AN INVESTIGATION OF SUBSPACE MODELING FOR PHONETIC AND SPEAKER VARIABILITY IN AUTOMATIC SPEECH RECOGNITION

Richard Rose¹, Shou-Chun Yin¹, Yun Tang²

¹Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
²Nuance Communications Inc., Montreal, Canada

rose@ece.mcgill.ca, sss123ca@yahoo.com, hottang@gmail.com

ABSTRACT

This paper investigates the impact of subspace based techniques for acoustic modeling in automatic speech recognition (ASR). There are many well known approaches to subspace based speaker adaptation which represent sources of variability as a projection within a low dimensional subspace. A new approach to acoustic modeling in ASR, referred to as the subspace based Gaussian mixture model (SGMM), represents phonetic variability as a set of projections applied at the state level in a hidden Markov model (HMM) based acoustic model. The impact of the SGMM in modeling these intrinsic sources of variability is evaluated for a continuous speech recognition (CSR) task. The SGMM is shown to provide an 18% reduction in word error rate (WER) for speaker independent (SI) ASR relative to the continuous density HMM (CDHMM) in the resource management CSR domain. The SI performance obtained from SGMM also represents a 5% reduction in WER relative to subspace based speaker adaption in an unsupervised speaker adaptation scenario.

Index Terms— automated speech recognition

1. INTRODUCTION

Subspace based methods in ASR and speaker verification have been successfully applied to modeling sources of signal variability. These methods attempt to identify low dimensional subspaces from training data which characterize variation among speakers or among other extrinsic factors that may affect the acoustic speech utterance. In subspace-based speaker adaptation, basis vectors of a linear subspace are trained using data from multiple speakers [3, 1, 2]. In text-independent speaker verification, two linear subspace matrices representing speech and session variability are trained using data from multiple speakers over multiple sessions [8]. More recently, the subspace based Gaussian mixture model (SGMM) has been proposed as a more general technique that was originally motivated by subspace-based speaker adaptation and speaker verification approaches [6, 4]. It is based on hidden Markov model state-level subspace projections and represents a significant departure from the parameterization of the CDHMM ASR [5].

This paper investigates the impact of the SGMM approach both in terms of its potential advantages with respect to continuous density HMM acoustic modeling and its relationship to existing subspace based speaker adaptation approaches. In the SGMM, multiple subspace matrices are trained to describe the allowable variation associated with individual distributions in the ASR acoustic model. Each state associated with a phonetic context in a hidden Markov model (HMM) is represented as one or more projections within these subspaces. So the SGMM can be loosely interpreted as a subspace representation of phonetic level variation in speech recognition [5]. The implementation of SGMM that is used in this paper is described in Section 2 and is very similar to that presented in [6]. To motivate the interest in the SGMM for this work a brief description of subspace based speaker adaptation approaches is provided in Section 2.1. In one class of implementations, both speaker subspace matrices and the speaker dependent projections applied within this subspace are estimated according to a maximum likelihood (ML) criterion using the Expectation-Maximization (EM) algorithm [2, 1]. The performance of the clustered maximum likelihood linear basis (CMLLB) approach to speaker space adaptation will be compared with SGMM in Section 4 The SGMM implementation used here can be viewed as an extension of these procedures and an attempt is made to compare the performance of this SGMM implementation with respect to an implementation of these ML subspace adaptation procedures.

The SGMM based acoustic model for ASR will be described in Section 2.2. ML estimates of subspace matrices and HMM state-based projection vectors are obtained using the EM algorithm [6]. It has been shown in [4, 5] and will also be shown here in Section 4 that lower ASR word error rates (WERs) can be obtained with the SGMM than the CDHMM with a smaller number of model parameters. The results presented in Section 4 demonstrate that it is a more efficient method for describing acoustic variability than the CDHMM or the speaker space adaptation approaches.

2. SUBSPACE-BASED MODELS

This section provides a brief introduction to both subspace based speaker adaptation and subspace based model adaptation. Section 2.1 describes subspace speaker adaptation as the estimation of a speaker dependent projection in a single global speaker space. Section 2.2 describes the SGMM as state specific projections within multiple linear subspaces.

