4

Analysis and quantification of accents

4.1 Introduction

Accent is a pattern of pronunciation used by the community of native/non-native speakers belonging to some cultural/geographical regions. The difference between accents is large for a language spoken in a wider geographical area. Hence, the effect of accent on recognition of speech is felt in many languages. The two principal components of variations in speakers correspond to the gender and accent. Accent is a major attribute that impacts automatic speech recognition systems. The impact of accent on the performance accuracy of speech recognition systems is higher than that of the gender. The models for ASR have been built to manage gender variability. However, the research on accented speech is relatively less, especially for native speakers with accents caused by the dialects of the regions. A systematic study of accents, quantitative as well as qualitative could be useful in speech recognition systems. To alleviate the problem of impact of accent on accuracy of ASR systems, the language as well as acoustic models are generally adapted to target accents. However, the amount of adaptation data available is less in general. Hence, it is desirable to have a systematic approach
to select a small amount of adaptation data that will adapt speech models well.

If we can quantify the degree of difference between two accents, this quantity could be used to estimate the amount of data needed to adequately adapt models to target accent. Further, if there is an approach for identifying phonemes whose pronunciations in two accent regions differ significantly, we may be able to select the adaptation data to include more of such phonemes. The adaptation data should be chosen such that the speech models trained with data of one accent region comes closer to the target region. Thus a scientific way of quantifying a pair of accents is desirable.

A qualitative analysis of Kannada vowels in different Kannada speaking regions was done to study the accents of spoken Kannada. Even though Kannada is the official language of Karnataka state, the accent, lexicon and prosody varies across geographical areas. We can identify three major accent regions, namely Coastal, South and North based on the native language and accent as shown in figure 3.1. Each region has specific prosodic and lexical features; the former defines the speaking style, while, the latter defines more prevalent words and their pronunciations. The study was conducted in the context of HMM based recognition of Kannada language speech of three accents. In the later sections of the chapter, we present the proposed method of accent quantification and results of experiments that adopts cross entropy as a means of quantifying difference between two accents.

Qualitative analysis of Kannada accents is discussed in section 4.2. Section 4.3 describes the usefulness of cross entropy in detail. Proposed method of accent quantification is discussed in section 4.5. The experimental details are presented in 4.7. Analysis of the experimental results are given in section 4.8.

4.2 Qualitative analysis of Kannada accents

An acoustic-phonetic analysis of two Kannada accents (spoken in southern and coastal regions of Karnataka) is presented in this section. The study is carried
out by analyzing the distribution of vowel formant frequencies. There is a marked difference in Kannada spoken in southern, coastal and northern regions, which is easily distinguishable by Kannada speakers. The analysis of vowels in formant space provides a method of assessing the influence of accent on formants of phonemes. The study aims at finding whether there is any significant difference in the acoustic space of vowels across the regions. If there is any difference, we can quantify and model difference between the accents so as to benefit from such a model in the context of Kannada speech recognition. Also, a study of formant characteristics of phonemes across various regions contributes to better understanding of spoken language. This section presents the results of a pilot study that analyzes the formants of vowels spoken by speakers of residents of Coastal and South Karnataka.

4.2.1 Computation of vowel formant frequencies

In this study, mid points of the vowels were automatically identified using a trained speech recognition system, and formant frequencies corresponding to the mid point were computed. The mid points of phonemes are less affected by the preceding and succeeding phonemes. Hence, the formants corresponding to the mid points better characterize the phoneme. Formant frequencies were estimated by conventional frame based Linear Prediction analysis using snack sound toolkit[74]. Linear Prediction Coefficient order was set to 18 and the frame size was set to 30 ms empirically. The schematic representation of the formant analysis method is shown in figure 4.1. The formant frequencies were extracted according to the following steps:

1. Compute the formant frequencies of all speech data.

2. Train a speech recognition system (generate HMM models for phonemes using transcriptions of speech data).

3. Align a speech file with a network of phoneme HMMs corresponding to the transcription of the speech file.
Figure 4.1: Schematic diagram of automatic computation of formant frequencies at the mid point of phonemes. Viterbi alignment of utterance with trained HMM models provides phoneme boundaries, which helps in computing the mid point of a phoneme. Formants corresponding to the mid point of a phoneme are selected, since mid point of a phoneme is less influenced by preceding and succeeding phonemes.

4. Determine boundaries of vowels from the alignment.

5. Select formant frequencies corresponding to the mid points of the vowels.

4.2.2 Speech corpus for acoustic-phonetic analysis

A text corpus of 200 Kannada proverbs was created. Speech data was collected from two speakers from each accent region. Each speaker read 40 sentences from a list of 200 sentences. Spoken sentences were recorded with 16000Hz sampling rate and 16 bit sample size with a unidirectional microphone. Transcription (text corresponding to speech data), pronunciation dictionary and word list was created for the Kannada proverb sentences, which are used by HTK toolkit to train the speech recognition system.

4.2.3 Results and analysis

We studied the characteristics of the first two formants of ten vowels, viz., /a/, /A/, /i/, /I/, /u/, /U/, /e/, /E/, /o/ and /O/.
we have considered two accent regions: Coastal and South. Analysis of variation in acoustic properties of vowels as a function of dialect was done by visual comparison of vowel clusters in space of the first two formants. This is similar to the famous plot of $F_1-F_2$ by Peterson & Barney[56] for American English vowels and similar plot by Agarwal et al. for Hindi vowels[3].

Figure 4.2 is a scatter plot of the ten vowels by speakers of coastal and south regions in the space of two formant frequencies, F1 and F2. There appear to be quite a few stray data, most likely due to errors in Viterbi alignment. Another reason for stray data is variation in pronunciation[57]. People pronounce sounds differently depending on dialect, context which will produce outlier data points. To tackle the above defined problem, we have ignored the obvious outlier points while fitting the ellipses using Principal Component Analysis (PCA)[33].

We can clearly see the difference in the size and shape of the clusters of vowels spoken by speakers of two accented regions. The spread of south data is bigger than the coastal data for most of the phonemes. There are two major reasons for this spread: Firstly, lexical constructs of Kannada speakers from southern region vary with those of the coastal region. Secondly, the coastal region Kannada has fewer variations in pronunciations. The native language of speakers of majority residents of coastal Karnataka is Tulu. Hence, the variations in pronunciations are minimal in coastal region Kannada.

