Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain

Aanchan Mohan a,*, Richard Rose a, Sina Hamidi Ghalehjegh a, S. Umesh b

a Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada
b Indian Institute of Technology – Madras, Chennai, India

Available online 24 July 2013

Abstract

In developing speech recognition based services for any task domain, it is necessary to account for the support of an increasing number of languages over the life of the service. This paper considers a small vocabulary speech recognition task in multiple Indian languages. To configure a multi-lingual system in this task domain, an experimental study is presented using data from two linguistically similar languages – Hindi and Marathi. We do so by training a subspace Gaussian mixture model (SGMM) (Povey et al., 2011; Rose et al., 2011) under a multi-lingual scenario (Burget et al., 2010; Mohan et al., 2012a). Speech data was collected from the targeted user population to develop spoken dialogue systems in an agricultural commodities task domain for this experimental study. It is well known that acoustic, channel and environmental mismatch between data sets from multiple languages is an issue while building multi-lingual systems of this nature. As a result, we use a cross-corpus acoustic normalization procedure which is a variant of speaker adaptive training (SAT) (Mohan et al., 2012a). The resulting multi-lingual system provides the best speech recognition performance for both languages. Further, the effect of sharing “similar” context-dependent states from the Marathi language on the Hindi speech recognition performance is presented.

© 2013 Elsevier B.V. All rights reserved.

Keywords: Automatic speech recognition; Multi-lingual speech recognition; Subspace modelling; Under-resourced languages; Acoustic-normalization

1. Introduction

With the proliferation and penetration of cellular telephone networks in remote regions, a larger proportion of the planet’s population has inexpensive access to meet its communication needs. This has resulted in a wider audience, especially under-served populations, that have access to telephony based information services. Spoken dialog (SD) systems provide a natural method for information access, especially for users who have little or no formal education. Among the challenges facing the development of such systems is the need to configure them in languages that are under-resourced.

Agriculture provides a means of livelihood for over 50% of India’s population. Most of India’s farming is small scale; 78% of farms are five acres or less (Patel et al., 2010; Singh et al., 1999). Indian farmers, a proportion of whom are illiterate, face a host of challenges such as water shortages, increasing cost of farm supplies and inequitable distribution systems that govern the sale of their produce. Access to information through information and communication technologies (ICT) is often seen as a solution to enable and empower rural Indian farmers. There have been many noted efforts in this direction to develop ICTs by members of the private sector, non-governmental organizations (NGOs) and the government. E-Choupal (Bowonder et al., 2003), Avaaj Otalo (Patel et al., 2010), and the Mandi Information System (Mantena et al., 2011; Shrishrimal et al., 2012) are examples of such efforts that have been undertaken in the past or are currently ongoing. In addition, it is noteworthy to mention the evaluation
done by Plauché and Nallasamy (2007) to assess the factors involved in setting up a low-cost, small-scale spoken dialogue (SD) system to disseminate agricultural information to farmers in rural Tamil Nadu, India.

The work reported on the Mandi Information System is part of a larger effort called “Speech-based Access for Agricultural Commodity Prices in Six Indian Languages” initiated by the Government of India (Mantena et al., 2011; Shrishrimal et al., 2012). The project involves development of spoken dialog systems to access prices of agricultural commodities in various Indian districts. This information, updated daily, is also made available online through an Indian government web-portal www.agmarknet.nic.in. Needless to say, configuring SD systems especially in any or most of these languages is hardly a trivial task. Also, a large investment in time, money and effort is needed to collect, annotate and transcribe speech data required to develop the automatic speech recognition (ASR) engine that the SD systems depend upon.

Our goal in this paper is not to describe the development of the dialog system or the data collection effort itself, but to describe acoustic modelling configurations that would make the process of developing such systems more efficient. We use a subset of the data provided to us by the teams involved in the project “Speech based Access for Agricultural Commodity Prices in Six Indian Languages” for our experimental study. In terms of the target population for the service, the recording conditions of the speech data, the nature of the task (small-vocabulary) and the service itself is similar to the work described by authors in Plauché and Nallasamy (2007). Section 2 provides a description of the data, which was provided to us at a time when the development of the dialog systems was still in progress. We restrict our multi-lingual experimental study to two out of six languages – Hindi and Marathi. With limited available training data in Hindi, and the existing sources of acoustic, background and channel variabilities associated with the data collected for each language, the data set provided to us from this task domain poses some unique challenges. The speech data is “real-world”, in that it has been collected in conditions that shall be experienced in the actual use of the service. It is worth noting that the speech data was collected from a population whose education levels were representative of the target population who would actually use the service.

Our work here is motivated by previous work in developing the subspace Gaussian mixture model (SGMM) (Povey et al., 2011) and the use of the SGMM for multi-lingual speech recognition reported by Burget et al. (2010) and Lu et al. (2011, 2012). Other approaches to training multi-lingual acoustic models involve the use of a so-called common phone set as a means of sharing data between languages (Schultz and Waibel, 2001; Schultz and Kirchhoff, 2006; Vu et al., 2011). Though the use of a common phone set is a popular approach to train multi-lingual acoustic models, it is also quite cumbersome. Since the parametrization of the SGMM naturally allows for the sharing of data between multiple languages, the use of a common phone set is unnecessary, and hence this approach is attractive. The results from previous work in Burget et al. (2010) imply that when there is a limited amount of data in a certain target language, the SGMM parametrization facilitates the use of “shared” parameters that have been estimated reliably from other well-resourced languages for speech recognition in the target language. Authors in Lu et al. (2011) examine the effect that limited training data can have on the estimation of the state-dependent parameters of the SGMM and suggest the modification of using a regularization term for this purpose. Further, authors in Lu et al. (2012) present a maximum a-posteriori (MAP) adaptation method to “adapt” an SGMM acoustic model trained in a related but well-resourced language to a target language in which limited training data is available. The speech data used in all of these studies were collected in noise-free environments, and over close-talking microphones. Given our task domain, it seems appropriate to perform an experimental study detailing some of the issues involved in building SGMM based systems for both mono-lingual and multi-lingual systems for speech data intended for a practical application and in languages that are of interest to a large-part of the developing world.

