CROSS-LINGUAL CONTEXT SHARING AND PARAMETER-TYING FOR MULTI-LINGUAL SPEECH RECOGNITION

Aanchan Mohan, Richard Rose

Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

ABSTRACT

This paper is concerned with the problem of building acoustic models for automatic speech recognition (ASR) using speech data from multiple languages. Techniques for multi-lingual ASR are developed in the context of the subspace Gaussian mixture model (SGMM)[2, 3]. Multi-lingual SGMM based ASR systems have been configured with shared subspace parameters trained from multiple languages but with distinct language dependent phonetic contexts and states[11, 12]. First, an approach for sharing state-level target language and foreign language SGMM parameters is described. Second, semi-tied covariance transformations are applied as an alternative to full-covariance Gaussians to make acoustic model training less sensitive to issues of insufficient training data. These techniques are applied to Hindi and Marathi language data obtained for an agricultural commodities dialog task in multiple Indian languages.

Index Terms— Low-resource speech recognition, Subspace Methods, Multi-lingual speech recognition, Semi-tied Covariances, Indian languages

1. INTRODUCTION

There has been a great deal of interest in the problem of rapid configuration of automatic speech recognition (ASR) systems in under-resourced languages [4, 5, 6]. The scenario that is of particular interest in this paper is one where a voice enabled service has been developed for a given task domain in one language, and there is a need to develop a service for the same domain in another language [7, 8, 9]. Of course, it would be desirable to leverage as many resources as possible from the system developed for a given language to use in developing a system in a new language. The project,"Speech-based access for agricultural commodity prices in six Indian languages", sponsored by the Government of India, is a good example of where this resource sharing would be desirable [10]. The project aims to develop a nation-wide spoken dialogue system that allows Indian farmers to obtain prices of daily agricultural commodities in multiple languages. The work described in this paper focuses on minimizing the need for acoustic training data in configuring an ASR system in a new language for the Indian commodities dialog task. Given a reasonably large amount of Marathi language speech data collected for this domain, the goal is to leverage this data for configuring an ASR system in the linguistically similar Hindi language with only a very small amount of Hindi language speech data.

There are two techniques that are investigated to achieve this goal. Both techniques are applied in the context of subspace Gaussian mixture model (SGMM) based ASR [2]. The SGMM parametrization is summarized in Section 2 as consisting of state level acoustic probabilities that are formed from multiple subspace projections. Multilingual SGMM based ASR systems have been configured with shared subspace parameters trained from multiple languages but having distinct language dependent phonetic contexts and states [11, 12, 2]. The general architecture of these multilingual systems is described in Section 2.

The first technique investigated here involves sharing state level target language (Hindi) and foreign language (Marathi) SGMM parameters to improve speech recognition performance in the Hindi language. A procedure for identifying Marathi language states that are "similar" to a given target language state, the process of sharing these states, and weighting them appropriately are discussed in Section 4. The issue of sharing state-level SGMM parameters is related to previous work presented by Qian et al. in [14]. Their work involved "borrowing" speech segments from a well-resourced foreign language that were acoustically "similar" to segments in a target language. The acoustic similarity of segments was determined by using a distance measure between the target language and foreign language state-dependent parameters that had been decoded for those segments. This "borrowed" data was then used to update the state-dependent parameters in the target language. The approach presented here differs from this previous work in that it involves borrowing model parameters from the non-target language model to improve performance in the target language rather than borrowing foreign language data.

The second technique investigated involves reducing the number of shared parameters in the SGMM in an effort to make acoustic model training less sensitive to issues of insufficient training data. Semi-tied covariance (STC) transformations are used to replace the shared full covariance matrices generally used in SGMMs. Section 5 describes how the use of STC transformations is adapted from the well known approach for STC estimation in continuous density hidden Markov models (CDHMMs) [15]. The impact of STC transformations on ASR performance and model size is also presented in Section 5.

The work presented in this paper represents an extension of previous work in leveraging Marathi data in a multilingual SGMM architecture to improve ASR performance in Hindi for the Indian agricultural dialog domain [1]. A review of this previous work and a summary of the task domain is provided in Section 3. The corpus used for training and evaluation consisted of narrowband speech collected from actual rural users under varying background and environmental conditions and over varying mobile handsets and channels. It was shown in [1] that the multilingual SGMM system provided a 14.5 percent relative reduction in word error rate compared to the best monolingual SGMM or CDHMM systems. One particularly interesting aspect of the multilingual scenario investigated in this work is the linguistic similarity of the two languages, Marathi and Hindi. It is believed that because of their similarity, this language pair is an ideal case for the study of multilingual acoustic modeling.

