ACOUSTIC MODELING USING TRANSFORM-BASED PHONE-CLUSTER ADAPTIVE TRAINING

Vimal Manohar, Bhargav Srinivas Ch., Umesh S

Department of Electrical Engineering,
Indian Institute of Technology Madras,
Chennai, India

ABSTRACT

In this paper, we propose a new acoustic modeling technique called the Phone-Cluster Adaptive Training. In this approach, the parameters of context-dependent states are obtained by the linear interpolation of several monophone cluster models, which are themselves obtained by adaptation using linear transformation of a canonical Gaussian Mixture Model (GMM). This approach is inspired from the Cluster Adaptive Training (CAT) for speaker adaptation and the Subspace Gaussian Mixture Model (SGMM). The parameters of the model are updated in an adaptive training framework. The interpolation vectors implicitly capture the phonetic context information. The proposed approach shows substantial improvement over the Continuous Density Hidden Markov Model (CDHMM) and a similar performance to that of the SGMM, while using significantly fewer parameters than both the CDHMM and the SGMM.

Index Terms— Acoustic Modeling, Subspace Gaussian Mixture Models, Phone-Cluster Adaptive Training

1. INTRODUCTION

Many GMM-based techniques widely used in speaker recognition and adaptation have proved to be successful when adopted for acoustic modeling in speech recognition. For example, [1] adapts a Universal Background Model (UBM) through a maximum a posteriori (MAP) scheme to each context-dependent phone in a way analogous to the MAP adaptation of UBM to each speaker during speaker recognition[2]. The Subspace Gaussian Mixture Model (SGMM) [3] tries to estimate basis vectors for the phonetic and speaker spaces. This approach is similar to the Joint Factor Analysis (JFA) [4], which tries to identify basis vectors for channel and speaker spaces. Recently, eigentriphones [5] was proposed, which develops an eigenbasis over context-dependent phones (triphones) and identifies each triphone as a point in the space spanned by that basis. This idea is adopted from the eigenvoices [6] approach for speaker adaptation.

We propose a new acoustic modeling technique, the transform-based Phone-Cluster Adaptive Training, hereafter referred to as *phone-CAT*. This method is inspired from the Cluster Adaptive Training (CAT), a speaker adaptation technique. In CAT, a speaker adapted model is formed by linear combination of several speaker cluster models that are adapted from a single speaker independent (SI) model. Similarly in phone-CAT, the context-dependent state models are obtained as a linear combination of monophone cluster models that are adapted from a canonical GMM through linear transformations. So, each tied context-dependent state is characterized by a linear interpolation vector, whose elements are weights assigned to the monophone cluster models.

Our technique, like the SGMM, models the HMM state parameters as vectors in a subspace of the total parameter space. But, instead of learning the subspace directly as in the case of SGMM, the structure of the subspace is defined in the form of linear transformations of a canonical model. This greatly reduces the number of free parameters to be estimated, which is advantageous for building robust acoustic model when less training data is available.

Our work also falls under the broad category of Canonical State Models (CSM)[7]. The MLLR-based CSM also adapts a canonical state through a linear transformation to a context-dependent state. But, rather than modeling the context-dependent state **distribution** as a linear combination of the **GMMs** of the transformed canonical states, we model the context-dependent state distribution **parameters** as linear combinations of the **parameters** of the GMMs of the transformed canonical states.

In this paper, we developed an adaptive training framework for the estimation of parameters of the canonical GMM, the linear transforms for monophone clusters and the linear interpolation vectors for context-dependent states. This type of adaptive training seems to preserve the phonetic context information in the interpolation vector, the plots of which are shown in the latter sections. We present the results on Aurora 4 [8] and Resource Management (RM) [9] databases. The model is shown to have a performance superior to that of CDHMM and on par with that of the SGMM.

The rest of the paper is organized as follows. Sections 2 and 3 give model description and training procedure of the phone-CAT model. Section 4 gives details of the experimental setup and results, followed by conclusions and future work in Section 5.

2. TRANSFORM-BASED PHONE-CAT MODEL

The block schematic diagram of phone-CAT model is shown in Fig. 1. The phone-CAT model consists of a set of P clusters corresponding to the P monophone models. Each cluster p has a cluster-specific mean $\boldsymbol{\mu}_i^{(p)}$ for each Gaussian component $1 \leq i \leq I$. The means of each context-dependent HMM state j are expressed as a linear combination of the P cluster means with interpolation weight vector $\mathbf{v}_j = \begin{bmatrix} v_j^{(1)} & v_j^{(2)} & \dots & v_j^{(P)} \end{bmatrix}^T$. Thus the mean of the i^{th} Gaussian of the j^{th} context-dependent state is modeled as follows:

$$\boldsymbol{\mu}_{ji} = \begin{bmatrix} \boldsymbol{\mu}_i^{(1)} & \boldsymbol{\mu}_i^{(2)} & \dots & \boldsymbol{\mu}_i^{(P)} \end{bmatrix} \mathbf{v}_j, \qquad (1)$$

This work was supported under the SERC project funding SR/S3/EECE/050/2013 of the Department of Science and Technology, India.