In this paper we demonstrate the effects of multi-lingual SGMM acoustic model training carried out after cross-corpus acoustic normalization. A study is performed using the Marathi and Hindi language pair since they are linguistically related languages. Our work shows the importance of compensating for the sources of acoustic variability between speech data collected from multiple languages while training multi-lingual SGMM models. The issue of handling speaker and environmental “factors” that are causes for variation in the speech signal has been addressed in Gales (2001a). Further, in Seltzer and Acero (2001) the authors propose the factoring of speaker and environment variability into a pair of cascaded constrained maximum-likelihood linear regression (CMLLR) transforms-one for each source of variability. They propose an adaptive training frame-work to train the acoustic model. Our approach to cross-corpus acoustic normalization is similar in spirit to the approach presented by the authors in Seltzer and Acero (2001) in using a pair of “factored transforms” to compensate for speaker and environmental variability. Further, our work addresses the issue of not having enough well-trained context-dependent states in the Hindi language. To address this issue, context-dependent states in the multi-lingual SGMM are borrowed from the more well-resourced Marathi language. This is complementary to the work presented by Lu et al. (2011), where a regularization term is introduced in the optimization criterion for the state-dependent parameters. This regularization is introduced as means of dealing with limited data in the target language.

To provide the context for this study, we present baseline results for mono-lingual continuous density hidden
Markov model (CDHMM) speech recognition performance for Hindi and Marathi in this task domain. We compare the ASR performance of the SGMM based mono-lingual system with respect to the CDHMM baseline (Mohan et al., 2012b). Our experimental study considers the impact of each of the stages of multi-lingual training on the final ASR system performance. Interesting anecdotal results are presented to show that the SGMM’s state-level parameters are able to capture phonetically similar and meaningful information across the two languages. Further, recognition errors made by the final multi-lingual SGMM system on the Hindi test set that are attributed to a lack of adequate context-dependent states are analysed. To this effect, an experimental study that demonstrates the impact of borrowing context-dependent states from the Marathi language is presented. The main contributions of this work include the development of the Multilingual SGMM system for Hindi and Marathi, cross-corpus normalization for multi-lingual training, an analysis of the linguistic similarity between the two languages and the cross-lingual borrowing of contexts from the Marathi (non-target, well-resourced) language.

This paper is organized follows. Section 2 presents a detailed description of the agricultural commodities task domain. Next, Section 3 briefly describes the SGMM in the context of the experimental study. In Section 4 we describe our experimental setup for both the CDHMM and SGMM systems. The comparative performance of baseline mono-lingual CDHMM systems and the mono-lingual SGMM systems is described in Section 5. In Section 6 we provide a description of training and the performance of the multi-lingual SGMM system for Hindi and Marathi, highlighting the effects of acoustic mismatch between the two languages. After describing an algorithm to obtain features normalized for cross-speaker and cross-corpus acoustic variation, we consider the impact of using these for multi-lingual SGMM training. We summarize the performance of the multi-lingual Hindi and Marathi ASR system obtained with these normalized features. Further, an anecdotal experiment is presented in Section 6.4, where it is shown that errors arising due to poor context modelling in Hindi can be mitigated by borrowing contexts from the Marathi language. Next, an analysis of cross-lingual similarity between the languages, based on the cosine distance measure between the individual state-dependent parameters is presented in Section 6.5. Finally, in Section 6.6, a method to borrow context-dependent states based on the cosine distance measure is discussed. The effect of appropriately weighting these Marathi language states and their impact on the Hindi language recognition performance is studied. We conclude this paper by summarizing our findings in Section 7.

2. Agricultural commodities task domain

This section provides a brief description of the data used in this experimental study. As mentioned in Section 1, we use a subset of the data which has been collected for the project titled “Speech-based Access for Agricultural Commodity Prices in Six Indian Languages” sponsored by the Government of India. The goal of this project is to implement and deploy a speech based system for farmers to access prices of agricultural commodities in various districts across India from inexpensive mobile phones. The dialog systems are being collectively developed by a consortium of Indian institutions. In this paper the consortium of Indian institutions responsible for spoken dialogue system development is referred to as the Indian ASR Consortium.

The data collection process involved consultants actually travelling to the regions of interest and soliciting the participation of the local residents to call a central location from their mobile phones and respond by voice to prompts requesting the names of agricultural commodities and local districts. One of the main issues associated with this speech data in each language, is the fact that it has been collected in environments that shall be experienced in the actual use of the service. These include quiet indoor home environments, noisy indoor home environments with competing speech and broadcast audio as well as outdoor environments with background vehicle and machine noise. In short, the environments encountered in the data collection effort for each individual language are varied and are difficult to specifically characterize.

Table 1 shows the subset of the above-mentioned data set used in this experimental study, along with the number of hours of data available in each language for both training and testing. The data is narrow band 8 KHz speech data collected over mobile telephones. In Table 1, Hindi has just one hour of data while the Marathi has about eight hours of data. The speech data in each language largely consists of names of agricultural commodities and local districts. Table 1 shows the amount of available speech data for both Hindi and Marathi. For Marathi, the collected data was divided into training and test sets by the organization (IIT-Bombay, Mumbai, India) involved in the development of the agricultural commodities spoken dialogue system. For Hindi, the available data was equally divided by us into training and test sets. Table 2 also gives a distribution of the number of speakers in the final training and the test sets for each of the languages. In each language, there was not enough data for an independent development set. The data collected by the Indian ASR Consortium only had an evaluation set to gauge the initial performance of the spoken dialogue systems. We used this evaluation set as our test set. Therefore a few (CDHMM), though not all (CDHMM and SGMM) parameters in this study are tuned to optimize ASR performance on the given evaluation sets. Discussions appear in Sections 4.1, 4.2 and 5 that mention the details on how the choices of the model and test parameters were arrived at for the various systems. Clarifications have been provided where necessary.