This work was partially supported by the FQRNT.

2. SUBSPACE GAUSSIAN MIXTURE MODEL FOR MULTI-LINGUAL SPEECH RECOGNITION

This section provides a brief description of the implementation used in this paper for SGMM acoustic modelling [2] in multi-lingual speech recognition. The description of the SGMM in Section 2.1 follows the work of Rose et al.[3]. Section 2.2 describes the SGMM parametrization for multi-lingual speech recognition.

2.1. The subspace Gaussian mixture model

For an ASR system configured with J states, the observation density for a given D dimensional feature vector, $\boldsymbol{x}(t)$ for a state $j \in 1 \ldots J$ can be written as,

$$p(\boldsymbol{x}(t)|j) = \sum_{i=1}^{I} w_{ji} \mathcal{N}(\boldsymbol{x}(t)|\boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma_i}), \quad (1)$$

where I full-covariance Gaussians are shared between J states. The state dependent mean vector, $\boldsymbol{\mu}_{ji}$, for state j is a projection into the ith subspace defined by a linear subspace projection matrix $\boldsymbol{M}_i$,

$$\boldsymbol{\mu}_{ji} = \boldsymbol{M}_i \boldsymbol{v}_j. \quad (2)$$

In Eq(2), $\boldsymbol{v}_j$ is the state projection vector for state j. The subspace projection matrix $\boldsymbol{M}_i$ is of dimension $D \times S$ where S is the dimension of the state projection vector $\boldsymbol{v}_j$ for state j. In this work, $S = D$. The state specific weights in Eq.(2), are obtained from the state projection vector $\boldsymbol{v}_j$ using a log-linear model,

$$w_{ji} = \frac{\exp \boldsymbol{w_i^T} \boldsymbol{v}_j}{\sum_{i'=1}^{I} \exp \boldsymbol{w_{i'}^T} \boldsymbol{v}_j}. \quad (3)$$

In addition, to add more flexibility to the SGMM parametrization at the state level, the concept of substates is adopted where the distribution of a state can be represented by more than one vector $\boldsymbol{v}_{jm}$, where m is the substate index. This "substate" distribution is again a mixture of Gaussians. The state distribution is then a mixture of substate distributions which are defined as follows:

$$p(\boldsymbol{x}(t)|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\boldsymbol{x}(t)|\boldsymbol{\mu_{jmi}}, \boldsymbol{\Sigma_i}), \quad (4)$$

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

$$\boldsymbol{\mu}_{jmi} = \boldsymbol{M}_i \boldsymbol{v}_{jm}, \quad (5)$$

$$w_{jmi} = \frac{\exp \boldsymbol{w_i^T} \boldsymbol{v}_{jm}}{\sum_{i'=1}^{I} \exp \boldsymbol{w_{i'}^T} \boldsymbol{v}_{jm}}. \quad (6)$$

2.2. SGMM for Multi-lingual ASR

The SGMM for multi-lingual ASR was first proposed by Burget et al[11]. From the description of the SGMM formalism in Section 2.1, it is clear that the context-dependent HMM states "share" parameters with only a small number of parameters attached to each state. The parametrization for multi-lingual SGMM training is described in Figure 1. The figure shows a separate set of sub-word units $P_1^H, \ldots, P_K^H$ for Hindi and $P_1^M, \ldots, P_L^M$ for Marathi. This gives rise to separate sets of context-dependent HMM states for each language, and hence separate sets of language specific state-projection vectors namely $\boldsymbol{v}_1^H, \ldots, \boldsymbol{v}_R^H$ for Hindi and $\boldsymbol{v}_1^M, \ldots, \boldsymbol{v}_S^M$ for Marathi. The figure shows that although language-specific context dependent states are maintained, *all* of these context-dependent states *share* parameters , $\boldsymbol{M}_i$, $\boldsymbol{w}_i$ and $\boldsymbol{\Sigma}_i$.

For multi-lingual acoustic modelling, the shared parameters $\mathbf{M}_i$, $\mathbf{w}_i$ and $\boldsymbol{\Sigma}_i$, are trained by pooling data from multiple languages. Separate phone sets are maintained by adding a language specific tag to each phone and thereby each clustered HMM state. Each state projection vector $\boldsymbol{v}_j$, which is attached to a clustered state, is then trained only with data specific to each language.