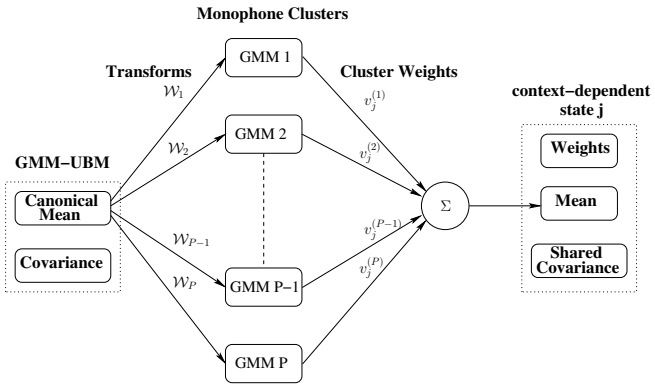


Fig. 1: Block diagram of phone-CAT

The phone cluster means $\boldsymbol{\mu}_i^{(p)}$ are not specified directly, but as linear transformations of the means of a canonical GMM. In the basic model, there is an MLLR transform, $\boldsymbol{\mathcal{W}}_p$, associated with each cluster p. The cluster-specific mean $\boldsymbol{\mu}_i^{(p)}$ for i^{th} Gaussian component is specified as:

$$\boldsymbol{\mu}_i^{(p)} = \boldsymbol{\mathcal{W}}_p \boldsymbol{\xi}_i = \boldsymbol{\mathcal{W}}_p \left[\begin{array}{cc} \boldsymbol{\mu}_i & 1 \end{array} \right]^T, \tag{2}$$

where $\boldsymbol{\xi}_i$ is the extended canonical model mean $\left[\begin{array}{cc} \boldsymbol{\mu}_i & 1 \end{array} \right]^T$ with $\boldsymbol{\mu}_i$ being the canonical mean of the i^{th} Gaussian. Using this, (1) can be rewritten as:

$$\begin{aligned} \boldsymbol{\mu}_{ji} &= \left[\begin{array}{ccc} \boldsymbol{\mu}_i^{(1)} & \ldots & \boldsymbol{\mu}_i^{(P)} \end{array} \right] \left[\begin{array}{c} v_j^{(1)} \\ \vdots \\ v_j^{(P)} \end{array} \right], \qquad (3) \\ &= \sum_{p=1}^{P} \boldsymbol{\mu}_i^{(p)} v_j^{(p)}, \\ &= \left(\sum_{p=1}^{P} v_j^{(p)} \boldsymbol{\mathcal{W}}_p \right) \boldsymbol{\xi}_i, \qquad (4) \end{aligned}$$

where $\mathbf{v}_j = \left[\begin{array}{ccc} v_j^{(1)} & \ldots & v_j^{(P)} \end{array} \right]^T$ is the linear interpolation vector, also called as the state vector.

The phone-CAT model has 3 distinct model sets. At the lowest level, there is a compact canonical model representing the average variability of all the speech data. At the intermediate level, there is a set of P clusters representing the speech subspace. The cluster means $\boldsymbol{\mu}_i^{(1)}$, $\boldsymbol{\mu}_i^{(2)}$, $\ldots$, $\boldsymbol{\mu}_i^{(P)}$ form the basis vectors of this subspace. These P models are linear transformations, represented by (2), of the canonical model. At the highest level, there is a set of J tied context-dependent states, whose models are obtained as linear interpolation of the P cluster models.

2.1. Model Description

The model can be expressed with the following equations:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} w_{ji} \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i\right), \tag{5}$$

$$\boldsymbol{\mu}_{ji} = \left(\sum_{p=1}^{P} v_j^{(p)} \boldsymbol{\mathcal{W}}_p \right) \boldsymbol{\xi}_i, \tag{6}$$

$$w_{ji} = \frac{\exp\left(\mathbf{w}_i^T \mathbf{v}_j\right)}{\sum_{i'=1}^{I} \exp\left(\mathbf{w}_{i'}^T \mathbf{v}_j\right)}, \tag{7}$$

where $\mathbf{x} \in \mathbb{R}^D$ is the feature vector of dimension D, $1 \leq j \leq J$ is the state index of the context-dependent state, $\mathbf{v}_j \in \mathbb{R}^P$ is the linear interpolation vector with P being the number of clusters and $v_j^{(p)}$ being the p^{th} element of it, $\boldsymbol{\mathcal{W}}_p$ is an MLLR Transform matrix corresponding to the p^{th} cluster, $\mathbf{w}_i \in \mathbb{R}^P$ is the weight projection vector, $\boldsymbol{\xi}_i = \left[\begin{array}{cc} \boldsymbol{\mu}_i & 1 \end{array} \right]^T$ is the extended mean of the i^{th} Gaussian component of the canonical model.