We analyze the range of formant frequencies ($F_1$ and $F_2$) for different vowels across two accent regions. The magnitude of major and minor axis of the ellipses derived from PCA indicate the spread of formants in the two principal directions. We can use the ratio of these magnitudes for the two ellipses corresponding to two accent regions as a measure of range of formant variation in acoustic space of vowels.

Let $V_{coastal}$ and $V_{south}$ denote magnitude of principal axis in two ellipses corresponding coastal and south respectively. The ratio $\frac{V_{south}}{V_{coastal}}$ gives a measure of spread. If the ratio is farther than 1, the difference in accents is more.

The value of variance ratios of formant frequencies of ten Kannada vowels are
Figure 4.2: Scatter plot of ten Kannada vowels /a/, /A/, /i/, /I/, /u/, /U/, /e/, /E/, /o/ and /O/ by speakers of Coastal and South Karnataka regions. The x-axis corresponds to the first formant frequency (\(F_1\)), while the y-axis corresponds to the second formant frequency (\(F_2\)). Top right corner of the subplot indicates the phoneme names corresponding to each subplot. The symbols “+”, “∗” indicate the presence of coastal and south data points in formant space respectively. Solid line ellipses show clusters fitted to Coastal data points, whereas dotted line ellipses indicate clusters around the South data in formant space.

The ratio \(V_{south}/V_{coastal}\) of phoneme /a/ in \(F_1\) and \(F_2\) are 1.3 and 1.1 respectively. This indicates that the spread of \(F_1\) of south region phoneme /a/ is greater than the coastal phoneme. Similarly, the spread of \(F_2\) of southern region phoneme /a/ is higher than that of coastal region. Figure 4.2 shows that the ellipses of phonemes /a/, /A/, /U/, /e/ and /O/ of southern region completely covers the coastal region ellipses. From these observations we can conclude that the southern region pronunciation for phonemes /a/, /A/, /U/, /e/ and /O/ varies more compared to the coastal region. However, such trends are not clear in other vowels. These observations are explained in following paragraph.

The south region phoneme /i/ has more spread across \(F_1\) and its variance
Table 4.1: Variance ratio $\frac{V_{\text{south}}}{V_{\text{coastal}}}$ of /a/, /A/, /i/, /I/, /u/, /U/, /e/, /E/, /o/, /O/ across $F_1$ and $F_2$. Here, $V_{\text{south}}$ and $V_{\text{coastal}}$ represent variances of formant frequencies around their means of South and Coastal Kannataka speakers.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>/a/</th>
<th>/A/</th>
<th>/i/</th>
<th>/I/</th>
<th>/u/</th>
<th>/U/</th>
<th>/e/</th>
<th>/E/</th>
<th>/o/</th>
<th>/O/</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>1.3</td>
<td>2.0</td>
<td>1.9</td>
<td>0.8</td>
<td>1.5</td>
<td>1.3</td>
<td>1.1</td>
<td>1.5</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>$F_2$</td>
<td>1.1</td>
<td>2.3</td>
<td>0.7</td>
<td>1.5</td>
<td>0.4</td>
<td>1.1</td>
<td>1.3</td>
<td>0.9</td>
<td>0.8</td>
<td>2.8</td>
</tr>
</tbody>
</table>

ratio is 1.9 whereas spread across $F_2$ of southern region is less than that of coastal region (variance ratio is 0.7). In case phoneme /I/, the spread of $F_2$ of south region is more than coastal region and its variance ratio is 1.2. However, the spread across $F_1$ of phoneme /I/ remains nearly the same for both the regions with little difference in the variance ratio (0.8). The phonemes /E/, /o/ has more spread across $F_1$ (variance ratio of /E/ is 1.1 and /o/ is 1.5) in southern region compared to coastal region whereas the spread across both the regions remains almost the same (variance ratio of /E/ is 0.9 and /o/ is 0.8). Finally, the spread of phoneme /u/ across both $F_1$ and $F_2$ is slightly lower than coastal region compared to southern region (variance ratio of $F_1$ is 0.6 and $F_2$ is 0.4).

From the above analysis, we note that the dimension of acoustic space of short vowel of coastal region phonemes are less than that of south region vowels across $F_1$ and $F_2$. From the figure 4.2, it can also be noted that the center of coastal and southern ellipses remains approximately the same despite the spread in $F_1$ and $F_2$. An inspection of the table 4.1 shows that, except for the short vowel /u/, the spreads of other southern phonemes across $F_1$ and $F_2$ are approximately equal to or greater than the coastal region phonemes. These results indicate that pronunciations of south region speakers vary more compared to the coastal region pronunciations. This phenomenon is expected, and can be explained as follows. Kannada is not the mother tongue of the two speakers from coastal region of Karnataka; Kannada is a second language for them. So, they are likely to pro-
nounce Kannada words much more consistently since it is an acquired language, and is used mostly on formal occasions. This results in lesser variation in acoustic properties of vowels across phonetic contexts even in natural (read) sentences.

4.3 Why quantification of accents?

Quantification of accents could be useful in speech recognition systems. Based on the knowledge of accent of a given input data, appropriate pronunciation models can be chosen in speech recognition systems, which increases the recognition accuracy of the system. Quantified accent information can be used in deciding the amount of training corpus for adaptation of an acoustical model to a new speaker or speakers of a new accent. Quantification of accents provides a scientific method of measuring the difference between given two accent models. This contributes to better understanding of the spoken language.

In later sections of this chapter, we present the results of experiments that adapts cross entropy as a means of choosing adaptation data that would enhance effectiveness of adaptation. In order to keep the size of the adaptation corpus small, the corpus should be rich in phonemes that have acoustic attributes specific to the accent for which the models need to be adapted. We propose to exploit cross entropy between two models of a phoneme trained with two accent data to identify those phonemes that have marked difference in two accents. The cross entropy of dissimilar models of a phoneme would be high. Hence, adaptation corpus can be created such that it is rich in phonemes with higher cross entropy values. In addition, the amount of adaptation data can also be decided by measuring cross entropy values of accent regions. More speech data are required for adaptation when the cross entropy of two accent regions is high.
4.4 Previous work on accent quantification

A number of research studies were reported on accent in English and Chinese languages. Most of the research work were conducted to classify or identify accents, and to study acoustic correlates of accents. Examples of features used as accent sensitive traits for accent classification are prosodic features (phoneme duration, pitch), slope of intonation contour \(^{27}\) and acoustic features (MFCC, log-energy).