Fig. 1. Parametrization of the Multi-lingual SGMM

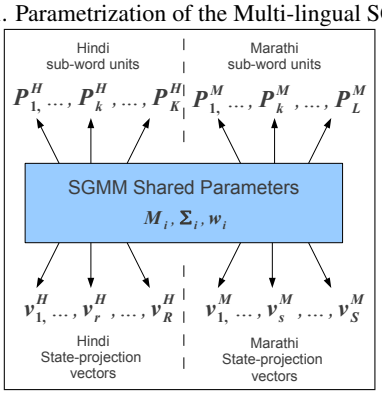


Table 1. Speech data used for experimental study

	Amount of data (Hrs.)		Characteristics	
Language	Train	Test	Words	Phones
Hindi	1.22	1.24	119	54
Marathi	7.94	2.5	551	60

3. HINDI AND MARATHI SYSTEMS

This section provides a description of the task domain and the data set, CDHMM mono-lingual systems, SGMM mono-lingual systems, and the multi-lingual SGMM system. The material in this section is an extension of earlier work appearing in[1, 12].

3.1. Agricultural Commodities Task Domain

Table 1 provides a summary of the subset of the Indian agricultural commodities spoken dialog corpus introduced in Section 1.The first two columns of Table 1 display the number of hours of speech data for each language, Hindi and Marathi. Hindi, spoken predominantly in northern India, and Marathi, spoken predominantly in the western part of India, are considered to be linguistically similar. The last two columns of Table 1 show the number of words and the number of phones in the Marathi and Hindi data sets. Note that this is a relatively small vocabulary task consisting primarily of the names of commodities and geographical regions.

3.2. Hindi and Marathi Continuous Density HMM Systems

The baseline CDHMM systems were based on conventional three state left-to-right HMM triphone models. Clustered states were obtained after decision tree clustering for the systems in each language. Acoustic models were trained with 161 clustered states for Hindi and 476 clustered states for Marathi. A fixed number of 16 Gaussians per state were used for Marathi, and 8 Gaussians per state were used for Hindi. The features used were 13 MFCC coefficients concatenated with first and second difference cepstrum coefficients. Mean normalization was performed for each utterance. The CDHMM baseline systems in each language were configured using the HTK Speech Recognition Toolkit [17].

The Hindi ASR performance was evaluated on utterances of commodity names [1]. For Hindi, a separate language model is trained for recognizing either commodities or districts, since these represent dialog states in the Hindi spoken dialog system. For Marathi on the other hand, a single language model is trained for recognizing both commodities and districts.

The CDHMM and SGMM ASR performance for the Hindi and Marathi language test sets are presented in Tables 2 & 3 respectively. The percentage of words correctly decoded (%Corr) is used as the

performance measure for recognition. Word Accuracy (%WAc.) figures have also been provided for comparison. The baseline CDHMM results are presented in the first row of Tables 2 & 3 for Hindi and Marathi respectively. The Hindi system with only one hour of data is seen to have a performance of 68.7% Corr., while the Marathi system with close to 8 hours of data is seen to have a performance of 82.2%Corr.

The impact of performing speaker adaptive training (SAT) during CDHMM training and constrained maximum likelihood linear regression (CMLLR) adaptation during recognition is displayed in the second row of Tables 2 & 3. In the case of Hindi, a relative improvement of 3.34%Corr. with respect to baseline CDHMM is observed. In the case of Marathi, a much larger relative improvement of 9.73%Corr. with respect to the baseline is observed.

3.3. Hindi and Marathi monolingual SGMM systems

The training for the monolingual SGMM in this work is summarized in [1]. For Hindi, the monolingual SGMM has $I = 256$ full covariance Gaussians while the Marathi system consists of $I = 400$ Gaussians. The Gaussians are obtained by training on speech-only segments from training corpora of the respective languages. Clustered tri-phone states for each language are obtained from the CDHMM. The basic setup is identical to that used for the CDHMM systems. For the SGMM, an implementation is used that is an extension to HTK with added libraries [3]. Training of the SGMM is carried out using a joint posterior initialization as described in [3]. The SGMM system performance in each language, with the number of mixtures I used in each system, is presented in row 3 of Tables 2 & 3 for Hindi and Marathi respectively. It is seen that the mono-lingual SGMM systems provide a performance of 71%Corr. in the case of Hindi and 84.7%Corr. in the case of Marathi. In both cases a consistent net performance gain is observed with respect to the CDHMM baseline.

3.4. Hindi-Marathi Multi-lingual SGMM system

SGMM training was carried out in a multi-lingual fashion, by using the Marathi and Hindi data to train the "shared" parameters, while maintaining distinct phone sets for the two languages. Maintaining separate phone sets, as illustrated in Figure 1, allows for the use of language specific states. It also allows for state specific parameters to be trained using data from a single language. The multilingual SGMM has a total of $J = 637$ states, with 161 states coming from Hindi and 476 coming form the Marathi system. The system was initialized using a UBM with $I = 400$ Gaussians trained on speech-only segments from both corpora. The system was initialized with a joint posterior initialization procedure similar to the monolingual SGMM systems trained for this task.