As given by (5), each context-dependent state is a GMM with I Gaussians with means $\boldsymbol{\mu}_{ji}$, covariances $\boldsymbol{\Sigma}_i$ and weights w_{ji}. As seen in (7), the idea of modeling Gaussian priors using a softmax function borrowed from SGMM [3] works well in phone-CAT model also. The model is similar to the SGMM, with covariances $\boldsymbol{\Sigma}_i$ being shared across all context-dependent states, and also the means $\boldsymbol{\mu}_{ji}$ and the weights w_{ji} being derived from the linear interpolation vector $\mathbf{v}_j$ and hence spanning a smaller P dimensional subspace of the total parameter space. But in this model, the subspace spanned by the means is not specified directly, but as linear transformations of a canonical model. Thus, a canonical model, which consists of I Gaussians with means $\boldsymbol{\mu}_i$ and covariances $\boldsymbol{\Sigma}_i$, is adapted through linear transformations to clusters of monophones. Hence, the model is termed as "Transform-based Phone-Cluster Adaptive Training" model.

A simple extension to the model can be to use piece-wise linear transformation. The Gaussian components of the canonical model are partitioned into Q disjoint transform classes, $M_t^{(1)}$ to $M_t^{(Q)}$. The mean $\boldsymbol{\mu}_{ji}$ is now calculated as:

$$\boldsymbol{\mu}_{ji} = \left(\sum_{p=1}^{P} v_j^{(p)} \boldsymbol{\mathcal{W}}_{pq} \right) \boldsymbol{\xi}_i, \tag{8}$$

where $\boldsymbol{\mathcal{W}}_{pq}$ is the MLLR matrix for transform class q of cluster p, and q is the transform class of the i^{th} Gaussian.

2.2. Parameter count: Phone-CAT Model vs SGMM

Table 1 compares the total number of parameters of a typical phone-CAT model against a typical SGMM for Aurora 4. Both the models have 3957 tied states and a full covariance UBM-GMM of 400 Gaussians. The phone-CAT model considered here has 4 MLLR transforms per cluster. When the subspace dimension (40) in SGMM is close to the number of clusters (42) in the phone-CAT model, the only significant difference in the parameter count is due to the defining of the subspace using 4 MLLR transforms for each of the 42 phone-clusters instead of the 39×40 dimensional matrices $\mathbf{M}_i$ for each of the 400 Gaussians in SGMM. From Table 1, it can be observed that the phone-CAT model consumes fewer parameters than the SGMM. Hence, the phone-CAT model can be used effectively even when the amount of training data available is less.

Table 1: Comparison of parameter count
Phone-CAT: 42 - Number of clusters, 39×40 - Size of MLLR matrix $\boldsymbol{\mathcal{W}}_{pq}$ (with bias), 4 - Number of transform classes
SGMM: 40 - Subspace dimension, 39×40 - Size of $\mathbf{M}_i$, 39 - Dimension of feature vector

Parameters	Phone-CAT			SGMM		
			Count			Count
State-specific	$\mathbf{v}_j$	3957×42	166,194	$\mathbf{v}_j$	3957×40	158,280
Global	μ_i	400×39	15,600	$\mathbf{M}_i$	$400 \times 39 \times 40$	624,000
	$\boldsymbol{\mathcal{W}}_{pq}$	$42 \times 4 \times 39 \times 40$	262,080			
	Σ_i	$400 \times 39 \times 40/2$	312,000	Σ_i	$400 \times 39 \times 40/2$	312,000
	$\mathbf{w}_i$	400×42	16,800	$\mathbf{w}_i$	400×40	16,000
Total			772,674			1,110,280

3. TRAINING OF THE MODEL

3.1. Training procedure

The model training starts with a conventional HMM-GMM system (CDHMM system), which provides the phonetic context information for parameter tying of context-dependent states, a set of Gaussian components to build a UBM as the canonical model and the Viterbi state alignments for the initial training iterations. The model is initialized and trained for a few iterations using the alignments obtained from the HMM-GMM system. In the subsequent iterations, the alignments are obtained from the phone-CAT system itself. There are three distinct parameter sets as in the case of the transform-based CAT. The linear interpolation vector parameters $\Lambda = \{\mathbf{v}_j\}, 1 \leq j \leq J$, canonical model parameters $\mathcal{M} = \{\{\boldsymbol{\mu}_1 \ \ldots \ \boldsymbol{\mu}_I\}, \{\boldsymbol{\Sigma}_1 \ \ldots \ \boldsymbol{\Sigma}_I\}\}$ and the subspace parameters $\mathcal{S} = \{\{\mathbf{w}_1 \ \ldots \ \mathbf{w}_I\}, \{\boldsymbol{\mathcal{W}}_{11} \ \ldots \ \boldsymbol{\mathcal{W}}_{PQ}\}\}$. The training scheme followed is analogous to the case of the transform-based CAT