A work was done on NATO N-4 corpus \(^2\), which showed that normalized to exhibit a potential discriminative value for classifying English Canadian, French Canadian and UK accents. Similar experiment on the same corpus with prosodical features showed that the sentence duration varies distinctively among these three accents. Accent analysis and classification were explored with Hidden Markov Model (HMM) codebooks of three types of accented English by non-native speakers and of standard American English \(^8\). Support vector machine and HMM were proposed for accent classification by Tang \(^75\). Also, spectral emphasis was used as feature for accent detection \(^83\). Recently, a perceptual assessment of accent variation in US native English was given by Lin \(^40\). The accent detection in nine diphthongs between standard Mandarin and Shanghai accented Mandarin with medium accent was studied by Li et al. \(^37\). They suggested that supra-segmental features played an important role in rating accents of three dialectal regions Shanghai, Wuhan and Xiamen.

In the above studies, accents have been analyzed without recourse to quantification of accent. In contrast, Seyed Ghorshi et al. introduced an “accent metric” that quantified the effect of American, Australian and British accents on acoustic realization of phonemes \(^72\). The accent metric was based on the cross entropy of probability models of speech from different accents. They demonstrated that grouping of phonemes in the three accents, as reflected in phonetic trees, had a high degree of correspondence with the cross entropies between pairs of accents. In the later work \(^71\), they showed similar correspondence between cross entropy and error in recognition of accented speech by a model trained with speech of another
accent. In both these studies, formant frequencies were used as the parameters for developing HMM models. Estimating formants for unvoiced sounds such as /k/ is difficult; formants are poorly defined for unvoiced sounds. In addition, reliable detection of phoneme boundaries is still a research problem. Quantitative measures of differences in accents that rely upon formant frequencies is not very reliable. It is desirable to have a more robust method of quantification. Hence, we propose to use MFCCs as features to train phoneme HMM models and use cross entropy between phoneme HMM models of two different accents to quantify the degree of dissimilarity between two accents.

**4.5 Proposed methods of quantification of accents using Cross Entropy**

Speech sounds are characterized mainly in terms of spectral properties in the context of speech recognition. Difference in two accent models will manifest as difference between probability distributions that constitute statistical models of speech sounds. If two accents are different, the density functions of the two models are likely to be dissimilar. So a quantitative measure of pairwise distance between densities of two accent models is likely to give an objective measure of separation of two accents. The difference between a pair of densities can be quantified by cross entropy. We propose to use pairwise cross entropy to derive a measure of dissimilarities of two accents.

The Kullback-Leibler[36] divergence is defined as the mean information discrimination between the probability density functions of random variables. Let $p_1$ and $p_2$ denote probability densities of two discrete random variables $M_1$ and $M_2$, the Kullback-Leibler divergence also known as Cross Entropy) of $p_2$ from $p_1$ is defined as

$$CE (M_1, M_2) = \sum_x p_1 (x) \log_2 \left( \frac{p_1 (x)}{p_2 (x)} \right)$$

This measure is also known as cross entropy. The Kullback-Leibler divergence is
not a true distance metric since it is not symmetric; a symmetric version can be constructed.

4.6 Cross entropy of two accent models

Speech recognition is a process of matching an utterance (characterized as a sequence of feature vectors) with all possible sentence HMMs (constructed as a sequence of phoneme HMMs) and finding the best match. A phoneme HMM consists of a sequence of states, each state associated with a time-invariant probability density function (pdf); probability distributions are assumed to be time invariant because speech signal corresponding to a phoneme is assumed to be piecewise stationary. Generally, probability distribution is assumed to be a multi-variate normal; the likelihood of a n-dimensional feature vector $x$ matching a state is given by

$$p(x \mid N(\mu, \Sigma)) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

(4.2)

where $N(\mu, \Sigma)$ is a normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$, associated with that state.

Recognition accuracy of accented speech data using an accent independent model will not be generally high. On the other hand, if we adapt an accent independent model with data of an accent region, this accent adapted model will recognize accented speech with better accuracy. Suppose there are two accent regions: say $A_1$ and $A_2$. By using data from accent region $A_1$, we can train a set of phoneme HMMs; Let $M_1$ denote this HMM model. Similarly, we can train a set of phoneme models, $M_2$ for the accent region $A_2$. A HMM model set (either $M_1$ or $M_2$) comprises of one phoneme HMM for every phoneme of the language. So, for each phoneme there will be two (trained) phoneme HMMs corresponding to the two accent regions, $A_1$ and $A_2$. The likelihood of a feature vector $x$ with respect to a phoneme HMM in either $M_1$ or $M_2$ can be computed using equation
Consider a feature vector \( x \) whose true phoneme identity is \( ph_i \). Let \( p_1(x) \) and \( p_2(x) \) denote likelihoods of feature vector \( x \) being generated by the phoneme as modeled by the pdfs in two HMM models \( M_1 \) and \( M_2 \) respectively. Then the cross entropy \( CE_i(M_1, M_2) \) of \( i^{th} \) phoneme \( ph_i \) with respect to the two accent models can be computed using equation 4.1. In this computation, summation is overall feature vectors \( x \) belonging to that phoneme. We label \( CE_i(M_1, M_2) \) as phoneme-wise cross entropy. We define cross entropy of \( M_1 \) (of accent region \( A_1 \)) with respect to model \( M_2 \) (of accent region \( A_2 \)) as the expectation of phoneme-wise cross entropies overall \( (P) \) phonemes:

\[
CE(M_1, M_2) = \frac{1}{P} \sum_{i=1}^{P} CE_i(M_1, M_2)
\]

In this work, we use the above definition of cross entropy as an objective measure of difference between two accent models. We propose three methods of computing the cross entropy between two accent models. These three methods differ in the approach to computation of phoneme-wise cross entropies. Specifically, these approaches differ in the way expectation is carried out in equation 4.1. If the expectation is overall feature vectors belonging to the phoneme, we call this approach as frame level cross entropy estimation approach.

A note about the term frame is in order. A feature vector is computed from every frame of speech (segment of about 25msec long). Henceforth, we use the terms frame and feature vector interchangeably.