2.1. Speaker Space Models for Speaker Adaptation

Speaker space based adaptation is performed on the supervector \( \mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K) \) formed by concatenating the \( D \) dimensional mean vectors associated with \( M \) diagonal covariance Gaussian densities in a continuous Gaussian mixture density HMM. In the training phase for a \( K \) dimensional subspace, a \( MD \times K \) subspace matrix, \( \mathbf{E} \), is estimated from multi-speaker training data where \( K < < MD \) and typically lies in the range \( 10 < K < 100 \). Methods based on both principal components analysis (PCA) [3] and maximum likelihood estimation using the EM algorithm [2, 1]
have been used for estimating $E$. During the adaptation stage, a $K$ dimensional weight vector, $u^s$, is estimated from the adaptation data for a given speaker $s$ and the adapted supervector, $\mathbf{m}$, is computed as

$$\mathbf{m} = \mathbf{m} + \mathbf{E} \mathbf{u}^s. \quad (1)$$

By effectively constraining the variation of HMM model parameters to lie in a very low dimensional space, Equation 1 facilitates efficient adaptation. Good unsupervised speaker adaptation performance has been obtained with under ten seconds of adaptation speech. However, with only a single projection vector, $u^s$, the asymptotic behavior of this class of approaches is such that there is little or no performance improvement observed when additional adaptation speech is available. The performance of the clustered maximum likelihood linear basis (CMLLB) approach will be summarized in the experimental study described in Section 4 [2]. CMLLB achieves a more efficient and robust representation for $E$ by forming the columns of $E$ as a concatenation of a small number of a clustered subvectors.

### 2.2. Subspace Gaussian Mixtures for ASR

The observation densities in the CDHMM are formed from a mixture of state dependent diagonal covariance Gaussians. In the SGMM, the distribution of the $D$ dimensional features, $x$, for HMM state, $j = 1, \ldots, J$ are formed from a set of $I$ shared full covariance Gaussians $N(x; \mu_i, \Sigma_i)$. In the simplest case these state densities are given by

$$p(x|j) = \sum_{i=1}^{I} w_{ji} N(x; \mu_{ji}, \Sigma_i) \quad (2)$$

where the state dependent mean vector, $\mu_{ji}$, for state $j$ is a projection into the $i$th subspace defined by linear subspace projection matrix $M_i$,

$$\mu_{ji} = \bar{\mathbf{m}} + M_i \mathbf{v}_i. \quad (3)$$

In Equation 3, $\mathbf{v}_i$ is the projection vector for state $j$. Typically, $I$ in Equation 2 may be on the order of 100 to 1000. The matrix $M_i$ in Equation 3 is of dimension $D \times S$ where $S$ is the dimension of the subspace associated with the mean vectors $\mu_{ji}$. In this work, $S = D$.

It is interesting to note that Equation 3 differs from Equation 1 in that state specific model means are formed as a weighted linear combination of multiple subspace projections whereas the adapted means in Equation 1 are obtained from a projection in a single subspace. In Equation 3, the subspace projection is actually an offset to the global mean vector, $\bar{\mathbf{m}}_i$. This is different from the expression for $\mu_{ji}$ in [6] where $\mu_{ji}$ is not dependent on the global mean. However, in practice the effect of the difference is minor. The state specific weights in Equation 3 are obtained from the state projection vector, $\mathbf{v}_i$, using a log-linear model,

$$w_{ji} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_i}{\sum_{k=1}^{I} \exp \mathbf{w}_k^T \mathbf{v}_j}. \quad (4)$$

Representing the mixture weights in Equation 4 in terms of a substate projection represents a significant departure from the speaker space model. It provides a mechanism for constraining the magnitude as well as the direction of movement with the model space. This representation is more robust than representing the weights as the normalized posterior mixture/state occupancy probabilities, $\gamma_{ji}(t)$ as is done in a CDHMM. With the number of mixtures in Equation 2 equal to over 100 and limited training data, a large portion of these posteriors will be near zero.