1. Re-estimate the linear interpolation vector parameters Λ using $\{\mathcal{M}, \mathcal{S}\}$ and the pre-update value of Λ.
2. Re-estimate the subspace parameters $\mathcal{S}$ given $\{\Lambda, \mathcal{M}\}$ and the pre-update value of $\mathcal{S}$.
3. Re-estimate the canonical model parameters $\mathcal{M}$ given $\{\mathcal{S}, \Lambda\}$ and the pre-update value of $\mathcal{M}$ by first updating the canonical means and then the covariances.

Practically, the different sets of parameters can be updated simultaneously to get a good estimate of the model in few iterations. The pre-update values of Λ, $\mathcal{S}$ and $\mathcal{M}$ are used to calculate the Gaussian posteriors. These values are used to accumulate the statistics required for updating parameters.

3.2. Model Initialization

The initialization of the phone-CAT model begins with the building of a Universal Background Model (UBM). The UBM is initialized using a bottom-up-clustering algorithm as in the case of SGMM ([10]) by repeatedly merging the Gaussians in all the states of the HMM-GMM system to get a diagonal GMM and then training a Full covariance GMM using all the training data. This UBM serves as the initial canonical model.

The phone-CAT model "is initialized such that the Gaussians in each state is identical to the Gaussians in the UBM." The MLLR transforms are all set to identity matrices with 0 bias so that all the cluster-specific means are initially identical to the UBM means. When using multiple transform classes, the Gaussians in the canonical model are clustered using the same bottom-up clustering algorithm into the required number of classes. The linear interpolation vectors $\mathbf{v}_j$ is assigned a vector giving a weight 1 to only one cluster depending on a mapping function $\mathcal{C}$ and 0 to every other cluster. In the simplest case, the mapping function can be defined such that $\mathcal{C}(j) = p$, where p is the index of the central phone of the context-dependent state j. Therefore the initialization is:

$$\boldsymbol{\mathcal{W}}_{pq} = \left[\begin{array}{cc} \mathbf{I}_{D\times D} & \mathbf{0}_{D\times 1} \end{array}\right], 1 \leq p \leq P, \ 1 \leq q \leq Q \quad (9)$$

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_i^{(UBM)} \quad , 1 \leq i \leq I \quad (10)$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i^{(UBM)} \quad , 1 \leq i \leq I \quad (11)$$

$$\mathbf{v}_j = \mathbf{e}_k \in \mathbb{R}^P \quad , 1 \leq j \leq J, \ k = \mathcal{C}(j) \quad (12)$$

$$\mathbf{w}_i = \mathbf{0} \in \mathbb{R}^P \quad , 1 \leq i \leq I \quad (13)$$

where $\mathbf{I}_{D\times D}$ is a $D \times D$ identity matrix with D being the dimension of the feature vector, $\mathbf{0}_{D\times 1}$ is a vector of D zeros, $\boldsymbol{\mu}_i^{(UBM)}$, $\boldsymbol{\Sigma}_i^{(UBM)}$ are the mean and the covariance matrix of the i^{th} Gaussian component of the UBM, $\mathbf{e}_k$ is a P dimensional unit vector with the k^{th} dimension as 1 and every other dimension 0, $\mathcal{C} : \{1, \ldots, J\} \rightarrow \{1, \ldots, P\}$ is a mapping from the state j to cluster p and all the other parameters are as defined in Section 2.1.

The model allows more complex mappings for $\mathcal{C}$. For example, when we have position and stress dependent phones, all those phones that are position and stress variants of the same 'real' phone can be assigned to one particular cluster.

3.3. Re-estimation of model parameters

The model parameters are re-estimated using the Expectation-Maximization algorithm, by maximizing the auxiliary function:

$$\mathcal{Q} = \sum_{j,i,t} \gamma_{ji}(t) \left[\log(w_{ji}) - \frac{1}{2}|\boldsymbol{\Sigma}_i| - \frac{1}{2}\left(\mathbf{x}(t) - \boldsymbol{\mu}_{ji}\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}(t) - \boldsymbol{\mu}_{ji}\right)\right], \quad (14)$$

where $\gamma_{ji}(t) = p(j, i|t)$ is the posterior probability of the j^{th} state, i^{th} Gaussian component at time t, $\mathbf{x}(t)$ is the feature vector at time t and w_{ji} and $\boldsymbol{\mu}_{ji}$ are expressed according to (4) and (7). The rest of the symbols are as defined in Section 2.1. The update equations for each of the parameters $\mathbf{v}_j, \boldsymbol{\mathcal{W}}_p, \mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ are obtained by optimizing $\mathcal{Q}$ with respect to that particular parameter keeping the other parameters fixed.

