

# STRUCTURED PRECISION MODELLING WITH CHOLESKY BASIS SUPERPOSITION FOR SPEECH RECOGNITION

Lei Jia<sup>†</sup>   Kai Yu<sup>\*</sup>   Bo Xu<sup>†</sup>

<sup>†</sup> Digital Media Content Technology Center  
Institute of Automation, C.A.S., 100190, Beijing, P. R. China

<sup>\*</sup>Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK

## ABSTRACT

Structured precision modelling is an important approach to improve the intra-frame correlation modelling of the standard HMM, where Gaussian mixture model with diagonal covariance are used. Previous work has all been focused on direct structured representation of the precision matrices. In this paper, a new framework is proposed, where the structure of the Cholesky square root of the precision matrix is investigated, referred to as *Cholesky Basis Superposition* (CBS). Each Cholesky matrix associated with a particular Gaussian distribution is represented as a linear combination of a set of Gaussian independent basis upper-triangular matrices. Efficient optimization methods are derived for both combination weights and basis matrices. Experiments on a Chinese dictation task showed that the proposed approach can significantly outperformed the direct structured precision modelling with similar number of parameters as well as full covariance modelling.

**Index Terms**— inverse covariance modeling, Cholesky square root, precision modelling

## 1. INTRODUCTION

The Hidden Markov model (HMM) is the most popular acoustic model for large vocabulary continuous speech recognition (LVCSR) systems. The state output probability distribution is typically represented by a multivariate Gaussian mixture models (GMM). For efficiency reasons, the covariance matrix of each Gaussian is normally assumed to be diagonal. Consequently, the spectral (intra-frame) correlation is poorly modelled. Although explicitly using full covariance can improve the intra-frame correlation modelling, it suffers from high computational cost and unreliable parameter estimation due to a significantly increased number of parameters compared to the use of diagonal covariance matrix. Hence, normally full covariance matrices are not directly used in LVCSR systems. Various approximation techniques have been proposed to get a better trade-off between effective intra-frame correlation modelling and the model complexity. Recently, approximation techniques which model *precision* matrices, i.e. *inverse covariance* matrices, are getting increasingly popular, which is the focus of this paper.

One way to achieve a compact representation of a precision matrix is to use a *precision basis superposition* (PBS) framework [1]. In this framework, a precision matrix is approximated by a set of common basis matrices and Gaussian component specific basis weights. A number of precision modelling techniques fall into this framework. Semi-tied covariance (STC) [2] and extended maximum likelihood linear transform (EMLLT) [3] employ global symmetric rank-1 bases with different basis order. Subspace for precision and mean

(SPAM) uses generic symmetric matrices as bases and have been shown to be the most powerful model [4]. Within the PBS framework, the estimation of the basis matrices and the basis coefficient are not trivial and can be very different for different models. In the case of STC, efficient closed-form solutions exist for both basis vectors (rank-1 matrices) and weights update. In contrast, there is no closed form solution for updating basis vectors of EMLLT. Generic gradient descent method is usually used instead. The basis weights of EMLLT can be easily updated using either multiplicative or additive approach [5]. Different from STC and EMLLT, there is no efficient update method for SPAM basis matrices. Even gradient descent method is not directly applicable as the gradient w.r.t. basis matrices can not be analytically evaluated. Hence, numerical optimization packages are normally used. As SPAM uses globally shared generic symmetric basis matrices, the convergence of the estimation process has been found to be poor. This limits the possible gain which could be achieved by SPAM. As for basis weights update, there is no closed form solution either, generic gradient descent approach or a more efficient conjugate gradient based algorithm [6] can be used. Due to high complexity of the optimization, it is recommended not to update the basis matrices of SPAM in LVCSR tasks [1]. Nevertheless, SPAM is the most powerful model within the PBS framework and has shown significant performance gain over other PBS approaches.

In this paper, a new framework of structured precision modelling is proposed. Here, precision matrices is first decomposed into its Cholesky square root. Then, the corresponding Cholesky matrix is approximated using a basis superposition framework. This is referred to as structured precision modelling with *Cholesky basis superposition* (CBS). An advantage of the CBS framework is that it is more compact than the PBS framework due to the combination effect of Cholesky matrices. Furthermore, as Cholesky matrix is an upper triangular matrix, whose determinant can be easily represented, there exist closed-form basis matrices update solution and efficient basis weights estimation formula. This will allow easy parameter update and consequently improved recognition performance.