This is a small vocabulary task. Table 3 shows the composition of the words and phones for this task. Column 2
of Table 3 lists the number of words in the lexicon for both languages. The lexicons in each language were prepared by each individual institution in-charge, to capture as many possible regional and dialectal variations of the pronunciation of a certain word as was deemed necessary. The Indian organizations involved in the development of the spoken dialogue systems for Hindi and Marathi did not consider it necessary to include pronunciation variants while developing the lexicons. The mapping of the consonants and vowels from native Indian scripts to their ASCII equivalent is achieved by using the ITRANS transliteration scheme (Chopde, 2006). Hindi, Marathi and Sanskrit use the Devanagari alphabet (Scharf and Hyman, 2009; Chopde, 2006; Central Hindi Directorate, 1977).

Most of the words that appear in the vocabulary correspond to names of commodities in each of the languages, while the rest are district names, “yes” and “no”, and spoken digits. To allow for unpredictable background events, special symbols appear in the transcriptions to mark events such as a sudden noise or a telephone ringing. The count of unique words in Table 3 does not include these symbols. These symbols map to a unique phonetic unit in the lexicon, and dedicated acoustic models are associated with them. The count of the phonetic units associated with these symbols are listed as “filler phones” in the last column of Table 3. The count of all of these filler phones include the mandatory sil phone.

3. The subspace Gaussian mixture model

This section provides a brief description of the subspace Gaussian mixture model (SGMM) implementation (Rose et al., 2011), proposed by Povey et al. (2011). The description here follows the work of Rose et al. (2011).

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

$$p(x|j) = \sum_{i=1}^{J} w_{ji} N(x|\mu_{ji}, \Sigma_{ji}),$$

(1)

where $I$ full-covariance Gaussians are shared between the $J$ states. The state dependent mean vector, $\mu_{ji}$, for state $j$ is a projection into the $i$th subspace defined by a linear subspace projection matrix $M_i$,

$$\mu_{ji} = m_i + M_i v_j.$$  (2)

In (2), $v_j$ is the state projection vector for state $j$. The subspace projection matrix $M_i$ is of dimension $D \times S$ where $S$ is the dimension of the state projection vector $v_j$ for state $j$. In this work, $S = D$. This choice of the subspace dimension was motivated primarily due to the ease of system configuration with this setting in our SGMM implementation (Rose et al., 2011). This parameter setting was consistent across our SGMM systems across our systems in Hindi, Marathi, Bengali and Assamese (Mohan et al., 2012b). The vector $m_i$, with dimension $D$, is used to denote the UBM mean of the Gaussian mixture component $i \in \{1, \ldots, I\}$. It is used as an optional offset to the term $M_i v_j$ in the expression for $\mu_{ji}$ in (2). The state specific weights in (1), are obtained from the state projection vector $v_j$ using a log-linear model,

$$w_{ji} = \frac{\exp w_{ji}}{\sum_{j'=1}^{J} \exp w_{j'i}}.$$  (3)

It is apparent that this acoustic modelling formalism has a large number of “shared” parameters and small number of state specific parameters. For multi-lingual acoustic modeling the shared parameters, namely $M, w$, and $\Sigma$, are trained by pooling data from multiple languages. Multi-lingual SGMM training involves maintaining separate phone sets for each language. This is made possible by adding a language specific tag to each phone, and thereby each clustered HMM state. Each state projection vector $v_j$, which is attached to a clustered state is then trained only with data specific to each language. The parameterization for multi-lingual SGMM training is described in Fig. 1.

In addition, to add more flexibility to the SGMM parametrization at the state level, the concept of substrates is adopted where the distribution of a state can be represented by more than one vector $v_{jiw}$, where $m$ is the substate


4. Experimental setup for the mono-lingual systems

This section gives a summary of the experimental setup for the mono-lingual systems. Section 4.1 describes the setup of the mono-lingual CDHMM system and Section 4.2 describes the setup of the mono-lingual SGMM systems.

4.1. Experimental setup for the mono-lingual CDHMM systems

This section gives an overview of the experimental setup and a description of the baseline CDHMM systems that we set up for this task. Baseline systems in each language were configured using the HTK Speech Recognition Toolkit (Young, 2006). Our CDHMM HTK ASR system setup in each language closely follows the setup of each regional institution of the Indian ASR Consortium. Each site configured their systems using the SPHINX (Lee, 1989) speech recognition system.

The baseline system was 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. Due to the lack of a development set, we used a fixed decision tree splitting threshold for the context-clustering procedure across the CDHMM systems that were built for Hindi, Marathi, Bengali and Assamese (Mohan et al., 2012b). The number of clustered states thus obtained in each language specific CDHMM system is listed in Table 5. We used 16 Gaussians per state for Marathi and 8 Gaussians per state for Hindi. The choice of the number of Gaussians per state was obtained by observing the performance of the CDHMM systems on the test set. The CDHMM systems for both languages used diagonal covariance Gaussians. The features used are 13 MFCC coefficients concatenated with first and second difference cepstrum coefficients. The features were extracted with an analysis window of size 25ms, with a hop size of 10ms used between successive frames of speech. Cepstrum mean normalization was performed for each utterance. We did not experiment with a variable number of Gaussians in each state.