When the parameters $\boldsymbol{\mathcal{W}}_p$ and $\boldsymbol{\mu}_i$ are fixed, the means $\boldsymbol{\mu}_{ji}$ as represented in (3) can be expressed as:

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j, \quad (15)$$

where $\mathbf{v}_j = \begin{bmatrix} v_j^{(1)} & v_j^{(2)} & \dots & v_j^{(P)} \end{bmatrix}^T$ is the linear interpolation vector, and $\mathbf{M}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{(1)} & \boldsymbol{\mu}_i^{(2)} & \dots & \boldsymbol{\mu}_i^{(P)} \end{bmatrix}$ is the matrix obtained by stacking the i^{th} mean of all the P phone clusters, where $\boldsymbol{\mu}_i^{(p)}$ is given as (2). Thus $\boldsymbol{\mu}_{ji}$ takes the same form as in SGMM. Hence, we re-estimate the parameters $\mathbf{v}_j$ and $\mathbf{w}_i$ using the same update equations as in SGMM [10].

3.3.1. Re-estimation of canonical model parameters

The re-estimation of the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ follows a procedure analogous to the re-estimation of the canonical model parameters in transform-based CAT[11]. The update equations are:

$$\boldsymbol{\mu}_i = \left[\sum_{p=1}^{P}\sum_{q=1}^{P} g_{pq}^{(i)} \mathbf{A}_p^T \boldsymbol{\Sigma}_i \mathbf{A}_q\right]^{-1} \left[\sum_{p=1}^{P} \mathbf{A}_p^T \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{k}_p^{(i)T} - \sum_{q=1}^{P} g_{pq}^{(i)} \mathbf{b}_q\right)\right], \quad (16)$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\sum_j \gamma_{ji}} \left[\mathbf{L}^{(i)} - \sum_{p=1}^{P} \mathbf{k}_p^{(i)} \mathbf{M}_i^{(p)T} - \sum_{p=1}^{P} \mathbf{M}_i^{(p)} \mathbf{k}_p^{(i)T} + \sum_{p=1}^{P}\sum_{q=1}^{P} g_{pq}^{(i)} \mathbf{M}_i^{(p)} \mathbf{M}_i^{(q)T}\right], \quad (17)$$

where $\mathbf{A}_p$ is the matrix consisting of the first D columns of $\boldsymbol{\mathcal{W}}_p$, $\mathbf{b}_p$ is the $(D+1)^{th}$ column of $\boldsymbol{\mathcal{W}}_p$, $\mathbf{M}_i^{(p)} = \boldsymbol{\mathcal{W}}_p \boldsymbol{\xi}_i$, $\mathbf{k}_p^{(i)}$ is the p^{th} row of the statistics $\mathbf{K}^{(i)}$, $g_{pq}^{(i)}$ is the $(p,q)^{th}$ element of statistics $\mathbf{G}^{(i)}$, and $\mathbf{G}^{(i)}$, $\mathbf{K}^{(i)}$ and $\mathbf{L}^{(i)}$ are statistics defined by

$$\mathbf{G}^{(i)} = \left[g_{pq}^{(i)}\right]_{1\le p,q\le P} = \sum_{j,t} \gamma_{ji}(t)\, \mathbf{v}_j \mathbf{v}_j^T, \quad (18)$$

$$\mathbf{K}^{(i)} = \left[k_{pk}^{(i)}\right]_{1\le p\le P,\, 1\le k\le D} = \sum_{j,t} \gamma_{ji}(t)\, \mathbf{v}_j \mathbf{x}(t)^T, \quad (19)$$

$$\mathbf{L}^{(i)} = \sum_{j,t} \gamma_{ji}(t)\, \mathbf{x}(t)\, \mathbf{x}(t)^T. \quad (20)$$

3.3.2. Re-estimation of Cluster Transforms $\boldsymbol{\mathcal{W}}_{pq}$ with Full Covariance model

If the covariance used in the model is diagonal, we can use an update procedure analogous to that in the transform-based CAT. But the performance of the model with diagonal covariance is not same as that of a full covariance SGMM. If full covariance model is used in phone-CAT, the standard update procedure becomes complex and computationally very expensive. So we have implemented a second-order gradient descent approach by extending the technique introduced in [12] to adaptive training. This technique is an iterative approach.