The rest of the paper is arranged as follows. Section 2 describes the details of the CBS framework, including the parameter update algorithm and implementation issues. Experiments are shown in section 3, followed by conclusion.

## 2. STRUCTURED PRECISION MODELLING WITH CHOLESKY BASIS SUPERPOSITION

GMM is normally used as state output distribution of HMMs. It can be expressed as

$$p(\mathbf{o}|s) = \sum_{m \in \mathcal{M}(s)} \omega_m \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (1)$$

where  $M(s)$  is the total number of Gaussian components of state  $s$ ,  $\mathcal{N}(\cdot)$  denotes Gaussian distribution and  $\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m$  are the mean and covariance matrix respectively,  $\omega_m$  denotes component prior. As indicated before, it may be more convenient to model the inverse covariance matrix, namely precision matrix. Here,  $\mathbf{P}_m$  is used to denote the precision matrix of Gaussian component  $m$ , i.e.  $\mathbf{P}_m = \boldsymbol{\Sigma}_m^{-1}$ . In the rest of the paper, for a matrix  $\mathbf{A}$ , the  $i^{\text{th}}$  row is denoted as  $\mathbf{A}^T(i)$  while the element of the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column is denoted as  $\mathbf{A}(i, j)$ . Sometimes, matrix  $\mathbf{A}$  is also denoted as  $[\mathbf{A}(i, j)]_{ij}$ .

## 2.1. Cholesky basis superposition (CBS)

As described in section 1, the PBS framework directly approximates precision matrix. The general form of PBS can be written as

$$\mathbf{P}_m = \sum_{i=1}^N \lambda_m^i \mathbf{S}_i \quad (2)$$

where  $N$  is the basis order,  $\lambda_m^i$  is the Gaussian component specific weight associated with the  $i^{\text{th}}$  basis matrix  $\mathbf{S}_i$ .

*Cholesky basis superposition (CBS)*, on the other hand, introduces the basis superposition structure for the Cholesky square root of precision matrix. Hence, the precision matrix is approximated as

$$\mathbf{P}_m = \mathbf{R}_m^T \mathbf{R}_m = \left( \sum_{i=1}^N \lambda_m^i \mathbf{U}_i \right)^T \left( \sum_{i=1}^N \lambda_m^i \mathbf{U}_i \right) \quad (3)$$

where  $\mathbf{R}_m$  is the Cholesky square root matrix of the precision matrix  $\mathbf{P}_m$ ,  $\mathbf{U}_i, i = 1, \dots, N$ , are the set of non-singular *upper triangular* basis matrices, and  $\boldsymbol{\lambda}_m = [\lambda_m^1, \dots, \lambda_m^N]$  is the basis coefficient vector associated with Gaussian  $m$ . The basis matrices  $\mathbf{U}_i$  are shared by all gauss components and basis weights are components specific, which is the similar to the PBS framework.

With basis superposition, the number of parameters of CBS is significantly smaller than that of full covariance modelling. By comparing equation (3) to (2) and considering that  $\mathbf{S}_i$  can be decomposed into Cholesky square root as well, it can be found that, given the same number of basis matrices, CBS actually has more effective basis matrices than PBS due to the combination effect. Hence, CBS is more compact than both full covariance modelling and PBS with similar model complexity. In order to avoid very small determinant of covariance matrix, a constraint is introduced, which has similar function as variance floor

$$0 < \left| \sum_{i=1}^N \lambda_m^i \mathbf{U}_i(k, k) \right| < \rho \quad k = 1, \dots, D \quad (4)$$

where  $D$  is the number of feature dimensions,  $\rho$  is a positive threshold. It is worth noting that, with constraint (4), no matter the weights or basis matrices are positive (definite) or not, the resultant  $\mathbf{P}_m$  is ensured to be positive definite. This naturally avoids the happening of negative covariance matrix which is possible within the PBS framework. Although equation (4) does not have sign constraint, in this paper,  $\sum_{i=1}^N \lambda_m^i \mathbf{U}_i(k, k)$  is constrained to be positive during parameter update as shown below.