### Table 1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Shared</th>
<th>States</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>1700</td>
<td>1700</td>
<td>635K</td>
</tr>
<tr>
<td>256</td>
<td>1700</td>
<td>5000</td>
<td>635K</td>
</tr>
<tr>
<td>256</td>
<td>5005</td>
<td>5005</td>
<td>635K</td>
</tr>
</tbody>
</table>

To provide additional flexibility in the parameterization of the SGMM, the notion of a substate was introduced. In this case, the distribution for the observation in state $j$ is a weighted combination of densities

$$p(x|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} N(x; \mu_{jmi}, \Sigma_i), \quad (5)$$

where $c_{jm}$ is the relative weight of substate $m$ in state $j$ and the means and mixture weights are now obtained from state projection vectors, $\mathbf{v}_{jmi}$,

$$\mu_{jmi} = \bar{\mathbf{m}} + M_i \mathbf{v}_{jmi}, \quad (6)$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jmi}}{\sum_{k=1}^{I} \exp \mathbf{w}_k^T \mathbf{v}_{jmi}}. \quad (7)$$

Multiple substate per state SGMM models are realized in multiple iterations by splitting $\mathbf{v}_{jmi}$ vectors with the highest occupancy counts.

### 2.3. SGMM Parameterization

One of the most important practical aspects of the SGMM model described in Section 2.2 is that most of the parameters in the model are dedicated to the subspace representation: $\{ \mathbf{m}_i, \Sigma_i, M_i, w_i \}$, $i = 1, \ldots, I$. Table 1 shows example SGMM parameterizations for an SGMM model with $I = 256$ Gaussians and $J = 1700$ to 5005 states. It is assumed in Table 1 and for the systems evaluated in Section 4 that the feature vector dimension and the state projection vector dimension are the same, $D = S = 39$. The first row of the table shows that for the SGMM model with a single substate per state, roughly 90 percent of the parameters are shared. The second row shows that, even with a very large number of substates, roughly two thirds of the parameters are shared. Note that a diagonal covariance Gaussian CDHMM model with $J = 1700$ and 6 mixtures per state consists of approximately 797K parameters with almost all of these parameters being associated with the state dependent Gaussian densities.

This may facilitate training scenarios where only a relatively small number of observations are available for training state specific parameters, $\mathbf{v}_{j,i}, j = 1, \ldots, J$. The initialization of the SGMM parameters can be done in many ways. It is assumed here that $J$ and the definition of the context clustered states are known in advance. This is obtained from the hierarchical state clustering procedure used in training a prototype CDHMM. In [4], the initial values for $\mathbf{m}_i$ and $\Sigma_i$ are obtained by clustering the Gaussians from a fully trained prototype CDHMM. In this work, full covariance Gaussian mixtures are trained separately from the same training set used for SGMM training.

The subspace model parameters, $M_i$, $w_i$, and $\mathbf{v}_i$, can be initialized from a flat start where $M_i$ is initialized as an identity matrix, and $w_i$ and $\mathbf{v}_i$ are initialized to the zero vector. An alternative
method for initializing SGMM parameter estimation is investigated here. The goal of the alternative initialization method is to exploit the alignment between HMM states and Gaussian mixture indices. Instead of initializing the subspace model parameters, initial estimates are obtained for \( \gamma_{ji}(t) \), the posterior probability of the state index, \( s_t \), being equal to \( j \) and the Gaussian mixture index, \( m_t \), being equal to \( i \) given feature vector \( x_t \). These posteriors are in general estimated from the forward backward algorithm and are used for estimating the model parameters in Equations 2 to 4 and the joint state, substate, mixture posteriors, \( \gamma_{jm}(t) \), are used to update the parameters in Equations 6-7. SGMM training is initialized here using initial estimates of the posteriors, \( \gamma_{ji}(t) \), that are approximated as

\[
\gamma_{ji}(t) = p_i^0(s_t = j, m_t = i|x_t) \approx p_i^0(s_t = j|x_t)p_u(m_t = i|x_t). \tag{8}
\]

In Equation 9, the posteriors \( p_i^0(s_t = j|x_t) \) are obtained from forward-backward decoding of the training utterances with respect to the prototype CDHMM. The posteriors \( p_u(m_t = i|x_t) \) are obtained from the training utterances using the initial Gaussian mixture models. The impact of this initialization strategy will be considered in Section 4.