In each iteration, the gradient of the auxiliary function (14) w.r.t. $\boldsymbol{\mathcal{W}}_{pq}$ is computed:

$$\begin{aligned} \boldsymbol{\mathcal{L}}_{pq} &= \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\mathcal{W}}_{pq}} \\ &= \sum_{j,i\in M_t^{(q)},t} \gamma_{ji}(t)\, \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x}(t) - \left(\sum_p \boldsymbol{\mathcal{W}}_{pq} v_j^{(p)}\right) \boldsymbol{\xi}_i\right) \boldsymbol{\xi}_i^T v_j^{(p)} \end{aligned} \quad (21)$$

Algorithm 1 Estimation of cluster transform parameters

1. For each cluster $1 \le p \le P$
   (a) For each transform class $1 \le q \le Q$
      i. Initialize learning rate $\alpha = 1$.
      ii. For n iterations, where $n < D$ is typically around 5-10
         A. Compute the first-order gradient $\boldsymbol{\mathcal{L}}_{pq}$ and the second-order gradients $\boldsymbol{\mathcal{G}}_{pq}^{(k)}$ for all dimensions $1 \le k \le D$.
         B. Re-estimate the k^{th} row of $\boldsymbol{\mathcal{W}}_{pq}$ using (23) for all dimensions $1 \le k \le D$.
         C. Compute the change in auxiliary function using (24) before and after the step 1(a)iiB.
         D. If the auxiliary function has increased, commit the updated $\boldsymbol{\mathcal{W}}_{pq}$ and use it for the subsequent iterations. Go to step 1(a)ii.
         E. If the auxiliary function has decreased, choose $\alpha = \alpha/2$. If $\alpha > \alpha_{min}$, go to step 1(a)iiB.

The second-order gradient $\boldsymbol{\mathcal{G}}_{pq}^{(k)}$ is also computed for the all dimensions $1 \le k \le D$. Here, we assume that the second-order gradient remains the same as that for diagonal covariance. It is obtained as:

$$\begin{aligned} \boldsymbol{\mathcal{G}}_{pq}^{(k)} &= \frac{\partial^2 \mathcal{Q}}{\partial \boldsymbol{\mathcal{W}}_{pq}^{(k)2}} \\ &= \sum_{j,i\in M_t^{(q)},t} \gamma_{ji}(t)\, \frac{v_j^{(p)2}}{\sigma_{kk}^{(i)2}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T, \\ &= \sum_{i\in M_t^{(q)}} \frac{g_{pp}^{(i)}}{\sigma_{kk}^{(i)2}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \end{aligned} \quad (22)$$

where $g_{pp}^{(i)}$ is the p^{th} diagonal element of (18) and $\sigma_{kk}^{(i)2}$ is the variance of the i^{th} Gaussian. Using (22), the entire k^{th} row of $\boldsymbol{\mathcal{W}}_{pq}$ can be estimated:

$$\begin{aligned} \hat{\boldsymbol{\mathcal{W}}}_{pq}^{(k)} &= \boldsymbol{\mathcal{W}}_{pq}^{(k)} + \alpha \left[\frac{\partial^2 \mathcal{Q}}{\partial \boldsymbol{\mathcal{W}}_{pq}^{(k)2}}\right]^{-1} \left[\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\mathcal{W}}_{pq}}\right]^{(k)^T}, \\ &= \boldsymbol{\mathcal{W}}_{pq}^{(k)} + \alpha\, \boldsymbol{\mathcal{G}}_{pq}^{(k)-1} \boldsymbol{\mathcal{L}}_{pq}^{(k)^T}, \end{aligned} \quad (23)$$

where $\boldsymbol{\mathcal{W}}_{pq}^{(k)}$ is the k^{th} row of $\boldsymbol{\mathcal{W}}_{pq}$, $\boldsymbol{\mathcal{L}}_{pq}^{(k)}$ is the k^{th} row of $\boldsymbol{\mathcal{L}}_{pq}$ and α is some learning rate. The re-estimation of all the rows of $\boldsymbol{\mathcal{W}}_{pq}$ can be done in parallel using (23).

The change in the auxiliary function (14) after the update of all rows is computed using:

$$\Delta\mathcal{Q} = \Delta \sum_{j,i\in M_t^{(q)},t} \left(\gamma_{ji}(t)\, \mathbf{x}(t)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ji} - 0.5\gamma_{ji}(t)\, \boldsymbol{\mu}_{ji}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ji}\right). \quad (24)$$

The procedure for sequential update of cluster transforms $\boldsymbol{\mathcal{W}}_{pq}$ is explained in Algorithm 1.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Experimental Setup

Word recognition accuracy obtained using the phone-CAT model is compared against that of the CDHMM and the SGMM for the

Aurora 4 and the RM continuous speech databases. The details of the experimental setup used for building all the three models for the aforementioned databases is below. All these models are trained and tested on clean data, and are built using Kaldi speech recognition toolkit [13]. For building the CDHMM and the SGMM for the RM database, the standard recipe from the Kaldi toolkit was used.

