Kannada speech recognition system

3.1 Introduction

Kannada, the official language of Karnataka state, is spoken by about 60 million speakers\textsuperscript{18, 50}. However, there is little research reported on Kannada speech processing compared to other languages of similar importance\textsuperscript{22, 34, 65, 67}. A Kannada number recognition system for recognizing a sequence of digits was reported by Sahana et al.\textsuperscript{63}. In their system, 400 Kannada numbers, each of seven digits long, were recorded from three speakers. HTK was used to develop a number recognition system using MFCC, delta and delta-delta coefficients. Experimental analysis showed that speaker dependent system gave word accuracy of 99\% on trained word data and 97\% on test word data. Sentence accuracy of speaker dependent system was 73\% on trained sentences and 74\% on test sentences. However, it performed poorly on speaker independent experiments with recognition accuracy up to 74\% on trained words and 74\% on test words. The failure of recognition system was mainly caused by insufficiency of speech data.

Automatic Speech Recognition (ASR) systems that recognize native speech have reached a certain maturity. However, the recognition of native speech with
accent variations is still a problem. Often, speech recognition system performance degrades drastically if an recognition system is exposed to accented speech. Accented speech is often encountered and it should not be neglected in speech recognition systems. In this chapter, implementation of accent independent and accent dependent Kannada speech recognition system are described. The methodologies to evaluate the speech recognition system is also described.

Successful research and development of practical speech recognition algorithms and systems depend on the quantity and quality of speech data available. There are large speech databases available on different languages such as English, French and Japan. However, such databases for Kannada language are not available for research except a few small scale databases such as [63]. One of the reasons for the little research in Kannada speech recognition is the lack of Kannada speech database. In this chapter, we also describe creation of Kannada text and speech corpora that were used to train and evaluate Kannada ASR system.

### 3.2 Kannada speech database

Kannada language is spoken by millions of people in Karnataka and overseas Kannada community. For the design, implementation and performance evaluation of speech processing systems, large amount of acoustic data are required. Hence, it is important to develop a speech database for research activities. Following sections describes a study on the design and collection of a speech database for Kannada.

#### 3.2.1 Kannada phonemes/phonology

Kannada language is one of the major south Dravidian languages. The phonemes of Kannada can be grouped into two categories[12].

1. Vowels
2. Consonants
Vowel

There are fourteen vowel phonemes in Kannada spoken language. These include eight short and six long vowels. They are always voiced and syllabic. The short vowels in Kannada are /a/, /i/, /u/, /e/, /o/ and /au/. The long vowels are /A/, /I/, /U/, /E/ and /O/. Below are given some examples of vowel phonemes.

Short vowel examples:
Vowel /a/, e.g. /a/ in Kannada, English translation: “Kannada”
Vowel /i/, e.g. /i/ in giri, English translation: “hill”
Vowel /u/, e.g. /u/ in kuri, English translation: “sheep”
Vowel /e/, e.g. /e/ in ellu, English translation: “where”
Vowel /ai/, e.g. /ai/ in aishwarya, English translation: “Aishwarya”
Vowel /o/, e.g. /o/ in ole, English translation: “oven”

Long vowel examples:
Vowel /A/, e.g. /A/ in Ame, English translation: “turtle”
Vowel /I/, e.g. /I/ in bhAraTIya, English translation: “Indian”
Vowel /U/, e.g. /U/ in Uta, English translation: “swelling”
Vowel /E/, e.g. /E/ in Enu, English translation: “what”
Vowel /O/, e.g. /O/ in Odu, English translation: “read”

Consonant

There are twenty nine consonants in Kannada spoken language. These are classified based on the place of articulation, manner of articulation and state of vocal cords. Table 3.2.1 lists the Kannada phonemes. Stop consonants are listed in a 5×5 table. First column of the table lists voiceless stop consonants /k/, /c/, /T/, /t/ and /p/. Third column names the voiced stop consonants /g/, /j/, /D/, /d/ and /b/. The second and fourth columns of the table names the aspirated stop consonants. The last column of the table lists the nasal sounds /ng/, /ng/, /n/ and /m/. Other non-stop consonants /y/, /r/, /l/, /w/, /sh/, /S/, /s/, /h/ and /L/ are listed in the last row.
<table>
<thead>
<tr>
<th>k</th>
<th>kh</th>
<th>g</th>
<th>gh</th>
<th>ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>ch</td>
<td>j</td>
<td>jh</td>
<td>nj</td>
</tr>
<tr>
<td>T</td>
<td>Th</td>
<td>D</td>
<td>Dh</td>
<td>N</td>
</tr>
<tr>
<td>t</td>
<td>th</td>
<td>d</td>
<td>dh</td>
<td>n</td>
</tr>
<tr>
<td>p</td>
<td>ph</td>
<td>b</td>
<td>bh</td>
<td>m</td>
</tr>
<tr>
<td>y</td>
<td>r</td>
<td>l</td>
<td>w</td>
<td>sh</td>
</tr>
</tbody>
</table>