To account for the acoustic mismatch across the two sets of data, multilingual SGMM training was carried out *after* cross-speaker and cross-corpus normalization. This acoustic normalization (AN) procedure is described in [12]. The results for the performance of multilingual SGMM for Hindi and Marathi test data are reported in row 4 of Tables 2 & 3 respectively. Cross-corpus acoustic normalization (AN)[12], was found to be critical for multi-lingual SGMM training. Acoustic normalization is denoted by +AN in Tables 2 & 3. The final multi-lingual systems are seen to provide performance of 76.2%Corr. on the Hindi test set, and performance of 92%Corr. on the Marathi test set. The multi-lingual system is thus seen to give the best performance across *both* languages.

4. CROSS-LINGUAL CONTEXT SHARING

The experimental study in Section 3 demonstrated that the multilingual SGMM is able to improve ASR performance by allowing a structured and shared parametrization for multiple languages. However, despite the improvements obtained from this cross-language data-sharing, regularly occurring recognition errors still exist for the target language. Examples of the acoustic confusions that cause these errors are given in Section 4.1. To mitigate these errors an experimental study is presented that considers the effect of borrowing "similar" context-dependent states from the more well-resourced Marathi language. Section 4.3 details the procedure employed for automatically selecting these "similar" states for each target language (Hindi) state. The non-target language (Marathi) states are then combined with the existing target language state using the SGMM concept of a sub-state. The sub-state weights that control the relative weighting of the target language and non-target language states are then globally varied over a range of values. The effect of varying these weights on the final system performance is presented in Section 4.4.

Table 2. ASR WAc for Hindi. I represents number of SGMM mixtures and AN indicates acoustic normalization.

Hindi Language ASR Performance		
System	I	% Corr. (%WAc)
CDHMM	n/a	68.7 (65.8)
CDHMM+SAT	n/a	71.0 (67.7)
Mono-lingual SGMM	256	71.0 (69.1)
Multi-lingual SGMM + AN	400	**76.2** (**74.4**)

Table 3. ASR WAc for Marathi. I represents number of SGMM mixtures and AN indicates acoustic normalization.

Marathi Language ASR Performance		
System	I	% Corr. (%WAc)
CDHMM	n/a	82.2 (74.3)
CDHMM+SAT	n/a	90.2 (85.3)
Mono-lingual SGMM	400	84.7 (77.7)
Multi-lingual SGMM + AN	400	**92.0** (**87.5**)

4.1. Typical acoustic confusions during Hindi ASR

It was found that the recognizer consistently made errors of the kind that are listed below:

- `TARBOOJ = t a r b oo j`, meaning watermelon in Hindi, is mis-recognized 6 times out of 13 occurrences of this term as musk-melon or `KARBUJA = kh a r b uu j aa`
- `ANANNAS = a n a nn a s`, meaning pineapple, is mis-recognized 13 times out of 18 occurrences of this term as pomegranate or `ANAR = a n aa r`
- `BADSHAH = b aa d sh a h`, a variety of potato, is mis-recognized 4 times out of 8 occurrences of this term as `BAJRA = b aa j r aa` or millet.
- `SUKHI MIRCH`, appearing as two separate words with the composite lexical expansion `s u kh ii m i r c` meaning dried red-chilli peppers, is mis-recognized 10 times out of 19 occurrences of this term as `SUPERIOR=s u p i r i-o r`, a variety of cardamom.

Errors of the kind listed above seem reasonable since the hypothesized word bears some likeness in pronunciation to the reference string. However, the repetitive nature of the errors for each of the examples listed here makes it clear that certain tri-phone contexts that are seen more often in training are favoured over more rarely occurring tri-phone contexts during recognition. For example it appears

that the context sil-kh+a is favoured over the context sil-t+a in the first example, or the context aa-r+sil is favoured over a-nn+a in the second example.

One possible approach for reducing these Hindi acoustic confusions would be to borrow context dependent state-projection vectors from the more well-trained Marathi states. These borrowed Marathi state projection vectors could then be combined with the more poorly trained Hindi state projection vectors. To do this, it is necessary to first determine which Marathi states should be associated with Hindi states. Then one must determine how the two state projection vectors must be combined.