Given an utterance, its transcription, and a trained model, we can forcibly align the feature vector sequence of the utterance with a sentence HMM constructed as a sequence of phoneme HMMs drawn from the trained model. Such a forced Viterbi alignment assigns sets of feature vectors to each state of each phoneme. Therefore, for a feature vector \( x \) belonging to \( i^{th} \) phoneme \( ph_i \), we can compute the probability \( p_1(x) \) with respect to the pdf of the state (to which the feature vector \( x \) is assigned) while aligning with respect to model \( M_1 \). We can
compute similar likelihood \( p_2(x) \) while aligning with respect to model \( M_2 \). In the following, we explain other two approaches of estimating phoneme-wise cross entropy.

1. State level: First, we compute the average of likelihoods of all the feature vectors assigned to a state. This average likelihood replaces \( p(x) \) in equation 4.1 and expectation is carried over all the states of the phoneme in order to compute phoneme-wise cross entropy. We call this approach as state level cross entropy estimation approach.

2. Phoneme level: The average of likelihoods of all feature vectors belonging to all states of a phoneme is computed. If we denote this average likelihood as \( p(x) \), the phoneme-wise cross entropy computation will reduce to

\[
p_1(x) \log_2 \left( \frac{p_1(x)}{p_2(x)} \right)
\]

We call this approach as phoneme level cross entropy estimation approach.

Computation of cross entropy values at different levels can help us in better analysis and understanding of accent variations. The following sub sections explain the methodology used in our work to compute phoneme-wise cross entropy using the above mentioned three approaches.

### 4.6.1 Frame level cross entropy estimation

In this approach, the phoneme-wise cross entropy is computed by direct application of equation 4.1 i.e., expectation is carried out over all feature vectors belong to that phoneme. Consider an utterance containing several phonemes spoken by a person belonging to accent region \( A_1 \). Let \( i^{th} \) phoneme contain \( N \) frames. If we align the feature vector sequence extracted from this utterance with the sentence HMM constructed from the model \( M_1 \), the \( N \) frames of the \( i^{th} \) phoneme ideally gets assigned to \( S \) HMM states of the \( i^{th} \) phoneme. Let \( n_1, n_2 ... n_S \) denote the number of frames assigned to \( s_1, s_2, ... s_S \) states respectively such that \( n_1 + n_2 + ... + n_S = N \).
Let \( p (A_1, f, s, M_1) \) denote the probability of \( f \text{th} \) feature vector of \( i \text{th} \) phoneme from accent region \( A_1 \) assigned to \( s \text{th} \) state of the phoneme of model \( M_1 \) as computed using equation 4.2. If the same feature vector sequence is aligned using the HMM model of the second accent region \( (M_2) \), then the frame state mapping could be different. However, in this work, we assume an ideal situation where the \( f \text{th} \) feature vector get assigned to the same state \( (s) \) while aligning with respect to model \( M_2 \) as well. Then we can compute \( p (A_1, f, s, M_2) \), the probability of the \( f \text{th} \) feature vector of region \( A_1 \) being matched with \( s \text{th} \) state of model \( M_2 \). We can quantify the difference in degrees of representation of speech from accent region \( A_1 \) by HMM models \( M_1 \) and \( M_2 \) as

\[
CE_i (A_1, M_1, M_2) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{f=1}^{n_s} p (A_1, f, s, M_1) \log_2 \left( \frac{p(A_1, f, s, M_1)}{p(A_1, f, s, M_2)} \right) \quad (4.4)
\]

Similarly, we can compute probabilities for feature vector of utterances from accent region \( A_2 \). Let \( p (A_2, f, s, M_1) \) and \( p (A_2, f, s, M_2) \) denote the probabilities of \( f \text{th} \) feature vector of region \( A_2 \) being matched with \( s \text{th} \) state of model \( M_1 \) and \( M_2 \) respectively. We can quantify the difference in degrees of representation of speech from accent region \( A_2 \) by HMM models \( M_1 \) and \( M_2 \) as

\[
CE_i (A_2, M_1, M_2) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{f=1}^{n_s} p (A_2, f, s, M_1) \log_2 \left( \frac{p(A_2, f, s, M_1)}{p(A_2, f, s, M_2)} \right) \quad (4.5)
\]

We define an objective measure of difference between two accent models \( M_1 \) and \( M_2 \) as

\[
CE_i (M_1, M_2) = CE_i (A_1, M_1, M_2) + CE_i (A_2, M_1, M_2) \quad (4.6)
\]

We label \( CE_i (M_1, M_2) \), defined by equation 4.6 as frame level cross entropy of \( i \text{th} \) phoneme.

### 4.6.2 State level cross entropy estimation

In this approach, the average of likelihoods of all feature vectors assigned to a state are computed first. This average likelihood replaces \( p (x) \) while computing
cross entropy (using equation 4.1) of one model (say model $M_1$) with respect to another (say model $M_2$). We define state level cross entropy as

$$CE_i (M_1, M_2) = \frac{1}{S} \sum_{s=1}^{S} p_1 (A_1, s, M_1) \log_2 \left( \frac{p_1 (A_1, s, M_1)}{p_2 (A_2, s, M_2)} \right) + \frac{1}{S} \sum_{s=1}^{S} p_1 (A_2, s, M_1) \log_2 \left( \frac{p_1 (A_2, s, M_1)}{p_2 (A_2, s, M_2)} \right)$$

(4.7)

### 4.6.3 Phoneme level cross entropy estimation

In phoneme level cross entropy estimation approach, the cross entropy is computed at a higher hierarchical level compared to frame and state level cross entropy estimation methods. Here, we compute the average of likelihoods of all feature vectors assigned to a phoneme. This average likelihood replaces $p (x)$ while computing cross entropy (using equation 4.1) of one model (say model $M_1$) with respect to another (say model $M_2$). We define phoneme level cross entropy as

$$CE_i (M_1, M_2) = p_1 (A_1, M_1) \log_2 \left( \frac{p_1 (A_1, M_1)}{p_2 (A_1, M_2)} \right) + p_1 (A_2, M_1) \log_2 \left( \frac{p_1 (A_2, M_1)}{p_2 (A_2, M_2)} \right)$$

(4.8)

### 4.7 Experimental details

In this section, we will first describe the general setup of accent quantification system. Then we explain speech recognition system details and how we conducted phoneme recognition experiments for comparison of cross entropy results. Finally, we will describe computation of cross entropy for given accent models.