One issue with setting up ASR systems in new languages is the definition of subword phonetic contexts. When these contexts are defined using decision tree clustering, it is necessary to define a list of phonetic contexts that make up the question set for the decision tree. Question sets prepared by expert phoneticians for decision tree clustering in under-resourced languages like the ones mentioned in this experimental study are difficult to obtain. Another approach, popular in grapheme-based speech recognition (Killer et al., 2003) is the use of the so-called singleton question sets. The questions used in the decision tree clustering procedure directly ask for the identity of the grapheme that might occur in the left or the right context, and not its linguistic class. Since Hindi and Marathi like most Indian languages are phonetic in nature with nearly a one-to-one mapping between the grapheme and the phoneme, it begets one to consider the use of grapheme based approaches to speech recognition. We did not consider this approach while building our system. Instead, we relied on an automated procedure that clusters CDHMM state distributions of context-independent phones is used to obtain a set of

![Diagram](image-url)

Fig. 1. Parameterization of the Multi-lingual SGMM.

Table 4

<table>
<thead>
<tr>
<th>Language</th>
<th>No. of words in training text for LM training</th>
<th>Perplexity</th>
<th>No. of 3-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marathi</td>
<td>28445</td>
<td>199.6</td>
<td>1624</td>
</tr>
<tr>
<td>Hindi</td>
<td>3807</td>
<td>49.8</td>
<td>361</td>
</tr>
</tbody>
</table>

Table 5

<table>
<thead>
<tr>
<th>Language</th>
<th>Clustered states</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marathi</td>
<td>476</td>
</tr>
<tr>
<td>Hindi</td>
<td>161</td>
</tr>
</tbody>
</table>
linguistic questions for clustering triphone contexts (Shrishrimal et al., 2012). The final linguistic question set that was obtained for each language was made compatible for use with HTK. We primarily chose the latter method to design our question set, to have our baseline HTK systems give an equivalent performance to the SPHINX based systems being configured by the Indian ASR Consortium. A comparison on the difference in ASR performance using singleton question sets though certainly interesting is not the focus of this experimental study and we leave it to address this issue in our future work.

Development of language models for spoken dialogue systems in a new task domain is difficult and language models are often bootstrapped (Weilhammer et al., 2011; Hakkani-Tur and Rahim, 2006). For this experimental study language models were built with the best available resources at hand. The language models (LM) used in the case of Marathi and Hindi were tri-gram language models. Training text consisted of transcriptions of the utterances used in acoustic model training. This is a closed vocabulary task. The training and test utterances consist of single or multiple words with sentence start and end markers describing the name of commodities or districts. The training text used for obtaining the N-gram LM was able to cover all of the words encountered in the vocabulary. The number of words used to train the language models along with the perplexity on the test set are included in Table 4. The evaluation we provide for our Hindi system is on recognition of the commodities since these span the majority of the words in the vocabulary. For Hindi, a separate language model is trained for recognizing either commodities or districts, as determined by the organization developing the Hindi spoken dialog system. For Marathi on the other hand a single language model is trained for recognizing both commodities and districts.

4.2. Experimental setup for the mono-lingual SGMM systems

For the SGMM, we use our implementation that is an extension to HTK with added libraries (Rose et al., 2011). The system is initialized with the clustered triphone states obtained from the CDHMM. The basic setup is identical to that used for the CDHMM systems.

The training for the SGMM in this work follows the procedure mentioned in Rose et al. (2011). Each language-specific SGMM system is initialized with Gaussians from a language-specific universal background model (UBM). Each language-specific UBM is trained from speech-only segments from the training portion of the data-set of each language. The nature of the speech data in this task which is available in limited quantities, forced us to experiment with empirical procedures to obtain the initial set of \( I_l \) Gaussians for each language. Here we use \( I_l \) to denote the total number of Gaussians in each language specific system \( l \in \{\text{Hindi, Marathi}\} \). For example, reference to Table 9, \( I_{\text{Hindi}} = 256 \). The number of \( I_l \) Gaussians in each language specific UBM were determined by looking at the distribution of the mixture dependent frame occupancy counts. We chose the final number of \( I_l \) Gaussians ensuring that there was an even distribution of speech frames across all mixtures. We trained each language specific UBM, with all \( I_l \) Gaussians in that UBM sharing a single grand full-covariance matrix. This allows for a robust estimation of the single full-covariance matrix, instead of estimating \( I_l \) individual full-covariance UBM matrices as is done traditionally (Povey et al., 2011; Rose et al., 2011). After obtaining the UBM Gaussians, each of the \( I_l \) individual SGMM Gaussians is initialized with the same full covariance matrix that was shared between the \( I_l \) UBM Gaussians. Training of the SGMM is then carried out using a joint posterior initialization (JPI) (Rose et al., 2011). A joint posterior initialization involves a process where the initial posterior probability of a speech frame \( x_t \) at time \( t \) belonging to a mixture component \( i \) in state \( j \) is approximated as:

\[
u_j^i(t) = p^i(s_t = j, m_t = i | x_t)
\]

\[
\approx p_h(s_t = j | x_t)p_h(m_t = i | x_t).
\]

In Eq. 8, the posterior \( p_h(s_t = j | x_t) \) are obtained from forward-backward decoding of the training utterances with respect to the prototype CDHMM. The posteriors \( p_h(m_t = i | x_t) \) are obtained from training utterances using the initial Gaussian mixture models (i.e. the language specific UBM in this case). For this task where the vocabulary is very small a JPI of the SGMM is extremely critical for the final system to give good performance. Updates for the full-covariance matrices were initially disabled for the first 8 iterations where initial estimates for the other parameters of the SGMM are obtained through the JPI process. For the 10 subsequent training iterations, parameter updates were allowed for all SGMM parameters. The results for the SGMM system performance in each language, with the number of mixtures \( I \) (\( I_l \) in this section) used in each system is presented in Table 9.