4.2. Impact of borrowing non-target language contexts

An anecdotal "cheating" experiment is presented here where the states of the context-dependent phonetic unit of the Hindi language directly use state-dependent parameters of "similar" context-dependent phonetic units of the Marathi language. The goal of this experiment is to determine the potential performance gains that might be achieved by combining state-level SGMM parameters across languages. The choice of which state-level parameters to combine is made by exploiting the actual misrecognitions observed on the test data. This is, of course, the cheating aspect of the experiment. The method for combining the state-level parameters is described below. The experiment provides a way to decouple the issue of which acoustic contexts to combine from the choice of the method used for combining them. In this experiment, the concept of the sub-state in the SGMM is used to allow Hindi context-dependent HMMs to use the state dependent parameters from context-dependent HMMs of the Marathi language. This allows well-trained "similar" phonetic context-dependent HMMs in the well-resourced language to "augment" existing context-dependent phones in the target language. The experiment proposed here allows us to investigate the potential for sharing sub-state projection vectors at the state-level between context-dependent models from a non-target language and context-dependent models in the target language.

This corpus offers a unique opportunity of this kind because a large proportion of tri-phone contexts appearing in the confusable Hindi words listed in the above example also appear in the Marathi language as well. This degree of cross-language context sharing is not at all typical. However, it provides a chance to observe the potential effects of cross-language parameter sharing when the determination of similar contexts is not an issue. After looking up the equivalent context-dependent models in Marathi, the state projection vector from each state from the Marathi model is "augmented" as a sub-state in the corresponding state for the Hindi model. The sub-state weights are set equally between the existing sub-state vector and the newly introduced sub-state vector from the Marathi model. The state dependent likelihoods for these "augmented" states are computed as given by Equation 4. The multi-lingual SGMM model thus obtained is referred to as the "augmented" model in Table 4. A similar procedure was carried out to modify the target language silence model.

N-Best lists with $N = 2$ were generated for the occurrences of misrecognized utterances listed in Section 4. This generated the possible hypotheses as being either the correct term or the closest incorrect term. The first row of Table 4 displays the percentage of these utterances that the multi-lingual SGMM model recognizes correctly. These N-Best lists were then re-scored with the "augmented" models. The proportion of utterances that were correctly recognized after this re-scoring procedure appears in the second row of Table 4.

It can be seen from Table 4 that the so-called "augmented" models provide a 20% absolute improvement over the multi-lingual SGMM Hindi models in correctly recognizing the originally misrecognized test utterances. While this result must clearly be considered as anecdotal, it suggests that there may be considerable potential for state-level cross-language parameter sharing in multilingual SGMMs. The next section presents a procedure for automatically associating the appropriate non-target language states with the appropriate target language states.

Table 4. Recognition accuracy for frequently mis-recognized utterances

ML-Hindi SGMM	56%
"Augmented" ML-Hindi SGMM	76%

4.3. Algorithm for selecting states from the non-target language

This section describes a method for selecting potential non-target language context-dependent states that are "similar" to target language states. The effectiveness of a cosine distance between state projection vectors as a measure of similarity between states is illustrated in [12]. This simple cosine-distance $s(h, m)$ between normalized (refer Appendix K of [18]) state projection vectors $\boldsymbol{v}_j^h$ and $\boldsymbol{v}_j^m$,

$$s(h,m) = \frac{\boldsymbol{v}_j^h . \boldsymbol{v}_j^m}{\| \boldsymbol{v}_j^h \| \| \boldsymbol{v}_j^m \|}, \tag{7}$$

is used here to identify similar cross-language states.

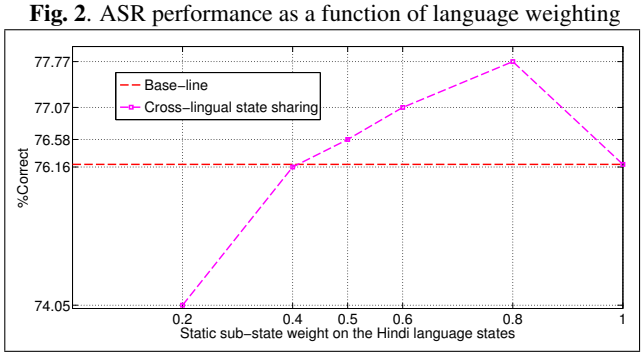
As a first step, for each target language (Hindi) state, a list of 10 similar non-target language (Marathi) states are picked based on the cosine distance metric. The evaluation of the similarity is restricted only to those non-target language states that appear in the same state position in a left-to-right HMM topology. As a second refinement step, a log-likelihood based re-ranking of this list of potential states for each target language state is carried out. The re-ranking procedure involves doing a forced alignment pass over the entire set of Hindi training utterances. During alignment, for each speech frame, when a certain target language state is encountered, a record of the frame log-likelihood for each potential non-target language state is maintained. This log-likelihood for the potential non-target state is accumulated over all such occurrences. Finally, the average log-likelihood for each potential non-target language state is calculated over all such occurrences and the list of potential non-target states is sorted. The top-ranking non-target language state is then selected as the non-target language state of choice.