#### 4.7.1 Cross Entropy quantification system

A Markov model based recognition system is formulated in order to generate phoneme HMM models. A flow diagram of the system for assigning feature vectors to appropriate phoneme is shown in figure 4.3. We train phoneme HMM models on a given accent database to generate the monophone HMM models. Next, a sentence HMM is constructed by concatenating phoneme HMMs corresponding to the transcription of the utterance. This sentence HMM is aligned with the
feature vector sequence of input speech signal to find the phoneme boundaries. This alignment gives information about features vectors assigned to each state of each phoneme. For each accent, phoneme HMM model are generated along with set of feature vectors assigned to corresponding phonemes. The cross entropy between two accents is computed using the HMM phoneme models and phoneme features of given two accents as given in equation 4.3. Figure 4.4 shows the block diagram illustration of cross entropy computation procedure. Cross entropy of two accent regions is the sum of cross entropies computed using features from both the regions. Calculation of probability density value using models and features from different accent region helps in knowing the difference which exists between the models. Computation of probability density value \( p_2 \) with features of accent region 1 and model2 gives the measure of how well the features of accent1 fits into model2 of accent2. Since the cross entropy is not symmetric (see equation 4.1), we use symmetric cross entropy as defined in equation 4.9. The symmetric cross entropy \( CE \) of two accent regions each modeled by \( P \) HMM phoneme models is computed as

\[
CE = \frac{1}{2P} \sum_{i=1}^{P} CE_i(M_1, M_2) + CE_i(M_2, M_1)
\]  

where \( M_1(i) \) and \( M_2(i) \) represent the HMM of \( i^{th} \) phoneme of accent region 1 and accent region 2 respectively. Equal importance (weight) has been given to all phonemes in computation of cross entropy in equation 4.9. However, frequencies of occurrence of different phonemes are different, in general. For example, vowels occur more frequently than consonants. So, we propose a weighted version of cross entropy that takes into account the frequency distribution of occurrences of phonemes. We define weighted symmetric cross entropy as

\[
CE = \frac{1}{2P} \sum_{i=1}^{P} \frac{CE_i(M_1, M_2) \times K_i(A_1) + CE_i(M_2, M_1) \times K_i(A_2)}{K_i(A_1) + K_i(A_2)}
\]  

where \( K_i(A_1) \) and \( K_i(A_2) \) denote the number of times \( i^{th} \) phoneme occurred in data of region \( A_1 \) and \( A_2 \) respectively.
Figure 4.3: A flow diagram of assigning feature vectors to each phoneme. Phoneme HMM models are trained using the transcription of speech data and feature vectors. Viterbi alignment of feature vector sequence with the sequence of trained HMMs phonemes of the utterance yields boundaries between the phonemes. Using the boundary information feature vectors are assigned to appropriate phonemes.

4.7.2 Illustration of cross entropy computation

Let an instance of a phoneme has 4 frames \( \{x_1, x_2, x_3, x_4\} \) assigned to 3 states \( \{s_1, s_2, s_3\} \) as follow

\[
(x_1) \rightarrow s_1, \ (x_2, x_3) \rightarrow s_2, \ (x_4) \rightarrow s_3.
\]

Let \( M_i \) be the probability distribution function of \( i^{th} \) state. Let the number of states, frames in a phoneme and frame-state alignment in accent regions \( A_1, A_2 \) be the same.
Figure 4.4: Illustration of cross entropy computation for two accent regions accent1 and accent2. The region accent1 is represented by the HMM model $M_1$ and feature vector set $x_1$. Similarly, Accent2 region is represented by the HMM model $M_2$ and feature vector set $x_2$. Cross entropy $CE(M_1, M_2)$ of accent2 from accent1 is computed as the sum of cross entropy values $CE_1$ and $CE_2$ (Equation 4.6).

Frame level cross entropy computation:

Compute cross entropy of frame $x_i$ which is aligned to state $s_i$ (with pdf $M_i$) and belong to region $A_1$ with respect to model $M_i$, state $s_i$ of region $A_2$ as explained in the section 4.6.1.

$$\begin{align*}
\alpha &= p(x_1|M_1, A_1) \log \frac{p(x_1|M_1, A_1)}{p(x_1|M_2, A_2)} \\
\beta &= \frac{1}{2} \left( p(x_2|M_2, A_1) \log \frac{p(x_2|M_2, A_1)}{p(x_2|M_2, A_2)} + p(x_3|M_2, A_1) \log \frac{p(x_3|M_2, A_1)}{p(x_3|M_2, A_2)} \right) \\
\gamma &= p(x_4|M_3, A_1) \log \frac{p(x_4|M_3, A_1)}{p(x_4|M_3, A_2)}
\end{align*}$$

where $\alpha$ is the cross entropy of frame $x_1$ corresponding to state $s_1$, $\beta$ is the average cross entropy of frames $x_2$ and $x_3$ corresponding to state $s_2$, $\gamma$ is the cross entropy of frame $x_4$ corresponding to state $s_3$. 


Similarly, compute the cross entropy taking the frame $x'_i$ of region $A_2$,

$$
\alpha' = \log \frac{p(x'_1|M_1, A_1)}{p(x'_1|M_1, A_2)}
\beta' = \frac{1}{2} \left( \log \frac{p(x'_2|M_2, A_1)}{p(x'_2|M_2, A_2)} + \log \frac{p(x'_3|M_2, A_1)}{p(x'_3|M_2, A_2)} \right)
\gamma' = \log \frac{p(x'_4|M_3, A_1)}{p(x'_4|M_3, A_2)}
$$

where $\alpha'$ is the cross entropy of frame $x'_1$ corresponding to state $s_1$, $\beta'$ is the average cross entropy of frames $x'_2$ and $x'_3$ corresponding to state $s_2$, $\gamma'$ is the cross entropy of frame $x'_4$ corresponding to state $s_3$. Frame level cross entropy of a phoneme belong to region $A_1$ with respect to $A_2$ can be computed as (Equation 4.6),

$$
CE_{frame}(A_1, A_2) = \frac{1}{3} (\alpha + \beta + \gamma) + \frac{1}{3} (\alpha' + \beta' + \gamma')
$$