The following Tables 6 and 7 give the counts of the number of parameters in the CDHMM and the SGMM systems for the Hindi language. In a small-vocabulary ASR system, the parameter count for the SGMM system surpasses the number of parameters in the equivalent CDHMM system. In this case the number of shared parameters in the SGMM dominates the total parameter count. This is true for the Hindi mono-lingual SGMM system described here. Referring to Table 6 the Hindi SGMM system configured with \( I = 256 \) full-covariance Gaussians and \( J = 161 \), tied-states, has a total of 605319 parameters. A feature dimension of \( D = 39 \) features was used with a subspace dimension \( S = 39 \). The counterpart CDHMM system, configured with \( J = 161 \) tied-states with 8 Gaussian mixtures per-state on the other hand has a total of 101752 parameters. On the other hand, in large-vocabulary continuous speech recognition systems (LVCSR) systems due to the presence of a larger number of states \( J \), the
CDHMM systems tend to have far more parameters compared to equivalent SGMM systems (Povey et al., 2011). The only way to increase the number of parameters in a CDHMM system is by increasing the number of state-dependent Gaussian densities. From our experiments when we tried increasing the number of state-dependent mixtures in the CDHMM beyond 8 mixtures per state the resulting ASR performance was seen to degrade significantly. This is to be expected as the amount of data per state-dependent Gaussian mixture in the CDHMM decreases as the number of state-dependent mixtures are increased. On the other hand, especially in a single sub-state per state SGMM system the state-dependent parameter is only a $D$ dimensional vector. With limited amounts of data therefore one could reliably estimate state-dependent parameters, with all the data in the corpus being used to train the shared parameters. The SGMM’s parametrization with its larger number of shared parameters (shared between the states, as depicted in Fig. 1) certainly allows for a more expressive model with the same amount of data that is used to train a CDHMM system.

### 5. Comparative performance of the mono-lingual CDHMM and the mono-lingual SGMM systems

This section summarizes the comparative performance between the baseline mono-lingual CDHMM systems and the mono-lingual SGMM systems. We use the Word Accuracy (WAc.) as the performance measure for recognition. We have also provided the percentage correct scores (%Corr.) for comparison. The percentage correct scores (%Corr.) perhaps carry more meaning in the context of this task, since for the spoken dialogue system it is important that the transcription of the words in the utterance be retrieved correctly by the ASR engine. The ASR result could then be post-processed to make it suitable for use with other components of the spoken dialogue system.

### 6. Hindi and Marathi multi-lingual SGMM system

This section summarizes the building of a multi-lingual SGMM system for the closely related language pair – Hindi and Marathi. Both Hindi and Marathi belong to the Indoeuropean (Cardona, 2003) group of languages. Both languages share a large degree of similarity in terms of syntactic structure, written Devanagari script (Central Hindi Directorate, 1977), and to some extent the vocabulary. The Devanagari script is phonetic in nature. There is more or less a one-to-one correspondence between what is being written and what is spoken. As a result, to a large extent both languages are in fact “phonetically similar”. Our interest here is to understand if this similarity between languages could be exploited in order to build a multi-lingual speech recognition system, whose performance is perhaps better than their mono-lingual counterparts. In addition a further question we seek to answer is, having built such a system, is the “phonetic similarity” between Hindi and Marathi apparent in terms of their acoustic realizations captured by statistical models?

### Table 6

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Parameter count</th>
<th>Example count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_i$</td>
<td>$IDS$</td>
<td>389376</td>
</tr>
<tr>
<td>$S_r$</td>
<td>$D(D+1)/2$</td>
<td>199680</td>
</tr>
<tr>
<td>$w_j$</td>
<td>$JS$</td>
<td>9984</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>605319</td>
</tr>
</tbody>
</table>

### Table 7

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Parameter count</th>
<th>Example count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{jr}$</td>
<td>$DJ \times #mixtures$</td>
<td>100464</td>
</tr>
<tr>
<td>$\Sigma_{jr}$</td>
<td>$D \times #mixtures$</td>
<td>100464</td>
</tr>
<tr>
<td>$w_{jr}$</td>
<td>$J \times #mixtures$</td>
<td>2576</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>101752</td>
</tr>
</tbody>
</table>

With reference to Table 8 the Marathi CDHMM system with a vocabulary of 551 words has a baseline performance of 74.2%. The performance for Hindi at 63.8% is perhaps due to a lack of data. The insertion penalty and language scale factors for the CDHMM decoder were tuned on the test set to give the best possible performance. The results for the performance of the mono-lingual SGMM systems is listed in Table 9. For the Marathi SGMM systems, a 3.5% absolute increase in the WAc. is observed with respect to the baseline system. For Hindi an increase of 3.4% WAc. is seen in the performance of the SGMM with respect to the baseline. The insertion penalty and language scale factors for the SGMM were not tuned on the test set. The same values for these parameters were maintained for the recognition experiments for all four languages Hindi, Marathi, Assamese and Bengali (Mohan et al., 2012b). From these results we could conclude that the SGMM is seen to provide better performance possibly due to its ability to model phonetic variability effectively.

### Table 8

<table>
<thead>
<tr>
<th>Language</th>
<th>%WAc (%Corr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marathi</td>
<td>74.2 (82.1)</td>
</tr>
<tr>
<td>Hindi</td>
<td>65.8 (68.7)</td>
</tr>
</tbody>
</table>

### Table 9

<table>
<thead>
<tr>
<th>Language</th>
<th>No. of Mixtures</th>
<th>%WAc (%Corr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marathi</td>
<td>400</td>
<td>77.7 (84.7)</td>
</tr>
<tr>
<td>Hindi</td>
<td>256</td>
<td>69.2 (71.0)</td>
</tr>
</tbody>
</table>
Multi-lingual SGMM systems in Burget et al. (2010) and Povey et al. (2011) have been built without explicitly accounting for possible mismatch of acoustic, background and environmental conditions between data-sets from multiple languages. From our previous work (Mohan et al., 2012a), in the absence of cross-corpus normalization, it has been found that these sources of mismatch between data sets is a significant issue in most practical scenarios for training systems from multiple data-sets. Without accounting for this mismatch the performance of the multi-lingual system is not any better than its mono-lingual counterpart. After accounting for cross-corpus mismatch the multi-lingual SGMM system is seen to perform better compared to the mono-lingual SGMM. The method for dealing with cross-corpus mismatch is a variation of Speaker Adaptive Training (SAT) and is described in greater detail in Section 6.2.