4.4. Experimental Results

Figure 2 summarizes the Hindi language ASR performance with cross-lingual state sharing. To recall, with the use of sub-states, the SGMM state-likelihood is calculated as mentioned in Equation 4. In this experiment, with reference to Equation 4, $M_j = 2$ and $m \in \{1 = \text{Hindi}, 2 = \text{Marathi}\}$. Here we let c_{jH} to denote the target-language (Hindi) sub-state weight and c_{jM} denote the non-target language (Marathi) sub-state weight. The performance curve is displayed as the static target language sub-state weight c_{jH} is varied between 0 and 1. The "closest" non-target language states are automatically determined using the procedure mentioned in Section 4.3. On the extreme end when the target language sub-state weight is set to 0.0, only the "closest" Marathi-language states are used to evaluate ASR performance. This is not shown in the figure, but as expected, the performance at 57.01%Corr. is well below the baseline. As the weighting towards the Hindi language states increase, the performance is seen to increase. The best performing system at 77.77%Corr. is obtained when the Hindi language states are weighted at 0.8. At this point, an improvement of 1.57%Corr. absolute is seen with respect to the SGMM baseline of 76.2%Corr.

The matched-pairs significance test described in [19] was run, and the improvement in performance with respect to the SGMM baseline was statistically significant at the chosen confidence level of 99.99%.

Fig. 2. ASR performance as a function of language weighting



Regarding the estimation of the weights, a maximum-likelihood estimation of the weights on the training data would lead to the degenerate solution where the target language state is weighted as 1.0 and the non-target language state is weighted as 0.0. A valid method to estimate these weights would be to use deleted-interpolation using N-fold cross-validation as mentioned in the work by Huang et al. [20]. The appropriate estimation of these weights is currently under study.

5. SEMI-TIED COVARIANCES FOR THE SGMM

Large-vocabulary continuous speech recognition (LVCSR) SGMM systems are known to have fewer parameters compared to their CDHMM system counterparts [2]. However, in small-vocabulary systems this is generally not true. This is predominantly due to the fact that the number of CDHMM states, J, is much smaller in small vocabulary systems. For example, the mono-lingual Hindi CDHMM system has 101752 parameters whereas the mono-lingual Hindi SGMM system with $I = 256$ mixtures has 605319 parameters. Furthermore, the parameter count is dominated in these small vocabulary SGMM system by the number of shared parameters.

In an attempt to reduce the shared parameter count and to make acoustic model training less sensitive to issues of insufficient training data, the effect of using semi-tied covariance (STC) matrices instead of full-covariance matrices is studied. STCs have predominantly been used with CDHMM systems and are introduced in Section 5.1. Section 5.2 describes the issues involved in using STCs with the SGMM. Finally, Section 5.3 presents a performance comparison between the use of STC transformations and full-covariance matrices in the multi-lingual SGMM system evaluated on the Hindi test set.

5.1. Semi-tied Covariance Modelling

In semi-tied covariance modelling, a full-covariance matrix for a mixture component i assigned to belong to one of $r = 1 \ldots R$ regression classes is expressed as:

$$\boldsymbol{\Sigma}_i = \boldsymbol{T}_r \boldsymbol{\Sigma}_i^{(diag)} \boldsymbol{T}_r^T \tag{8}$$

The matrices $\boldsymbol{T}_r$ for a regression class r are called the semi-tied transforms. The model structure changes by allowing dedicated diagonal components for each covariance mixture component. It is usually easier to work with the inverse of the semi-tied transform $\boldsymbol{A}_r = \boldsymbol{T}_r^{-1}$. The semi-tied covariance transforms are trained in a maximum-likelihood sense given the current acoustic model parameters. The optimization procedure based on the formulation of the Expectation Maximization (EM) algorithm for estimating STC transformations in the CDHMM is presented in [15].

5.2. Integrating semi-tied covariances with the SGMM

This section mentions some of the practical issues involved in using semi-tied covariance modelling with the SGMM.