## 2.2. Basis weights update

Expectation maximum (EM) algorithm can be used to update basis weights of CBS given fixed basis matrices. The auxiliary function

for basis weights can be expressed as (ignoring irrelevant constant)

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\lambda}) &= \frac{1}{2} \sum_{m,t} \gamma_m(t) \left( \log |\mathbf{P}_m| - (\mathbf{o}_t - \boldsymbol{\mu}_m)^T \mathbf{P}_m (\mathbf{o}_t - \boldsymbol{\mu}_m) \right) \\ &= \frac{1}{2} \sum_m \left( \beta_m \log \left| \sum_{i=1}^N \lambda_m^i \mathbf{U}_i \right|^2 - \text{tr}(\boldsymbol{\lambda}_m \boldsymbol{\lambda}_m^T \mathbf{W}_m) \right) \end{aligned} \quad (5)$$

where  $\gamma_m(t)$  is the component posterior occupancy at time  $t$ ,  $\text{tr}(\cdot)$  denotes trace function,  $\beta_m = \sum_t \gamma_m(t)$ , and

$$\mathbf{W}_m = \left[ \sum_t \gamma_m(t) (\mathbf{o}_t - \boldsymbol{\mu}_m)^T \mathbf{U}_i^T \mathbf{U}_j (\mathbf{o}_t - \boldsymbol{\mu}_m) \right]_{ij} \quad (6)$$

Differentiating equation (5) w.r.t.  $\boldsymbol{\lambda}_m$  yields

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\lambda}_m} = \beta_m \mathbf{b}_m - \mathbf{W}_m \boldsymbol{\lambda}_m \quad (7)$$

where  $\mathbf{b}_m = \left[ \sum_{k=1}^D \frac{\mathbf{U}_1(k, k)}{\mathbf{R}_m(k, k)}, \dots, \sum_{k=1}^D \frac{\mathbf{U}_N(k, k)}{\mathbf{R}_m(k, k)} \right]^T$ ,  $\mathbf{R}_m$  is defined in equation (3). Similar to the basis weights update in the PBS framework, there is no closed-form solution for equation (7). However, it is interesting to further investigate the property of the above auxiliary function. By taking second order derivative of equation (5), the Hessian matrix can be derived as

$$\mathbf{H}(\boldsymbol{\lambda}_m) = -\beta_m \mathbf{B}_m - \mathbf{W}_m \quad (8)$$

where  $\mathbf{B}_m = \left[ \sum_{k=1}^D \frac{\mathbf{U}_i(k, k) \mathbf{U}_j(k, k)}{\mathbf{R}_m^2(k, k)} \right]_{ij}$ . Since  $\mathbf{B}_m$  and  $\mathbf{W}_m$  are both non-singular symmetric matrices, the Hessian matrix is ensured to be positive definite. Therefore, the maximization of auxiliary function (5) can be transformed into a convex optimization problem with constraint (4), which does not have local optimum problem. The guarantee to find global optimum is a nice property, which is not easily achieved by the equivalent PBS SPAM approach. Iterative numerical solution [7] is used in this paper for optimization. Assuming  $v_m$  is the current search direction to find next  $\boldsymbol{\lambda}_m$ , the optimal step size along direction  $v_m$  can be found by performing one-dimensional line search along  $v_m$  within the interval  $(a, b)$  determined by the constraint (4). As mentioned in section 2.1, only positive constraint is considered in this paper. Hence, the upper bound  $b$  and the lower bound  $a$  are defined as

$$\begin{aligned} a &= \max \left( 0, \max_{1 \leq k \leq D} \left\{ - \frac{\sum_{i=1}^N \lambda_m^i \mathbf{U}_i(k, k)}{\sum_{i=1}^N v_m^i \mathbf{U}_i(k, k)} \right\} \right) \\ b &= \min_{1 \leq k \leq D} \left\{ \rho - \frac{\sum_{i=1}^N \lambda_m^i \mathbf{U}_i(k, k)}{\sum_{i=1}^N v_m^i \mathbf{U}_i(k, k)} \right\} \end{aligned} \quad (9)$$

## 2.3. Basis matrices update

As indicated in section 1, basis matrices update is very difficult within the PBS framework if the rank of these basis matrices is greater than 1. This has been a great problem for complex PBS models, such as SPAM. However, by using the CBS framework, this problem can be effectively solved due to the property of upper triangular matrix. This section will detail the derivation of closed form solution for basis matrices update using EM algorithm.