6.1. Multi-lingual SGMM training

First, SGMM training was carried out in a multi-lingual fashion, by using the Marathi and Hindi data to collectively train the “shared” parameters $M_i, w_i$ and $\Sigma_i$ as was first done in Burget et al. (2010). Here $i = \{1, \ldots, I\}$ is the index over the Gaussian mixture component. As mentioned in Section 3 separate phone sets are maintained for each language. Maintaining separate phone sets allows for training the state specific parameters i.e. $v_j, j = \{1, \ldots, J\}$ with data from each of the two languages individually. Our multi-lingual SGMM has a total of $J = 637$ clustered states, with $161$ states coming from Hindi and $476$ coming form the Marathi CDHMM systems respectively. The system was initialized with a UBM with $I = 400$ Gaussians trained on speech-only segments of speakers from both corpora. The system was initialized with a joint posterior initialization procedure, similar to the monolingual SGMM systems trained for this task. No feature normalization was applied prior to multi-lingual training to account for cross-corpus mismatch. The results for the multi-lingual system trained in this manner is reported in Table 10.

<table>
<thead>
<tr>
<th>Language</th>
<th>%WAc ( %Corr )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marathi</td>
<td>78.7 (85.2)</td>
</tr>
<tr>
<td>Hindi</td>
<td>50.5 (60.1)</td>
</tr>
</tbody>
</table>

It can clearly be seen from the results of Table 10 that the multi-lingual SGMM system for Hindi performs no better than the mono-lingual Hindi SGMM system. This is consistent with our earlier observations stated in Mohan et al. (2012a). Building a multi-lingual system without accounting for the systematic acoustic differences between sources of multi-lingual data could decrease rather than help performance. Since the data used in this experimental study is “real-world”, where the acoustic environments are difficult to characterize, the impact of this mismatch appears to be severe. In comparing this multi-lingual training result for the Hindi language test set, a decrease in WAc. of 8% absolute is seen with respect to the CDHMM baseline. Further, with respect to the mono-lingual Hindi SGMM system a decrease in WAc. of 19% is seen. For the Marathi system on the other hand, the performance is about the same as the Marathi mono-lingual SGMM system, with an improvement of about 1% that can be seen for the multi-lingual training.

6.2. Cross-corpus feature normalization

This section presents an acoustic normalization procedure for dealing with this cross-corpus mismatch (Mohan et al., 2012a). The procedure is applied as part of SGMM training and corresponds to a straight-forward variant of speaker adaptive training (SAT) (Gales, 2001b) and constrained maximum likelihood linear regression (CMLLR) (Gales, 1998). Referring to Fig. 2 the acoustic normalization procedure is described here as being performed in two major steps as given in Sections 6.2.1 and 6.2.2 below.

6.2.1. Compensating for speaker variation

- Given our training feature sets $X = \{X_H, X_M\}$ and the speaker independent (SI) CDHMM models $\Lambda = \{\Lambda_H, \Lambda_M\}$, we first estimate speaker dependent CMLLR matrices $\Lambda_H$ and $\Lambda_M$, as is done in standard SAT for both languages Hindi (H) and Marathi (M). Here we use the index $r$ to denote a speaker in the Hindi training set and $s$ to denote a speaker in the Marathi training set.
- Interpreting the CMLLR transformation as a feature space transformation (Gales, 1998), we obtain a transformed feature set $X = \{A_H X_H, A_M X_M\}$ by transforming our original feature set $X$ using the speaker dependent transformations $A_H$ and $A_M$.
- Speaker-normalized mono-lingual CDHMM models $\Lambda = \{\Lambda_H, \Lambda_M\}$ are trained using our new feature set $X$.
- Using the mono-lingual speaker-normalized models $\Lambda = \{\Lambda_H, \Lambda_M\}$ and the test features $Y = \{Y_H, Y_M\}$, language-specific speaker dependent CMLLR transforms are generated in an unsupervised fashion for each test speaker. Here $r$ is used to denote a speaker in the Hindi test set, and $s$ is used to denote a speaker in the Marathi test set.
- $Y = \{Y_r, Y_s\}$ are then transformed with the corresponding test speaker CMLLR transforms to obtain a new set of test features $Y$. Recognition results for this SAT procedure for both Marathi and Hindi are listed in row 2 (CDHMM+SAT) of Tables 11 and 12, respectively.
Since this first step normalizes only inter-speaker variation and not inter-corpus acoustic mismatch, we perform a modified version of the SAT procedure in a second step.

### 6.2.2. Compensating for inter-corpus acoustic mismatch

- Using the models $\hat{A}$ and the feature sets $\hat{X}$, two language specific CMLLR transformation matrices $B^H$ and $B^M$ are trained. The transforms are now generated as a modified version of SAT, where we treat the data from each language as if it were data coming from two “speakers” – Hindi and Marathi. Two language-independent regression classes are used namely “speech” and “silence”. The “speech” regression class is assigned all the non-silence Gaussian components coming from CDHMM states of both languages. We believe that large amounts of data are required to characterize the acoustic environments, and therefore all the training data from each language is required to train the transformation matrices $B^H$ and $B^M$.

- We use $B^H$ and $B^M$ to obtain an updated set of training features $\hat{X}$ by transforming the features $\hat{X}$. These features are then used to obtain an updated CDHMM model $A$.

- A new set of test features $\hat{Y}$ is obtained by transforming $\hat{Y}$ using the language specific CMLLR transforms $B^H$ and $B^M$.