3. TASK DOMAIN

Unsupervised and supervised adaptation was performed on the Resource Management (RM) corpus under the following scenario. Acoustic SI CDHMMs and SGMMs were trained using 3990 utterances from 109 speakers taken from the standard RM SI-109 training set. Mel frequency cepstrum coefficient (MFCC) feature analysis was used. Feature vectors included 12 MFCC coefficients, normalized energy, and their first and second difference coefficients for a 39-dimensional feature vector. The subspace matrices for MLLB and CMLLB techniques were also trained from this 109 speaker training set. ASR WER was evaluated using 1200 utterances from 12 speakers taken from the RM speaker dependent evaluation (SDE) set. Baseline speaker independent (SI) CDHMM’s contained left-to-right 3-state clusters of triphones with 6 diagonal Gaussian mixtures per state for a total of 10,224 Gaussians.

The interest in this relatively constrained task domain is due to the fact that the major issues affecting performance are intrinsic sources of variability. The effects of acoustic environment and channel variability in this corpus are relatively minor. As a result, one can attribute reductions in WER to the impact of modeling techniques on these targeted sources of variability. Furthermore, the baseline CDHMM word error rate (WER) for this task domain is under five percent which is already reasonably low. Of course, the utterances in the corpus do not reflect the intrinsic sources of variability and the level of co-articulation variability that are present, for example, in conversational telephone speech domains. Future work will involve extending the study presented in Section 4 to more general and less constrained application domains.

The experimental results obtained in Section 4 for the CDHMM acoustic model relied on the HTK Toolkit for model training and recognition [7]. SGMM training and recognition was also implemented by updating HTK training and recognition tools. While differences between CDHMM and SGMM acoustic model parameterizations resulted in significant modifications to internal and external model representations, there is no fundamental reason why the SGMM acoustic model could not eventually be integrated into the existing HTK framework.

4. EXPERIMENTAL STUDY

This section describes the experimental study performed to evaluate the performance of the SGMM system on the resource management (RM) task domain. The study is described in two parts. First, the SGMM parameterization is considered in terms of its impact on efficiency, trainability, and performance with respect to HMM based systems. Second, the performance of this state based system is compared to the performance of the well known subspace based speaker adaptation described in Section 2.1.

4.1. Impact of SGMM Parameterization

Table 2 displays the word error rates obtained using CDHMM and SGMM systems configured with a range of parameter allocations for model states and substates. The first four rows of Table 2 show the WERs obtained using a baseline CDHMM with 1700 states and the SGMM models configured using the same number of states and with parameter allocations given in the first two rows of Table 1. The third row of Table 2, labeled “SGMM-FSinit”, displays the WER for the flat start initialization described in Section 2.3. All other SGMM results shown are initialized from joint state/mixture posteriors.

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>States</th>
<th>Subst.</th>
<th>Percent WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDHMM</td>
<td>1700</td>
<td>-</td>
<td>4.91</td>
</tr>
<tr>
<td>SGMM-FSinit</td>
<td>1700</td>
<td>1700</td>
<td>4.48</td>
</tr>
<tr>
<td>SGMM</td>
<td>1700</td>
<td>1700</td>
<td>4.26</td>
</tr>
<tr>
<td>SGMM</td>
<td>1700</td>
<td>5000</td>
<td>3.99</td>
</tr>
<tr>
<td>CDHMM</td>
<td>5005</td>
<td>-</td>
<td>6.24</td>
</tr>
<tr>
<td>SGMM</td>
<td>5005</td>
<td>5005</td>
<td>4.24</td>
</tr>
</tbody>
</table>

Table 2. WERs for multiple parameter allocations of SGMM and CDHMM

There are several observations that can be made from the top portion of Table 2. The first observation is that SGMM configurations with \( J = 1700 \) states obtain a WER reduction ranging from 8 percent to 18 percent. Second, comparing rows two and three, there is a small 5 percent WER reduction obtained by initializing SGMM training from joint posteriors relative to flat start initialization. Finally, by comparing the third and fourth rows of Table 2, it is clear that increasing from a single substate per state to approximately three substates per state, the SGMM WER is reduced by 6 percent.

With only a small number of projection vectors representing state level information, one might expect that it would be efficient to train these state level parameters with a relatively small number of effective observations per state. To investigate this conjecture, CDHMM and SGMM models with a much larger number of context clustered states were trained on the same data set used to train the original models. This was thought to be a better means for evaluating training efficiency than simply reducing the overall number of training utterances. There is less of a chance in this case of introducing artifacts that can arise from the highly skewed distribution of phonetic contexts that can occur with a very small corpus size. The efficiency of the SGMM models is demonstrated by comparing the WER for the 5005 state systems in rows four and five of Table 2. The WER obtained for the SGMM system represents a 32 percent reduction compared to the 5005 state baseline HMM model. This is a much greater reduction than was obtained for the 1700 state case and illustrates the robustness of the SGMM model with respect to sparseness in training data.
4.2. SGMM and Subspace Speaker Adaptation

In Section 2 the SGMM was described as being in the same class of modeling approaches as subspace based speaker adaptation. The purpose of this section is to demonstrate that the SGMM does in practice provide a more powerful representation by making direct comparisons to the CMLLB speaker space adaptation approach [2].