The update of the basis matrices is performed in a row-by-row fashion. Considering that the determinant of an upper triangular matrix is the product of its diagonal elements, the auxiliary function (5)

can be rearranged for updating  $U_i$  as

$$\mathcal{Q}(U_i) = \sum_m \left( \beta_m \sum_{k=1}^D \log(\lambda_m^i U_i(k, k) + C_m^i(k, k)) - \frac{1}{2} \text{tr}((\lambda_m^i U_i + C_m^i) G_m (\lambda_m^i U_i + C_m^i)^T) \right) \quad (10)$$

where  $\log(\cdot)$  is directly due to the positive value assumption as explained near equation (9) and

$$C_m^i = \sum_{j=1, j \neq i}^N \lambda_m^j U_j \quad (11)$$

$$G_m = \sum_t \gamma_m(t) (\mathbf{o}_t - \boldsymbol{\mu}_m)(\mathbf{o}_t - \boldsymbol{\mu}_m)^T \quad (12)$$

Differentiating equation (10) w.r.t.  $U_i^T(k)$ , the  $k^{\text{th}}$  row of  $U_i$ , yields

$$\frac{\partial \mathcal{Q}}{\partial U_i(k)} = \sum_m \left( \beta_m \mathbf{k}_m - (\lambda_m^i)^2 G_m U_i(k) - \lambda_m^i G_m C_m^i(k) \right) \quad (13)$$

where  $\mathbf{k}_m = [0, \dots, \frac{\lambda_m^i}{\lambda_m^i U_i(k, k) + C_m^i(k, k)}, \dots, 0]^T$ , only the  $k^{\text{th}}$  element is not zero. Note that partial derivative (13) is a vector, equating it to zero will lead to  $D$  equations, each one corresponding to one particular feature dimension. This equation group can be effectively solved using the property of upper triangular matrix.

Considering the last row  $U_i(D) = [0, \dots, 0, U_i(D, D)]^T$ , there is only one non-zero equation with one variable from the derivative (13). The estimation of  $U_i(D, D)$  then becomes a one-dimensional quadratic optimization problem with the constraint (9). If  $k$  is less than  $D$ , a forward and backward substitution algorithm can be used to solve the equation group, which has a similar idea as the standard Cholesky solution to solve equation groups. Firstly each of the last non-zero  $D - k$  elements of  $U_i(k)$  is expressed as a linear function of  $U_i(k, k)$  by rearranging the last  $D - k$  linear equations derived from (13). Secondly, these linear functions are backward substituted into the only quadratic equation for  $U_i(k, k)$ . Optimal value of  $U_i(k, k)$  can be obtained by solving the quadratic equation with the constraint (9). Then  $U_i(k, k)$  can be forward substituted into the above linear functions to get the optimal values for the last non-zero  $D - k$  elements of  $U_i(k)$ . The optimization process is iterative, since the solution of  $U_i(k)$  is dependent on the  $k^{\text{th}}$  row of the other upper triangular basis matrices. The optimization needs to iterate over the  $k^{\text{th}}$  row vector of all basis matrices interleavingly until convergence. The optimization of different rows of basis matrices are performed independently. With the above forward backward substitution algorithm, closed form solution for updating Cholesky basis matrices is obtained and convergence is easily obtained. Compared to SPAM where basis matrices update is very difficult, the CBS framework allows more efficient and effective way to improve model accuracy.

#### 2.4. Likelihood calculation and other issues

With the CBS framework, likelihood calculation can be efficiently performed. For a particular Gaussian with the mean  $\boldsymbol{\mu}_m$ , the log likelihood is expressed as

$$\log p(\mathbf{o} | \boldsymbol{\mu}_m, \mathbf{P}_m) = K + \frac{1}{2} \left( \log |\mathbf{P}_m| - \left( \sum_{i=1}^N \lambda_m^i (\tilde{\mathbf{o}}^i - \tilde{\boldsymbol{\mu}}_m^i) \right)^T \left( \sum_{i=1}^N \lambda_m^i (\tilde{\mathbf{o}}^i - \tilde{\boldsymbol{\mu}}_m^i) \right) \right) \quad (14)$$

where  $\tilde{\mathbf{o}}^i = U_i \mathbf{o}$  and  $\tilde{\boldsymbol{\mu}}_m^i = U_i \boldsymbol{\mu}_m$ . By separating the model independent part and caching the transformed observations, efficient likelihood calculation can be achieved.