- The new training features $\hat{X}$ are now used for multi-lingual SGMM training. This involves training the UBM, generating the JPI (joint posterior initialization, ref. Section 4.2) estimates, and finally the SGMM training. The updated CDHMM model $A$ is used for generating the new JPI estimates before multi-lingual SGMM training.

Clearly there are many ways of compensating for variability across speakers and across corpora. The advantage of transforming features from both data sets as described in this section is that it serves to reduce mismatch across all stages of SGMM training.

### 6.3. Impact of speaker variability and the cross-corpus feature normalization procedure

In this section we provide some experimental results to assess the effectiveness of the cross-corpus normalization procedure presented in Section 6.2, and its impact on multi-lingual training. The results for Marathi language ASR performance is summarized in Table 11 and Hindi language ASR performance is summarized in Table 12.
In this corpus, speaker variability seems to be an issue affecting the speech recognition system performance. Referring to the second row of Tables 11 and 12 it can be seen that a large gain in performance with respect to the CDHMM baseline for both Marathi and Hindi is obtained by using speaker adaptive training with the CDHMM models.

Row 3 of Tables 11 and 12 lists the mono-lingual SGMM performance in both Marathi and Hindi. Row 4 of Tables 11 and 12 lists results of multi-lingual training without any cross-speaker or cross-corpus normalization. This result, as mentioned before in Section 6.1 highlights the issues of training a multi-lingual system without accounting for cross-corpus mismatch. The label AN in the first column of Tables 11 and 12 is used to denote the use of the acoustic normalization procedure described in Section 6.2. In Row 5 of both tables, results are shown for language-specific SGMM models which have been trained with the normalized features. In other words the language-specific systems (the results for which are displayed in Row 5 of both tables), are trained after speaker normalization and cross-corpus acoustic normalization. A drastic improvement is seen with respect to the mono-lingual SGMM trained without the normalization procedure (row 3 of Tables 11 and 12).

These results thus show the impact of this normalization procedure on the final ASR system performance in both languages. As noted earlier in Section 2 a great deal of environmental variability exists in the speech data within a language. Therefore this improvement could be attributed to the fact that the CMLLR transform $B^f$ trained from all the features in Hindi training set is able to compensate for the intra-corpus environmental variability to yield the improvement that is seen by the mono-lingual+AN Hindi system. Further, Row 6 of Tables 11 and 12 show the performance of the multi-lingual SGMM models trained with the normalized acoustic features. With respect to row 5 of the Table 12, a 2% absolute improvement in the Hindi ASR system can be observed. This improvement can therefore be attributed to multi-lingual training. We applied the matched-pairs significance test described in Gillick and Cox (1989), and the improvement in performance for the multi-lingual system (Row 6 of Table 12) versus the Hindi SGMM (mono-lingual+AN) (Row 5 of Table 12) was statistically significant at the chosen confidence level of 99.5%. Section 6.4 analyzes the recognition errors made by the multi-lingual Hindi system.

### 6.4. Analysis of recognition errors made by the multi-lingual Hindi system

While the SGMM is able to provide additional gains by allowing a structured and shared parameterization between multiple languages, useful insights for SGMM-ASR system construction can be obtained by analyzing the errors made on the Hindi utterances by the multi-lingual SGMM system. It was found that the recognizer consistently made errors of the kind that are listed below:

- **TARBOOJ** = *tarbooj* meaning watermelon in Hindi is mis-recognized as *musk-melon*.
- **KARBUJA** = *karbujaa*.
- **BADSHAH** = *b a a d s h a a* a variety of potato is mis-recognized as *BAJRA = b a a j r a a* or millet.
- **ANANNAS** = *a n a a n a s* meaning pineapple is mis-recognized as *pomegranate* or **ANAR** = *a n a a r*.
- **SUHUMIRCH** appearing as two separate words with the composite lexical expansion *su kh i m i r c* meaning dried red-chilli peppers is mis-recognized 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 favored 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 $a-nm+a$ in the third example. Further, the number of context-dependent states in Hindi, $J = 161$, is only slightly higher than the number of $K = 147$ context-independent states for this task.

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.

The issue of sharing state-level SGMM parameters for multi-lingual ASR has received limited attention in the literature, except for the work presented in Qian et al. (2011). An algorithm is presented by the authors in that work to “borrow” data from segments of speech data of the well-resourced language that are acoustically “similar” to that in the target language. The acoustic similarity was determined by using an appropriate distance measure between the state-dependent parameters of the low-resource target language SGMM and the non-target language SGMM. This “borrowed” data was then used to update the state-dependent parameters of states that are then shared between the target and non-target languages. For the low-resource acoustic model training condition that was studied, a limited gain in ASR performance over the base-line SGMM were reported by using this strategy.

Here an anecdotal “cheating” experiment is presented where the states of the context-dependent HMM of the Hindi language directly use state-dependent parameters of “similar” context-dependent HMMs 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 should be combined 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.

All the SGMM based systems described in Sections 6.1, 6.2, 6.3 rely on single sub-state per state models. 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. A technique for sharing Gaussians at the state-level between context-dependent models was successfully applied in Saraclar and Khudanpur (2000) to model pronunciation variation in conversational speech. 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 confusables 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 issue of estimation of these weights is not considered here and is a topic of future research. The state dependent likelihoods for these “augmented” states are computed as given by Eq. 4. The multi-lingual SGMM model thus obtained is referred to as the “augmented” model in Table 13. A similar procedure was carried out to modify the target language silence model.

N-Best lists were generated for all the utterances corresponding to the four mis-recognized examples with a value of $N = 2$. This generated the possible hypotheses as being either the correct term or the closest incorrect term. The first row of Table 13 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 13.

It can be seen from Table 13 that the so-called “augmented” models provide a 20% absolute improvement over the multi-lingual SGMM Hindi models in correctly recognizing the originally mis-recognized 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 multi-lingual SGMMs.