5.2.1. SGMM initialization

In the conventional SGMM the shared full-covariance matrices are initialized from a so-called universal background model(UBM). In a similar manner, the SGMM component covariances $\boldsymbol{\Sigma}_i^{(diag)}$ and the associated STC matrix $\boldsymbol{A}^r$ is initialized from the estimates obtained from a UBM with semi-tied covariance Gaussians (STC-UBM) as:

$$\boldsymbol{\Sigma}_i^{(diag)} = \bar{\boldsymbol{\Sigma}}_i^{(diag)} \tag{9}$$

$$\boldsymbol{A}^r = \bar{\boldsymbol{A}}^r \tag{10}$$

In the above equation $\bar{\boldsymbol{\Sigma}}_i^{(diag)}$ and $\bar{\boldsymbol{A}}^r$ denote the covariances and STC transformations in the STC based UBM. While training the STC-UBM, Gaussians can be clustered into regression classes $r \in \{1, \ldots, R\}$ by agglomerative clustering [21] of the diagonal covariance UBM Gaussians. Also, estimates of the initial posteriors $\gamma_{j,i}(t)$, where j is used to denote a context-dependent state, are approximated using the STC-UBM.

5.2.2. Likelihood calculation

In order to compute likelihoods efficiently it is useful to maintain a version of the subspace matrices $\boldsymbol{M}_i$ pre-multiplied with the associated regression matrices $\boldsymbol{A}^r$, given by $\boldsymbol{B}_i = \boldsymbol{A}^r \boldsymbol{M}_i$. The quantities for the Gaussian likelihood computation as given in [2] are computed as follows,

$$\boldsymbol{x}_r(t) = \boldsymbol{A}^r \boldsymbol{x}(t) \tag{11}$$

$$n_{ji} = \log w_{ji} - \frac{1}{2}(D \log 2\pi + \log |\boldsymbol{\Sigma}_i^{(diag)}| + \boldsymbol{v}_j^T \boldsymbol{B}_i^T \boldsymbol{\Sigma}_i^{(diag)-1} \boldsymbol{B}_i \boldsymbol{v}_j) \tag{12}$$

$$n_i(t) = \log |\boldsymbol{A}^r| - 0.5 \boldsymbol{x}_r^T(t) \boldsymbol{\Sigma}_i^{(diag)-1} \boldsymbol{x}_r(t)$$

$$\boldsymbol{z}_i(t) = \boldsymbol{B}_i^T \boldsymbol{\Sigma}_i^{(diag)-1} \boldsymbol{x}_r(t) \tag{13}$$

$$\log p(\boldsymbol{x}(t), i|j) = n_{ji} + n_i(t) + \boldsymbol{z}_i(t).\boldsymbol{v}_j \tag{14}$$

Equation (14) gives the expression for the log-likelihood for state j, and Gaussian i. The process of evaluating the likelihoods over all possible mixtures $i \in \{1, \ldots, I\}$, for each state j is time-consuming. Instead, an approximation over the top-N mixture components for each observation $\boldsymbol{x}(t)$ is used. These top-N Gaussians are picked by evaluating likelihoods against the UBM in a process called Gaussian pre-selection[2, 3]. Here the Gaussian pre-selection is carried out by evaluating likelihoods against the STC based UBM.

5.2.3. SGMM statistics accumulation and model update

Following [2], care needs to be taken to compute the quantities that depend on $\boldsymbol{\Sigma}_i$ correctly during the EM (and sub-state splitting) computations. These are:

$$\boldsymbol{y}_j = \sum_{t,i} \gamma_{j,i}(t) \boldsymbol{z}_i(t) = \sum_{t,i} \gamma_{j,i}(t) \boldsymbol{B}_i^T \boldsymbol{\Sigma}_i^{(diag)-1} \boldsymbol{x}_r(t) \tag{15}$$

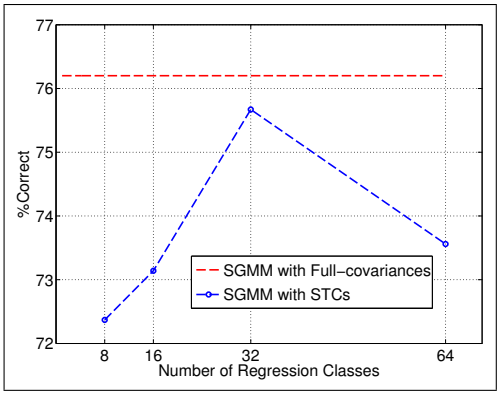
$$\boldsymbol{H}_i = \boldsymbol{B}_i^T \boldsymbol{\Sigma}_i^{(diag)-1} \boldsymbol{B}_i \tag{16}$$

The updates for the $\boldsymbol{\Sigma}_i^{(diag)}$ and $\boldsymbol{A}^r$ are carried out after all of the SGMM model parameters have been updated. The same STC iterative update algorithm as mentioned by Gales [15] is used to update these quantities.