The Gaussian mean update can be easily performed given the current estimation of  $\mathbf{P}_m$  with standard update formula. Due to the efficient parameter estimation, the update of basis weights, basis matrices and Gaussian means can be embedded into the component splitting procedure of HMM training. During the Gaussian component split process, the perturbation of mean vector can not be done at dimension level as no diagonal covariance matrix is used. Instead, the combined Cholesky matrix is used to transform the mean vector and the result is used as the perturbation direction in this paper. It is possible to first initialize and estimate all parameters for a single component system. Then the number of Gaussian can be increased by only perturbation of the mean vector while keeping the other parameters unchanged. The new components obtained from the splitting will share the same basis weights and basis matrices as before. After that, all parameters can be re-estimated. This process will continue until the number of Gauss components reach the target. This process is expected to lead to better model estimate than directly initializing from a multiple component HMM system.

### 3. EXPERIMENTAL RESULTS

The proposed CBS approach was evaluated on a large vocabulary Mandarin dictation task. A total number of 190K sentences from 403 speakers (about 230h) recorded in a clean environment are used for training. 12-dimensional MFCC cepstral coefficients and log energy together with their first and second-order derivatives was used as the feature (39 dimension in total). Gender-dependent cross-word triphone systems were built. Decision-tree based state tying was applied leading to about 3780 tied-states. On average each state has about 24 Gaussian components. The test set consists of 20 speakers with known gender information, each uttering 60 sentences. A trigram model with 40K word vocabulary is used during decoding.

The first experiment gave the performance of the CBS approach with different basis orders. Here, a diagonal covariance system and a full covariance system were used as the baselines. The decoding speed of the full covariance system was improved by using the method proposed by IBM [8], in which Cholesky decomposition of precision matrix is performed to allow pruning of component calculation first and then hierarchical Gaussian computation was used. In order to have robust parameter estimation for the full covariance system, each state only has 4 Gaussian components and the eigen-values of full covariance matrices were floored by 0.01. The initialization of basis weight and basis matrix of CBS used the same scheme as that mentioned in [1] for SPAM initialization. For the estimation of the basis weights of CBS, numerical solution of BFGS [7] was used. Since the computation load involved in the basis weight estimation is mainly incurred by derivatives calculation, the search of the optimal step size can be carried out easily. The elegant convex property guarantees the stable increase of the objective function. Normally the weights optimization converges after 50 iterations with most of the likelihood gain in the first 20 iterations. The convergence property of basis matrices update is similar to the row-by-row optimization scheme of STC [2]. The likelihood of training data increases monotonically and optimization will be converged in 10 iterations. The basis weights and basis matrices were updated alternatively and 3 iteration loops is enough to get most of the gain.

Figure 1 illustrates the character error rate (CER) of different CBS systems as a function of basis order (the unit is the number of feature dimension). From figure 1, it can be observed that CER

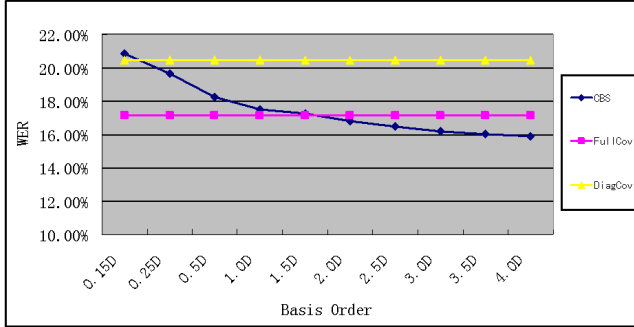


Fig. 1. CER of CBS, diagonal covariance and full covariance with the change of Basis Order(from 0.15D to 4D)

decreases with the increase of Cholesky basis order. When the basis order is above 78 (2D), the CER of CBS model is lower than the full covariance system.