Table 13
Recognition accuracy for frequently mis-recognized utterances.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML-Hindi SGMM</td>
<td>56%</td>
</tr>
<tr>
<td>“Augmented” ML-Hindi SGMM</td>
<td>76%</td>
</tr>
</tbody>
</table>

6.5. Analysis of cross-lingual phonetic similarity captured by the SGMM model parameters

As mentioned at the beginning of this section, Marathi and Hindi are languages that are “phonetically similar”. This is because both languages share the Devanagari script (Central Hindi Directorate, 1977), where there is almost a one-to-one correspondence between what is being written and spoken. This section demonstrates that this cross-lingual similarity is in fact reflected in the state-dependent parameters of the multi-lingual SGMM.

Fig. 3 shows a heatmap to visualize the similarities “discovered” between the phones of Hindi and the phones of Marathi. The plot was obtained as follows. We first extracted the normalized (refer Appendix K of Povey (2009)) state projection vectors $v$ from the centre state of clustered cross-word tri-phone models of both languages Hindi and Marathi. Then, a similarity $s(h,m)$ was calculated as the cosine distance between state $h$ from the Hindi language model and state $m$ from the Marathi language.

In Fig. 3, the vertical axis bears the labels of the phones that correspond to the clustered states of the Hindi language, and the horizontal axis bears the labels of the phones that correspond to the clustered states of the Marathi language. It is necessary to note that in this task domain, Hindi and Marathi do not have the identical phone sets, even though they are derived from the same script. In Fig. 3, for the purposes of viewing not all of the center context labels are displayed along the axes of the plot. The brightness of a point on this heatmap is proportional to the “closeness” between the two corresponding states. The labels are ordered alphabetically by the ITRANS ASCII equivalent of the Devanagari alphabet. As the ASCII equivalents (of the phones) in both languages (i.e. Hindi and Marathi) correspond to the same Devanagari alphabet some of similarities are immediately apparent from the plot. For example, aa (ARPABET AA) in Marathi is seen to be similar to a (ARPABET A) in Hindi. Other common confusables pairs are also apparent:
o (ARPABET AO) is seen to be confused with uu (ARPABET UW) in both Marathi and Hindi and s in Hindi (ARPABET S) is seen to be confused with cc (ARPABET CH) in Marathi. The heatmap in Fig. 3 serves to show that the state projection vectors vj are able to capture information that is phonetically and linguistically meaningful.

6.6. Cross-lingual context-dependent state borrowing

This section describes an experimental to study the effect of borrowing context dependent states from the Marathi language to improve the Hindi ASR performance. This in effect includes two steps. The first step is to automatically select the appropriate Marathi language context-dependent state for each Hindi language state. The second step is to weight these context-dependent states appropriately, as sub-states, to yield the final state-likelihood. The first issue is the subject of Section 6.6.1. After having observed cross-lingual similarities using a simple cosine distance metric in Section 6.5, we use this distance to select context-dependent states from the Marathi language. In order to understand the effect of weighting the sub-states properly, the sub-state weights are varied globally over a range of values and the effect on the percentage of Hindi words correctly decoded is observed in Section 6.6.2.

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

As a first step, for each target language (Hindi) state, a list of 10 similar states is picked based on the cosine distance metric. The evaluation of the similarity is restricted only to those non-target language (Marathi) 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 involved 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 was 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.

6.6.2. Experimental analysis of language-weighting

Fig. 4 summarizes the Hindi language ASR performance with cross-lingual state sharing. The performance curve is displayed as the static target language sub-state weight is varied between 0 and 1. 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% is well below the baseline multi-lingual SGMM performance on the Hindi test set. Also as expected, as the weighting towards the Hindi language states increase, the performance is seen to increase. The best performing system at 77.77% is obtained when the Hindi language states are weighted at 0.8. At this point, an improvement of 1.57% absolute is seen with respect to the SGMM baseline of 76.2%. The matched-pairs significance test described in Gillick and Cox (1989) was run, and the improvement in performance with respect
to the SGMM baseline was statistically significant at the chosen confidence level of 99.99%.

7. Conclusions

An experimental study was presented that investigated acoustic modelling configurations for speech recognition in the Indian languages – Hindi and Marathi. The experimental study was performed using data from a small vocabulary agricultural commodities task domain that was collected for configuring spoken dialogue systems. Two acoustic modelling techniques for mono-lingual ASR were compared namely – the conventional CDHMM and the SGMM acoustic modelling technique. The SGMM mono-lingual models were seen to outperform their CDHMM counterparts when there is insufficient acoustic training data. For the Hindi and Marathi language pair, a multi-lingual SGMM training scenario was presented. It was concluded that cross-corpus mismatch is an important issue that needs to be addressed while building systems of this nature. Not accounting for cross-corpus mismatch is seen to decrease the performance in the target language Hindi which has limited amounts of training data. After accounting for this cross-corpus mismatch a gain in multi-lingual SGMM performance is observed. Further, interesting anecdotal results have been obtained to show that the parameters in the multi-lingual SGMM system are able to capture the phonetic characteristics in a structured and meaningful manner especially across the the “similar” languages used in this experimental study. Finally, it was found beneficial to share “similar” context-dependent states from Marathi in order to improve Hindi language speech recognition performance.

Acknowledgements

The authors would like to thank all of the members involved in the data collection effort and the development of the dialogue system for the project “Speech-based Access for Agricultural Commodity Prices in Six Indian Languages” sponsored by the Government of India. We would also like to thank M.S. Research Scholar Raghavendra Bilgi at IIT Madras for his timely help with providing resources for this experimental study.

References


Povey, D., 2009. A Tutorial-Style Introduction to Subspace Gaussian Mixture Models for Speech Recognition. Microsoft Research Redmond, WA.