**State level cross entropy computation:**

The Gaussian probability values of frames assigned to a state are averaged before computing the cross entropy corresponding to a state. Calculate average Gaussian probability values of feature vectors ($x_i$ from region $A_1$, $x'_i$ from region $A_2$)
correspond to a state,

\[
\begin{align*}
p(s_1, A_1) &= p(x_1 | M_1, A_1) \\
p(s_2, A_1) &= \frac{1}{2} (p(x_2 | M_2, A_1) + p(x_3 | M_3, A_1)) \\
p(s_3, A_1) &= p(x_4 | M_4, A_1) \\
p(s_1, A_2) &= p(x_1 | M_1, A_2) \\
p(s_2, A_2) &= \frac{1}{2} (p(x_2 | M_2, A_2) + p(x_3 | M_3, A_2)) \\
p(s_3, A_2) &= p(x_4 | M_4, A_2) \\
p'(s_1, A_1) &= p(x'_1 | M_1, A_1) \\
p'(s_2, A_1) &= \frac{1}{2} (p(x'_2 | M_2, A_1) + p(x'_3 | M_3, A_1)) \\
p'(s_3, A_1) &= p(x'_4 | M_4, A_1) \\
p'(s_1, A_2) &= p(x'_1 | M_1, A_2) \\
p'(s_2, A_2) &= \frac{1}{2} (p(x'_2 | M_2, A_2) + p(x'_3 | M_3, A_2)) \\
p'(s_3, A_2) &= p(x'_4 | M_4, A_2)
\end{align*}
\]

Compute the cross entropy of state,

\[
\begin{align*}
\alpha &= p(s_1, A_1) \log \frac{p(s_1, A_1)}{p(s_1, A_2)} \\
\beta &= p(s_2, A_1) \log \frac{p(s_2, A_1)}{p(s_2, A_2)} \\
\gamma &= p(s_3, A_1) \log \frac{p(s_3, A_1)}{p(s_3, A_2)} \\
\alpha' &= p'(s_1, A_1) \log \frac{p'(s_1, A_1)}{p'(s_1, A_2)} \\
\beta' &= p'(s_2, A_1) \log \frac{p'(s_2, A_1)}{p'(s_2, A_2)} \\
\gamma' &= p'(s_3, A_1) \log \frac{p'(s_3, A_1)}{p'(s_3, A_2)}
\end{align*}
\]

where \(\alpha, \beta, \gamma\) are the cross entropy values computed with feature vectors \(x_i\) from region \(A_1\), \(\alpha', \beta', \gamma'\) are the cross entropy values computed with feature vectors \(x'_i\) from region \(A_2\).

State level cross entropy of a phoneme belong to region \(A_1\) with respect to \(A_2\) can be computed as (Equation 4.11),

\[
CE_{state}(A_1, A_2) = \frac{1}{3} (\alpha + \beta + \gamma) + \frac{1}{3} (\alpha' + \beta' + \gamma')
\]
Phoneme level cross entropy computation:

Phoneme level cross entropy is computed by averaging the probability values of all the three states. From equation 4.11 we have computed \( p(s_1, A_1), p(s_2, A_1), p(s_3, A_1), p(s_1, A_2), p(s_2, A_2), p(s_3, A_2), p'(s_1, A_1), p'(s_2, A_1), p'(s_3, A_1), p'(s_1, A_2), p'(s_3, A_2). \)

Compute the average probability values of all the states of corresponding to the phoneme.

\[
\begin{align*}
    p_1 &= \frac{1}{3} (p(s_1, A_1) + p(s_2, A_1) + p(s_3, A_1)) \\
    p_2 &= \frac{1}{3} (p(s_1, A_2) + p(s_2, A_2) + p(s_3, A_2)) \\
    p'_1 &= \frac{1}{3} (p'(s_1, A_1) + p'(s_2, A_1) + p'(s_3, A_1)) \\
    p'_2 &= \frac{1}{3} (p'(s_1, A_2) + p'(s_2, A_2) + p'(s_3, A_2))
\end{align*}
\]

Phoneme level cross entropy of a phoneme belong to region \( A_1 \) with respect to \( A_2 \) can be computed as (Equation 4.6.3),

\[
CE_{\text{phoneme}} (A_1, A_2) = p_1 \log \frac{p_1}{p_2} + p'_1 \log \frac{p'_1}{p'_2}
\]

4.7.3 Speech recognition/adaptation experimental conditions

The speech modeling used in this study is HMM based and relies on a context-independent modeling of the phonemes. The monophone HMM models were created for all the phonemes with five states including start and exit state. Input speech signals from database discussed in section 3.2.4 were pre-emphasized with pre-emphasis coefficient 0.97. Hamming window of length 25 ms was applied to the speech segments. The acoustic analysis computes MFCC features as explained in Section 2.3.2. MFCC features were computed with a frame width of 25 ms and frame shift of 10 ms. A single Gaussian model was used to model 39 MFCC acoustic features: MFCC and energy, plus their first and second order temporal derivatives.
Table 4.2: Phoneme recognition accuracies (%) when a trained model of a given accent region is used to recognize speech data from same accent region and other two accent regions. It can be observed that the recognition accuracy decreased while recognizing data from one accent region with phoneme models of other accents. However, the decrease in recognition accuracy is marginal. In addition, the phoneme recognition accuracy itself is very low due to lack of good phoneme representation (monophones).

The overall goal of the experiment is to quantify the accent differences between two given accent regions and selectively choose accent adaptation data based on quantified accent information. So, we computed the phoneme recognition accuracy that depends only on acoustic information of input speech data. Kannada dictionary described in section 3.2.5 was used for all the accent regions. Phoneme recognition experiments were run with ergodic language model. Round robin phoneme recognition experiments were conducted on the three accent region speech data set.

In order to measure the accent difference based on acoustical features, we conducted a round robin experiment. For each target accent two sets of experiments were conducted. The target accent was compared to remaining two accent regions to compute cross entropy. Figure 4.8 shows quantified accent values of phonemes. Based on the quantified values of phonemes which belonged to the given two regions, we conducted adaptation of accent regions using selective speech data. Results of adaptation experiments based on the quantified accent values are shown in table created for three accent regions.
Figure 4.5: Phoneme labels vs phoneme-wise cross entropy values $C E_i (M_1, M_2)$. Boxes indicate the phoneme cross entropy value while, solid line shows phoneme repetition count ($\log_{10}$ (phoneme count)). (a) Top panel: Cross entropy values of south region phoneme models with respect to north region phoneme models. (b) Bottom panel: Cross entropy values of coastal region phonemes with respect to north region phonemes.
Figure 4.6: Phoneme labels vs phoneme cross entropy values. Boxes indicate the phoneme cross entropy value, while, solid line shows phoneme repetition count ($\log_{10}(\text{phoneme count})$). (a) Cross entropy values of north region phonemes with respect to south region phonemes. (b) Cross entropy values of coastal region phonemes with respect to south region phonemes.
Figure 4.7: Phoneme labels vs phoneme cross entropy values. Boxes indicate the phoneme cross entropy value while, solid line shows phoneme repetition count ($\log_{10}$ (phoneme count)). (a) Cross entropy values of north region phonemes with respect to coastal region phonemes. (b) Cross entropy values of south region phonemes with respect to coastal region phonemes.
Figure 4.8: Speech recognition accuracy for different approaches for selection of adaptation data. +: solid lines indicate recognition results of adaptation without cross entropy selection criterion. ○: dotted lines indicate the recognition results when adaptation data contain only those phonemes whose cross entropy is in the range [-10:10]; ▲: double dotted lines shows the recognition results when adaptation data contain only those phonemes whose cross entropy is in the range [-20:20]. (a)Top-left panel: North model - South adaptation data, (b)Top-right panel: North model - Coastal adaptation data, (c)Middle-left panel: South model - North adaptation data, (d)Middle-right panel: South model - Coastal adaptation data, (e)Lower-left panel: Coastal model - North adaptation data, (f)Lower-right panel: Coastal model - South adaptation data.
Average symmetric cross entropy

