covariance are estimated using maximum likelihood criterion. We will keep the structure, the mixture weight, and the mean vectors of the baseline model set unchanged in following stages.
2. All the tied- states are used to build a tree. The tree can be built according to the full covariance's K-L distance with the top-down clustering. Or we can use the decision tree generated from the previous baseline model training stage. We use tied-states as base elements in tree-building since the full covariance matrices of Gaussian components may not be reliable for the clustering. After the tied-state tree is built, for each tied-state node we expand all its Gaussian components as another layer of its child.
3. Estimate a covariance matrix for each node in the tree. For all leaf nodes, we estimate diagonal covariance matrices. For each upper-level node, a full covariance matrix is estimated from all of its child nodes.
4. For each Gaussian component in a leaf node, the estimated full covariance matrices of all the nodes along the upward path from the leaf node to the global are used to estimate the off-diagonal components in its full covariance matrix. Based on a linear combination scheme, where the combination weights are estimated by the maximum likelihood criterion.

For the i th Gaussian component, all intermediate nodes along the upward path from this node to the root is defined as the set

$$\Psi(i) = \left\{ \begin{matrix} i's \text{ parent} , i's \text{ parent' s parents} , \\ ...., ..., root \end{matrix} \right\} \quad (9)$$

Thus the new full covariance $\hat{\Sigma}_i$ of the i th Gaussian components is estimated by

$$\hat{\Sigma}_i = diag(\Sigma_i) + \sum_{m \in \Psi(i)} \lambda_{i,m} \left[\Sigma_{node,m} - diag(\Sigma_{node,m}) \right]$$

After above four steps, we obtained the new full covariance matrices.

To combine HCC with STC and multiple HLDA, we can simply apply the full covariance estimated by HCC in the objection function of STC (Equ 4) and multiple HLDA (Equ 6).

4. EXPERIMENTS

4.1 Experiment setup

A standard speech recognition task, the DARPA Resource Management (RM) task, is used. A total of 3990 sentences is used for training. The baseline system uses mixture Gaussian densities with 1603 tied HMM states determined by standard decision tree. Cross word triphone models that ignore the word boundaries in the context are used. The baseline system is produced by standard iterative mixture splitting using four embedded training per mixture configuration. 6 mixture components with diagonal covariance are trained for every tied-triphone state. A total of 1199 sentences with a simple word-pair grammar are used for decoding and the word error rate of the baseline system is 4.09%.

4.2 Tying on different levels

STC and multiple HLDA provide good frameworks to tie the linear transformation at different levels. Here we use four different tying methods:

- Global: all the Gaussian components share the same transformation matrix.
- Monophone: all the Gaussian components belong to the same monophone share the same transformation matrix. There are totally 49 monophone.
- Monophone state: There are 3 states for each monophone (except silence and sp), there are totally 143 monophone states.
- Tied-triphone state: Tied-triphone stated is the leaf nodes of the tree-based clustering. There are totally 1603 tied-triphone states, so totally 1603 linear transformation matrices are used different tying is also shown in Fig 1.

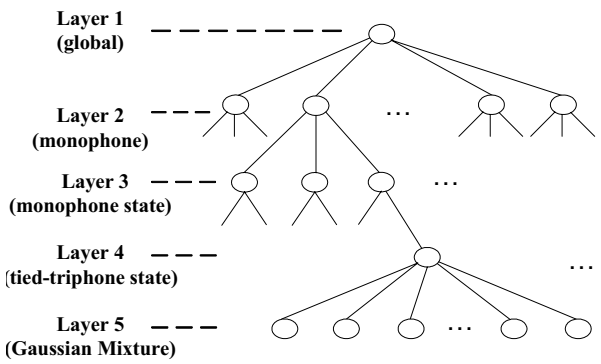

Figure 1. Tree structure: HMM structure

4.3 STC+HCC

Table 1 shows the performance of STC+HCC at different levels, comparing with that of STC only. If we only using global transform, the word error rate (WER) of calculating $W^{(m)}$ directly by (4) is 3.94%, and that of using HCC to calculate new full covariance is 3.91%. There is no main difference. However, with the increase of linear transforms number, the performance of STC+HCC significantly outperforms that of STC. When we tie the linear transformations at tied-triphone state, the word error rate of STC is 5.74%, which is much higher

than that of baseline. However, we get 3.30% word error rate (19.32% ERR reduction) by using STC+HCC at the same level. That means the performance improved consistently with the increase of linear transformation number.

Table 1. Word error rate of STC and STC+HCC on RM database (The baseline word error rate 4.09%)

different tying	transform number	word error rate	
		STC	STC+HCC
Global	1	3.94%	3.91%
Monophone	49	3.79%	3.77%
Monophone state	143	3.59%	3.40%
Tied-triphone state	1603	5.74%	3.30%

4.4 Multiple HLDA+HCC

Table 2 shows the performance of multiple HLDA+HCC comparing to multiple HLDA. When we tie the linear transformations at tied-triphone state, the word error rate of multiple HLDA is 5.62%, which is much higher than that of baseline. However, we get 3.50% word error rate (14.43% ERR reduction) by using multiple HLDA+HCC at the same level. We can also that the performance improved consistently with the increase of linear transformation number.

Table 2. Word error rate of multiple HLDA and multiple HLDA+HCC on RM database (source dimensions is 39 and destination dimensions is 30) , The baseline word error rate 4.09%.

different tying	transform number	word error rate	
		MHLDA	MHLDA+HCC
Global	1	3.85%	3.99%
Monophone	49	3.9%	3.79%
Monophone state	143	3.58%	3.65%
Tied-triphone state	1603	5.62%	3.50%

5. CONCLUSIONS AND FUTURE WORK

In this paper, we compared the performances of STC and multiple HLDA when tying the transformations at different levels. By combining HCC algorithm with STC and multiple HLDA, we improved the robustness of linear transformation, especially when the number of linear transformation matrices is large.

Experiments show that standard STC can achieve 12.47% error reduction on RM database, while our HCC+STC can achieve 19.32% error rate reduction. In the future, we will the test the effective of this combination on larger database.

6. REFERENCES

[1] K. Fukunaga, “Introduction to Statistical Pattern Recognition”. New York: Academic, 1972.

[2] R.Haeb-Umbach, “Linear discriminant analysis for large vocabulary speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, San Francisco, 1992, pp. 13-16

[3] R. Gopinath, “Maximum likelihood modeling with Gaussian distributions for classification,” in *Proc. ICASSP*, vol. II, 1998, pp. II-661–II-664.

[4] N. Kumar, “Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition,” *Ph.D. dissertation*, Johns Hopkins Univ., Baltimore, MD, 1997.

[5] M. J. F. Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 272–281, 1999.

[6] M. J. F. Gales, “Maximum Likelihood Multiple Subspace Projections for hidden Markov models,” *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 37–47, 2002.

[7] Vanhoucke, V.; Sankar, A.; "Mixtures of inverse covariance", *IEEE Trans. Speech Audio Processing*, vol. 12 , pp. 250 - 264, May 2004

[8] S. Axelrod, R. Gopinath, and P. Olsen, "Modeling with a subspace constraint on inverse Covariance matrices", in *Proc. ICSLP 2002*.

[9] Olsen, P.A.; Gopinath, R.A., "Modeling inverse covariance matrices by basis expansion", *IEEE Trans. Speech and Audio Processing*, vol. 12, pp.37 - 46, Jan. 2004

[10] Hui Lin, Ye Tian, Jian-Lai Zhou, Hui Jiang, “Hierarchical correlation compensation for hidden markov models,” Submitted to *Proc. ICASSP2005*.