It is difficult to make direct comparisons between a subspace acoustic modeling formalism like the SGMM and subspace based speaker adaptation. In general, even unsupervised adaptation approaches assume that multiple speaker labeled utterances per speaker are available in training and that speaker dependent adaptation utterances are available prior to decoding. In speaker independent (SI) ASR, there is no speaker labeled training data and no adaptation data of any kind is used for updating model parameters prior to recognition. To minimize these differences, an unsupervised single utterance speaker adaptation scenario was used so no additional labeled or unlabeled data was used for adaptation other than the test utterance. Still, even a single utterance speaker adaptation scenario enjoys the basic advantage of estimating projection vectors from the test utterance while the SGMM parameters remain fixed.

Table 3 provides a comparison of the WER obtained for the SGMM with that obtained by applying the unsupervised speaker space adaptation CMLLB procedure [2] described in Section 2.1 to the baseline CDHMM model. MLLB and CMLLB were implemented using a $K=20$ dimensional subspace and were used to update the mean supervector corresponding to the concatenated means of the baseline 1700 state CDHMM model. For both MLLB and CMLLB, the speaker projection vector, $u^s$, was estimated from the test utterance as described in [2]. The number of parameters listed for these systems include the parameters associated with the CDHMM and the parameters associated with $E$ and $u^s$ in Equation 1.

<table>
<thead>
<tr>
<th>System</th>
<th>Parameters</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDHMM (1700 St.)</td>
<td>816K</td>
<td>4.91</td>
</tr>
<tr>
<td>CDHMM+MLLB</td>
<td>8.2M</td>
<td>4.53</td>
</tr>
<tr>
<td>CDHMM+CMLLB</td>
<td>1.62M</td>
<td>4.18</td>
</tr>
<tr>
<td>SGMM (1700 St.,5000 SubSt.)</td>
<td>975K</td>
<td>3.99</td>
</tr>
</tbody>
</table>

Table 3. WER comparison of SGMM with subspace adaptation

The following observations can be made from Table 3. First, there are an extremely large number of parameters used for MLLB adaptation. This is because of the dimensionality of the subspace matrix in Equation 1. Even though the subspace dimension is only 20, the supervector dimension is nearly 400K. Since CMLLB forms the columns of $E$ as a concatenation of a small number of a clustered subvectors, the parameters allocated to the subspace matrix is much smaller. Second, there is a six percent reduction in WER obtained for CMLLB system. This is attributed to better generalization properties associated with the more efficient representation of the subspace vectors. The last observation is that the SGMM obtains lower WER than the speaker space adaptation approaches despite the fact that the adaptation procedures require multiple passes over the test utterances to update projection vectors.

5. CONCLUSIONS

The SGMM model was described and compared to CDHMM based ASR and subspace based speaker adaptation on an LVCSR task. The SGMM demonstrated a reduction in WER of as high as 18 percent with respect to CDHMM for speaker independent ASR. These relative improvements are believed to result from the efficient parameterization of the SGMM which describes phonetic variability in terms of state specific projections in multiple linear subspaces. This was particularly important for the corpus used in this task which included only four hours of training speech. The SGMM also demonstrated reduced WER with respect to CMLLB subspace based speaker adaptation when CMLLB was applied to single utterance unsupervised speaker adaptation. In fact, the SGMM WER obtained for speaker independent ASR was lower than all adaptation scenarios, whether supervised or unsupervised, evaluated in [2].

6. ACKNOWLEDGEMENTS

The authors would like to thank all of the members of the 2009 CLSP summer workshop on this topic lead by Dan Povey for their valuable advice and suggestions. The authors would also like to thank Aanchan Mohan at McGill for his contribution to Gaussian mixture training and SGMM initialization.

7. REFERENCES