- *Feature extraction*: Mel frequency cepstral coefficients (MFCC) are used for parametrizing the speech data. 13 dimensional MFCCs are extracted using the standard signal processing steps. Delta and acceleration coefficients are appended to make composite 39 dimensional MFCC vector. Cepstral mean normalized MFCCs are used as feature vectors for acoustic modeling.
- *Details of the databases:* Aurora 4, which is sampled at 8kHz, has 7138 train utterances and 330 test utterances in the clean set. A 5000 vocabulary bigram language model is used for testing. RM database, sampled at 16kHz, has a train set of 3990 utterances and 1460 test utterances split into six different test sets – Feb'89, Feb'91, Oct'89, Mar'87, Sept'92 and Oct'87.
- *Acoustic Modeling*: A total of 3957 and 1560 tied states are used to model the entire data in Aurora 4 and RM respectively. The following are the other specifications of the acoustic models.
  - *CDHMM:* A three state HMM is used to model crossword triphones and five state HMM for silence. A total of 24,000 Gaussians are used to model the entire data of Aurora 4 as compared to 9,000 Gaussians for the RM task.
  - *SGMM:* A full covariance UBM consisting of 400 Gaussians is built (from CDHMM) for each database. A subspace dimension of 40 is chosen. Although the usage of substates and speaker space gives better performance than the basic SGMM, the same extensions can also be applied to the phone-CAT and hence we compare the results of the proposed method (which is devoid of all such extensions) with the basic version of SGMM.
  - *Phone-CAT:* The number of full covariance UBM mixtures is kept same as in SGMM for comparison purposes. As Aurora 4 uses dictionary comprising of 42 monophones, the number of phone-clusters is taken as 42. Similarly, the number of phone-clusters for RM database is taken as 48.

4.2. Results

Tables 2 and 3 compare the performances of the phone-CAT model, the SGMM and the CDHMM for Aurora 4 and RM databases respectively in terms of word recognition accuracy. It can be seen that the phone-CAT consistently performs better than the CDHMM and is on par with the SGMM. An absolute improvement of 1.94% and 2.2% over CDHMM is attained by the SGMM and the phone-CAT model (with 4 transforms per cluster) respectively in the case of Aurora 4. Similar results can be seen for the case of RM, where an absolute improvement of 0.69% and 0.66% over CDHMM is obtained for the SGMM and the phone-CAT model (with 4 transforms per cluster) respectively. We notice that the phone-CAT model and the SGMM have comparable performances, although the former has $\approx 300k$ less parameters compared to the latter.

4.3. Discussion

4.3.1. Analysis of interpolation vector $\mathbf{v}_j$

Figures 2a and 2c depict the interpolation vectors $\mathbf{v}_j$ for the first state of triphone /s/-/p/+/iy/ and the second state of the triphone /ax/-/m/+/iy/ in phone-CAT model. Both these triphones are however tied with several other similar triphones, which have the same center phone. In Figure 2a, the monophone cluster /p/, which is the center phone of the corresponding triphone /s/-/p/+/iy/, receives the highest weight in the corresponding interpolation vector. Similarly, in Figure 2c, the monophone cluster /m/, which is the center phone of the corresponding triphone /ax/-/m/+/iy/, receives the highest weight in the corresponding interpolation vector. This follows the intuition that a triphone state is a linear combination of monophone clusters with a large contribution from its center phone. This characteristic is observed across all the triphone state interpolation vectors.

For the case of SGMM, plots of the state vectors, $\mathbf{v}_j$, for same context-dependent states – the first state of triphone /s/-/p/+/iy/ and the second state of triphone /ax/-/m/+/iy/ – are shown in Figures 2b and 2d. Unlike the previous case, the SGMM does not give a similar intuition for the distribution of elements of vector $\mathbf{v}_j$.

Table 2: Aurora 4 (clean test case) results (in % Word Recognition Accuracy)

Model	No. of Transform Classes	% Acc	Parameters	
			State	Global
CDHMM	-	87.60	1800k	0
SGMM	-	**89.54**	**158k**	**952k**
Phone-CAT	1	89.24	166k	410k
	2	**89.73**	**166k**	**475k**
	3	89.5	166k	541k
	4	**89.80**	**166k**	**606k**
	5	89.67	166k	672k

5. CONCLUSIONS AND FUTURE WORK

A new acoustic model named phone-CAT is proposed, in which each context-dependent state model is formed by a linear combination of monophone models. The proposed model is shown to have substantial improvement over the CDHMM. It also gave a performance similar to that of the SGMM. It has an added advantage of having fewer parameters to be estimated as compared to the CDHMM and the SGMM.