5.3. Experimental results

Figure 3 compares the performance of the full-covariance SGMM multi-lingual model to the STC-SGMM multi-lingual models with a varying number of regression classes. All systems are configured with $I = 400$ mixtures and the evaluation is run on the Hindi test set. The following conclusions can be drawn from this experiment: (1) There appears to be an optimum number of regression classes, $R = 32$, that are required to be created when using semi-tied covariances. (2) The performance of the STC-SGMM at 75.8% with $R = 32$ regression classes is comparable with the performance of the full-covariance SGMM at 76.2%. (3) The total number of STC-SGMM covariance parameters when $R = 32$ is 64,272 which is far fewer compared to the total number of full-covariance parameters which is 312,000. It is clear that the SGMM trained using STC is far more compact than the full-covariance SGMM. However, it was surprising to find that the ASR performance for the STC based model, trained on the multi-lingual data set, is almost identical to the full-covariance system.

Fig. 3. STC SGMM models vs Full-covariance SGMM models



6. CONCLUSION

This paper has investigated approaches for state level cross-lingual parameter sharing and shared Gaussian parameter tying in multi-lingual SGMM based ASR. Experiments were performed using Hindi speech data as the target language and Marathi speech data as the foreign language, both obtained for an Indian agricultural commodities spoken dialog task. A relative decrease in the percentage of words incorrectly decoded of 4.7% for the Hindi language was obtained by combining similar Hindi language and Marathi language SGMM sub-state projection vectors. Applying STC transformations resulted in a reduction in the number of Gaussian parameters by a factor of five relative to full-covariance Gaussians while maintaining similar ASR accuracy.

7. REFERENCES

[1] A. Mohan, S. Umesh, and R. C. Rose, "Subspace based acoustic modelling in Indian languages," in *IEEE Conference on Information Science, Signal Processing and their Applications, Montreal, Canada*, 2012.

[2] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, et al., "The subspace Gaussian mixture model A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[3] R C Rose, S. Yin, and Y. Tang, "An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition," in *ICASSP*, 2011.

[4] R. Rasipuram, P. Bell, and M. Magimai.-Doss, "Grapheme and multilingual posterior features for under-resource speech recognition: A study on Scottish Gaelic," in *ICASSP*, 2013.

[5] N. T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *ICASSP*, 2011.

[6] A. Mandal, D. Vergyri, M. Akbacak, C. Richey, and A. Kathol, "Acoustic data sharing for Afghan and Persian languages," in *ICASSP*, 2011.

[7] M. Plauché, O Cetin, and U. Nallasamy, "How to build a spoken dialog system with limited (or no) language resources," in *Proc. IJCAI Workshop on AI in ICT for Development*, 2006.

[8] A. Waibel, P. Geutner, L.M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

[9] U. Bub, J Kohler, and B. Imperl, "In-service adaptation of multilingual hidden Markov models," in *ICASSP* 1997.

[10] G.V. Mantena, S. Rajendran, B. Rambabu, S.V. Gangashetty, B. Yegnanarayana, and K. Prahallad, "A speech-based conversation system for accessing agriculture commodity prices in Indian languages," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2011.

[11] Lukas Burget et al., "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *ICASSP* 2010.

[12] A. Mohan, S. Hamidi Ghalehjegh, and R. C. Rose, "Dealing with acoustic mismatch for training multilingual subspace Gaussian mixture models for speech recognition," in *ICASSP* 2012.

[13] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* 2011.

[14] Y. Qian, D. Povey, and J. Liu, "State-level data borrowing for low-resource speech recognition based on subspace GMMs," in *Interspeech*, 2011.

[15] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[16] India. Central Hindi Directorate, *Devanagari: development, amplification, and standardisation*, Central Hindi Directorate, Ministry of Education and Social Welfare, Government of India, 1977.

[17] S. Young et al., "The HTK book (for HTK version 3.4)," 2006.

[18] D. Povey, "A tutorial-style introduction to subspace Gaussian mixture models for speech recognition," *Microsoft Research, Redmond, WA*, 2009.

[19] L. Gillick and SJ Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP* 1989.

[20] XD Huang, Mei-Yuh Hwang, Li Jiang, and Milind Mahajan, "Deleted interpolation and density sharing for continuous hidden Markov models," in *ICASSP* 1996.

[21] H. Gish, M.H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *ICASSP* 1991.