<table>
<thead>
<tr>
<th>Accent regions</th>
<th>North-South</th>
<th>North-Coastal</th>
<th>South-Coastal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame level</td>
<td>6.81</td>
<td>5.04</td>
<td>0.30</td>
</tr>
<tr>
<td>State level</td>
<td>9.27</td>
<td>6.25</td>
<td>0.86</td>
</tr>
<tr>
<td>Phoneme level</td>
<td>2.66</td>
<td>2.10</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Weighted average symmetric cross entropy

<table>
<thead>
<tr>
<th>Accent regions</th>
<th>North-South</th>
<th>North-Coastal</th>
<th>South-Coastal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame level</td>
<td>1.23</td>
<td>0.85</td>
<td>0.006</td>
</tr>
<tr>
<td>State level</td>
<td>2.67</td>
<td>1.80</td>
<td>-0.071</td>
</tr>
<tr>
<td>Phoneme level</td>
<td>1.22</td>
<td>1.14</td>
<td>-0.148</td>
</tr>
</tbody>
</table>

Table 4.3: Symmetric cross entropy values, $CE(M_1, M_2)$ of pairs of accent regions. Average symmetric cross entropy and weighted average symmetric cross entropies follow the similar trend in distance between the accent regions. It is clear that southern accent region is more close to coastal region than northern accent region. Also, distance between North and southern region is high compared to other accent region distances.

Average asymmetric cross entropy

<table>
<thead>
<tr>
<th>Accent regions</th>
<th>N-S</th>
<th>N-C</th>
<th>S-N</th>
<th>S-C</th>
<th>C-N</th>
<th>C-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame level</td>
<td>13.15</td>
<td>9.73</td>
<td>0.45</td>
<td>0.18</td>
<td>0.35</td>
<td>0.42</td>
</tr>
<tr>
<td>State level</td>
<td>41.40</td>
<td>30.15</td>
<td>-22.87</td>
<td>2.44</td>
<td>-17.63</td>
<td>-0.72</td>
</tr>
<tr>
<td>Phoneme level</td>
<td>15.43</td>
<td>16.1</td>
<td>-10.12</td>
<td>3.97</td>
<td>-11.90</td>
<td>-3.19</td>
</tr>
</tbody>
</table>

Weighted average asymmetric cross entropy

<table>
<thead>
<tr>
<th>Accent regions</th>
<th>N-S</th>
<th>N-C</th>
<th>S-N</th>
<th>S-C</th>
<th>C-N</th>
<th>C-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame level</td>
<td>2.01</td>
<td>1.37</td>
<td>0.22</td>
<td>0.01</td>
<td>0.22</td>
<td>0.001</td>
</tr>
<tr>
<td>State level</td>
<td>11.79</td>
<td>11.28</td>
<td>-9.24</td>
<td>1.73</td>
<td>-9.61</td>
<td>-1.72</td>
</tr>
<tr>
<td>Phoneme level</td>
<td>6.97</td>
<td>9.18</td>
<td>-6.30</td>
<td>3.34</td>
<td>-8.53</td>
<td>-3.36</td>
</tr>
</tbody>
</table>

Table 4.4: Asymmetric cross entropy values of accent region pairs.

4.8 Results and discussion

The goal of the study is to quantify the accents. The underlying concept of accent quantification approach is to measure the distance between HMM phoneme models of accent regions. Cross entropy is a quantitative description of the description between the accents. If the two accents ($\text{accent}_1$, $\text{accent}_2$) are farther then a model...
Table 4.5: Coefficients of correlation between phoneme recognition accuracy and asymmetric cross entropy values of accent region pairs. Correlation coefficients were computed for average asymmetric and weighted average asymmetric cross entropy values (Equation 4.10). Negative correlations demonstrates that asymmetric cross entropy is a good measure of difference between two accent regions because higher the cross entropy (i.e., higher difference) implies lower recognition accuracy when data of one accent region is recognized using models of another accent region.

There are six different ways of characterizing the difference between accents as shown in table 4.3 and 4.4. Frame level, state level and phoneme level symmetric cross entropy computation were computed as given in equation 4.9. In addition, weighted symmetric cross entropies were computed by multiplying asymmetric cross entropy with corresponding phoneme repetition count as shown in equation 4.10. Weighted cross entropies at frame, state and phoneme levels were computed to characterize the difference in accents between regions. It can be observed from the symmetric cross entropy results given in tables 4.3 that the coastal region is more closer southern region and northern region is closer to coastal. The trend in all six methods of symmetric cross entropy are the same. However, the phoneme level asymmetric cross entropy computation shows that the northern region is closer to southern region compared to coastal region. Table 4.4 shows the asymmetric cross entropy results. Similar to the symmetric cross entropy results, southern region is closer to coastal compared to northern region in all the six asymmetric cross entropy computations.

To substantiate the above points, table 4.2 show phoneme recognition results.
It can be observed from the results that phoneme recognition accuracy is high when a model is trained and tested on the speech data from the same accent region. The difference between recognition accuracies when northern accent region is tested on coastal and southern accents is marginal. Southern region is marginally closer to northern region which matches the general perception regarding the accent regions. Also, coastal region is marginally closer to northern region compared to southern region. The difference in recognition accuracy when coastal model is tested on northern and southern regions is 0.5. Since the difference in recognition accuracies are small, it is difficult to draw strong conclusions. However, the trend in phoneme recognition accuracies along with cross entropy results can be useful in comparing the distances between the accent regions.