In future, we intend to extend this work to multiple interpolation vectors for single triphone state. Extensions like substates and speaker space, as in the case of SGMM, can be attempted in phone-CAT also. We would like to apply this technique for building acoustic models for low resource languages.

6. REFERENCES

[1] D. Povey, S. M. Chu, and B. Varadarajan, "Universal background model based speech recognition," in *Proc. ICASSP*, 2008, pp. 4561–4564.

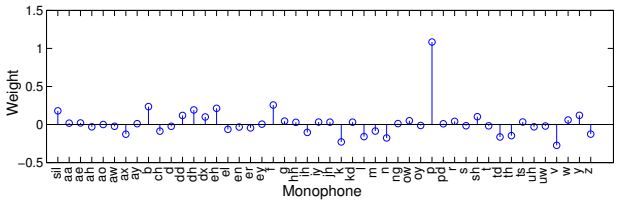
[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[3] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiat, A. Rastrow *et al.*, "Subspace Gaussian mixture models for speech recognition," in *Proc. ICASSP*, 2010, pp. 4330–4333.

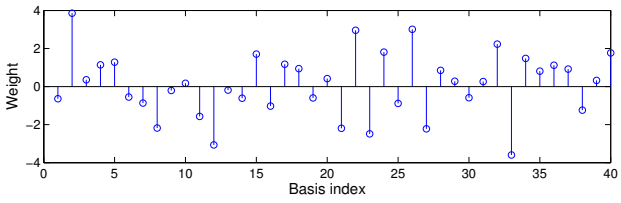
[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

Table 3: RM results (in % Word Accuracy)

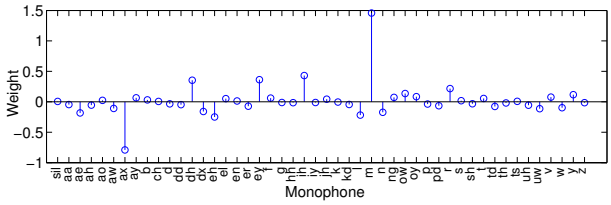
Model	No. of Transform Classes	All	Feb'89	Feb'91	Oct'89	Mar'87	Sep'92	Oct'87	Parameters	
									State	Global
CDHMM	-	96.83							711k	0
SGMM	-	**97.52**	**97.77**	**98.35**	**97.17**	**99.76**	**95.9**	**98.37**	**64k**	**952k**
Phone-CAT	1	97.29	97.70	97.46	97.24	99.16	95.74	98.3	76k	422k
	2	**97.42**	**97.81**	**97.50**	**97.28**	**99.52**	**95.90**	**98.51**	**76k**	**497k**
	3	97.33	97.77	97.42	97.21	99.64	95.90	98.01	76k	571k
	4	**97.49**	**97.97**	**97.67**	**97.62**	**99.64**	**96.01**	**98.44**	**76k**	**646k**
	5	97.18	97.81	97.38	97.06	99.52	95.66	98.37	76k	721k



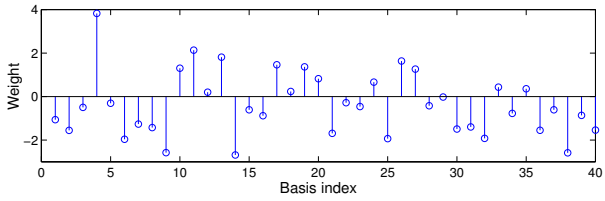
(a) phone-CAT: /s/-/p/+/iy/



(b) SGMM: /s/-/p/+/iy/



(c) phone-CAT: /ax/-/m/+/iy/



(d) SGMM: /ax/-/m/+/iy/

Fig. 2: Comparison of $\mathbf{v}_j$ for transform based phone-CAT and SGMM: Sub-Figures 2a, 2b for the first state of the triphone /s/-/p/+/iy/ and Sub-Figures 2c, 2d for the second state of triphone /ax/-/m/+/iy/

[5] T. Ko and B. Mak, "Eigentriphones: A basis for context-dependent acoustic modeling," in *Proc. ICASSP*, 2011, pp. 4892–4895.

[6] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. ICSLP*, vol. 98, 1998, pp. 1771–1774.

[7] M. J. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Proc. Interspeech*, 2010, pp. 58–61.

[8] D. Pearce, H.-G. Hirsch *et al.*, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Interspeech*, 2000, pp. 29–32.

[9] P. Price, W. M. Fischer, J. Bernsteina, and D. S. Pallett, "Resource management complete set 2.0," from Linguistic Data Consortium, Philadelphia, USA.

[10] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model - a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011.

[11] M. J. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[12] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians," in *Proc. Interspeech*, 2006, pp. 1145–1148.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.