The next experiment further compared the CBS technique to several commonly used covariance modelling techniques, in particular the PBS approaches. Full covariance modelling with diagonal backup[9] was adopted in this experiment, where full covariance were estimated only for the components with sufficient training data and diagonal covariance were used when training data is sparse. As for the PBS framework, state dependent STC system was built, i.e. there are 3780 STC transforms in total. SPAM is the PBS equivalence of the proposed CBS method in this paper. Therefore, the basis order of SPAM was set to be identical to the CBS system, which is 117(3d). The SPAM system only optimized the basis weights with all basis matrices fixed, as suggested in [1]. Several CBS systems were built. The first one initialized the CBS system from the multiple component system and only updated basis weights. The second one used similar initialization but updated both basis weights and basis matrices. The third one initialized the CBS system from a single component system and embedded the update of basis weights, basis matrices and Gaussian means in the whole Gaussian splitting process during training. All systems in this experiment have 24 Gaussian components per state on average.

Covariance Model		Free Param. (K)	CER (%)
Diagonal		3538.1	20.46
Full		70761.6	17.20
PBS	STC	9434.9	17.67
	SPAM	10705.5	17.89
CBS	$\lambda$	10705.5	16.77
	$\lambda + U$	10705.5	16.30
	$\lambda + U + 1$ Comp. Init	10705.5	16.15

Table 1. CER of different covariance modelling techniques

From table 1, it can be observed that the proposed CBS technique can obtain significant improvements over the other covariance modelling methods, especially the PBS techniques. Comparing the CBS systems to SPAM shows that the elegant update scheme of CBS has led to great advantages. About 10% relative improvement was obtained over SPAM with the same model complexity. Also consistent improvements have been obtained over full covariance modelling with diagonal backup, which shows that CBS can effectively address the issue of data sparseness and provide a compact model

representation with high accuracy. It is worth noting that though there exist caching scheme for CBS as indicated before, the computational cost of the CBS approach during decoding is still much heavier than the PBS SPAM system due to the combination effect. Future work will investigate the computation load reduction for the CBS models as well as applying discriminative training to the CBS framework.

#### 4. CONCLUSION

This paper has described a new framework for structured precision modelling. Here Cholesky square root of precision matrix is represented as a linear combination of basis matrices with corresponding basis weights. This representation is more compact than the direct approximation to precision matrix. Efficient update scheme for basis weights and closed form solution for basis matrices can be derived. Experimental results showed that significant improvement can be obtained over traditional PBS SPAM modelling and robust full covariance modelling.

#### 5. ACKNOWLEDGEMENT

This research is supported by Natural Science Foundation of China (Grant No. 90820303).

#### 6. REFERENCES

- [1] K.C.Sim and M.J.F. Gales, "Precision matrix modeling for large vocabulary continuous speech recognition," Tech. Rep. CUED/F-INFENG/TR485, Cambridge University Engineering Department, 2004.
- [2] M.J.F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.
- [3] P.Olsen and R.A.Gopinath, "Extended mllt for gaussian mixture models or modeling inverse covariances by basis expansion," *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 37 – 46, 2004.
- [4] K. Visweswariah, P.Olsen, R.A. Gopinath, and S.Axelrod, "Maximum likelihood training of subspaces for inverse covariance modeling," in *Proc. ICASSP*, 2003.
- [5] P.Olsen and R.A.Gopinath, "Modelling inverse covariance matrices by basis expansion," in *Proc. ICASSP*, 2002.
- [6] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, 1971.
- [7] C.G. Broyden, J.E. Dennis, Jr., and Jorge J. More, "On the local and superlinear convergence of quasi-newton methods," *Journal of the Institute of Mathematics and its Applications*, vol. 12, pp. 223–245, 1973.
- [8] H.Soltau, B.Kingsbury, L.Mangu, D.Povey, G.Saon, and G.Zweig, "The ibm 2004 conversational telephone system for rich transcription," in *Proc. ICASSP*, 2005.
- [9] Ye Tian, Jian-Lai Zhou, Hui Lin, and Hui Jiang, "Tree-based covariance modeling of hidden markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 14, pp. 2134 – 2146, 2006.