The results of the cross entropy experiment and phoneme recognition experiment are compared by computing correlation coefficient. A correlation describes the strength of association between variables cross entropy and phoneme recognition experiments. We have used Pearson product-moment correlation coefficient [54] which is obtained by dividing the covariance of the two variables by the product of their standard deviations. Six correlation coefficients computed between phoneme recognition accuracy, asymmetric and weighted asymmetric cross entropy are given in table 4.5. High correlation (-0.91) is observed for phoneme level(weighted) cross entropy computation. A low correlation value is observed for frame level(weighted) approach. However, results clearly show high correlation in five cases. Hence, it establishes that the cross entropy computed using above mentioned approaches is a good quantitative measure.

In order to selectively train/adapt HMM based speech recognizers, a huge database is required. However, collection of speech data and preparation of transcriptions, are very time consuming. Hence, it is impractical to provide enough training data for model training/adaptation process. Quantified accent information can be useful in choosing the speech data from a large speech data pool. Asymmetric cross entropy values of phonemes gives information about the distance between phoneme of respective accent regions. These cross entropy values
can be useful in selective adaption. Figure 4.5, 4.6, 4.7 shows the asymmetric cross entropy values of respective accent regions. The cross entropy values are positive in north-south, north-coastal comparisons, while the values are negative in south-north, coastal-north comparisons. However, in case of south-coastal and coastal-south cross entropy computation values are mixture of both negative and positive. Also, the range of cross entropy values of south-coastal and coastal-south regions are very smaller compared to ranges of other combinations of regions. The closures of consonants are not included in cross entropy computation and are not used in adaptation experiments. Most of the aspirated phone mes have high value of cross entropy as depicted in figure 4.5. It is expected since aspiration is produced by turbulence noise generated in the vicinity of the glottis.

Phoneme models of accent regions were adapted based on the cross entropy information. The adaptation speech data was chosen based on state level asymmetric cross entropy information. Asymmetric cross entropy of phonemes defines the distance of a given accent region phoneme from a reference accent region phoneme. Hence, a subset of speech data chosen based on the asymmetric cross entropy values were used for adaptation of models to a given data set. For adaptation of speech data on trained models three selection conditions were applied.

1. Phoneme count (without cross entropy)

2. Cross entropy 0 to 10 and Phoneme count

3. Cross entropy 0 to 20 and Phoneme count

Figure 4.8(a) shows the recognition accuracy for selective adaptation of southern accent speech data on northern speech model. The phoneme recognition accuracy for adaptation criterion “Cross entropy 0 to 20 and Phoneme count” was the highest compared to other two conditions. However, phoneme recognition accuracy for adaptation criterion “Cross entropy 0 to 10 and Phoneme count” is relatively less than criterion “Cross entropy 0 to 20 and Phoneme count”. Also, adaptation based on condition “Phoneme count” performs better as the phoneme repetition count increased above 500 compared to condition “Cross entropy 0
to 20 and Phoneme count”. When a model is adapted based on the phoneme repetition count, the maximum number of vowels are adapted as they occur frequently. Figure 4.5 shows that the cross entropy is high for most of the aspirated phonemes. The phoneme repetition count of aspirated phoneme is less compared to the vowels. Furthermore, when a model is adapted based on the phoneme count it is actually adapted to the vowels. Hence, if a model is adapted based on high cross entropy which lists dissimilar phonemes between accent regions, the phoneme recognition accuracy may not increase much as the system is adapted to less frequently occurring phonemes. On the contrary, selective adaptation decreases the adaptation data size and thereby reducing the adaptation complexity.

Trends in phoneme recognition accuracy after the selective adaptation are similar in figure 4.8(a) and 4.8(b). However, the trends in figure 4.8(c)-4.8(f) are different from 4.8(a) and 4.8(b). The phoneme recognition accuracies overlap when southern speech data was adapted to a northern model based on the above mentioned selection procedure. It can be observed in figure 4.6 and 4.7 that the range of cross entropy values differs. Hence, adaptation data selection criterion should be changed based on the cross entropy values of accent regions. Table 4.6 shows phoneme recognition accuracy of adapted models (south and coastal) when tested on speech data from adapted region (coastal and south). South speech model was adapted on coastal speech data and Coastal speech model was adapted on south speech data selectively based on cross entropy conditions. Two cross entropy criterion were used to choose speech data,

1. Cross entropy 0.5 to 5

2. Cross entropy 0.5 to 10

An increase in the phoneme recognition accuracy is observed when adapted selectively compared to the model adapted with all speech data from adaptation region. However, marginal decrease (0.15%) in phoneme recognition accuracy is seen when coastal model is adapted with southern speech data using adaptation data selection criterion “Cross entropy 0.5 to 5”. One of the reasons for decrease
Table 4.6: Speech recognition accuracy results of adapted models. South speech model was adapted on coastal speech data and Coastal speech model was adapted on south speech data selectively based on cross entropy condition. An increase in accuracy is observed with less speech data. Even though the increase in accuracy is marginal, it shows that selective speech adaptation can be useful to adapt the speech data.

in phoneme recognition accuracy is very less speech data adapted using the condition “Cross entropy 0.5 to 5”. Since the low cross entropy values correspond to the similar phonemes in both the accent regions, the accuracy of adapted may not improve.

The cross entropy approach based can be useful for selectively preparing adaptation data. It was shown, that it is possible to select relevant adaptation data based on cross entropy while adapting southern speech data on northern model and while adapting northern speech data on southern model. However, significant improvements in recognition is not observed. Hence, further improvements in approach has to be done for better results.

4.9 Summary

Methods of selecting adaptation data scientifically are explored in this chapter. A qualitative analysis of Kannada vowels in different Kannada speaking regions was done to study the accents of spoken Kannada. The study aimed at finding whether there is any significant difference in the acoustic space of vowels across the regions. If there is any difference, we can quantify and model difference between the accents so as to benefit from such a model in the context of Kannada speech recognition. Experiments showed difference between Kannada accent regions. The results of the proposed cross entropy experiments and phoneme recognition experiment are compared by computing correlation coefficient. High correlation
is observed for phoneme level (weighted) cross entropy computation. Analysis of proposed methods of accent quantification established that the cross entropy computed using above mentioned approaches is a good quantitative measure.