Table 3.1: Kannada consonants

| a   | A   | i   | I   | u   | U   | e   | E   | ai  | o   | O   | au  |

Table 3.2: TIFR label set for Kannada alphabet

### 3.2.2 Transliteration: Baraha fonts to TIFR symbols conversion

Baraha is a free Kannada software intended to popularize Kannada in the cyber-world. Baraha software supplies fonts encoding which can be used in Microsoft word processors for editing etc. The readable Kannada text was generated from online Kannada text using Baraha fonts during the text data collection for Kannada speech recording. Baraha transliteration rule can be used to convert text written in Kannada script, written using Baraha font encoding to roman characters. Baraha transliteration scheme is similar to writing our names in English. Tata Institute of Fundamental Research (TIFR) speech lab follows a standard labeling scheme for phonemes which is shown in table 3.2.2.
3.2.3 Major accents of Kannada

Since Kannada is spoken by a large number of people in vast areas, it has several accents/dialects, namely

1. Northern Kannada
2. Southern Kannada
3. Coastal Kannada

Northern, Southern and Coastal Kannada has several unique peculiarities. These peculiarities are mainly due to the cultural, geographical and linguistic variations. Northern Kannada is spoken by Hubli and Bagalakot districts. Mysore district Kannada is among the Southern Kannada dialects with several peculiarities. Coastal Kannada is spoken by Udupi and South Canara districts. Figure 3.1 shows the three accent/dialect regions. The geographical distance between these regions is approximately 300 Kms. Hence, the chances of observing mixed accent in these regions is low.

3.2.4 Speech corpus

Spoken language corpora was designed and collected with the primary objective of development of Kannada accent database. The accent corpus consists of 89 speakers who belong to three Kannada accent groups of Karnataka namely Coastal, South and North. The coastal and south has 32, 23 speakers respectively whereas North has 34 speakers. Each speaker read 30 Kannada sentences from a list of 3500 Kannada sentences, which approximately correspond to 3 minutes of speech. In sum, 4.5 hours of data was recorded. Approximately 20650 words were spoken, covering a vocabulary of about 11600 words. The Kannada sentences were collected from Kannada text books. The sentences were neutral and non political. Majority of the speakers were natives of each accent region. In order to avoid the mixed accent, the speech data was recorded at the place of each accent region itself. The average age of the speakers ranged between 20 to 50. The corpus was
Figure 3.1: Karnataka state map and three Kannada accent regions

recorded using an unidirectional microphone in Linux environment. The corpus was digitized at 16000Hz and stored in 16-bit PCM format without headers in the speech file. (i.e., raw audio format).

3.2.5 Language model

A language model describes the rules of constructing sentences in a language. It is very useful in speech recognition systems to improve the speech recognition accuracy. Language models help a speech recognizer figure out how likely a word sequence is, independent of the acoustics. Two components of the language model are pronunciation dictionary and bigram grammar model. They are discussed in this section.
Pronunciation dictionary

Pronunciations tend to be more variable in spontaneous speech than in careful read speech. In the later case, where pronunciations of words are more likely to adhere to their citation forms. There is a near one-to-one correspondence between the scripts used by Indian languages and corresponding speech sounds. However, there are certain exceptions in pronunciations of words. These exceptions are captured using a pronunciation dictionary. Firstly, if we consider the plosive sounds which are in the first four column of the table they are characterized by at-least two acoustic events. The plosives are characterized by the closure of the air flow through the mouth at some point in the vocal tract. Hence, in case of unvoiced sounds no energy (silence) is observed in the waveform but a voice bar is observed in the case of voiced sounds which is known as closure. The closure is followed by a release wherein air rushes out, there can be a burst, frication and aspiration. Since the characteristic of the closure and the later events of plosive are acoustically different, we represent a plosive sound by two acoustic units.

Figure 3.2 shows the waveform of voiced and unvoiced plosives. The utterance “kabaka” has a unvoiced plosive sound /k/ and voiced plosive sound /b/. It can be observed from the figure that voiced plosive sound /b/ has a sinusoidal closure whereas unvoiced plosive /k/ has a clear silence before the burst. The acoustic basic units for spoken Kannada are given in table 3.2.5.

Certain words in Kannada spoken language has context dependent multiple pronunciations. Such information has to embedded into the pronunciation dictionaries for a better modeling of speech sound units. A few entries in the pronunciation dictionary look like this,

```
... lekka l e clkk k a ...
... mele m e l ae ...
```
Figure 3.2: Waveform and Spectrogram depiction of utterance “kabaka”. The symbol clk and clb represents the closures of plosives /k/ and /b/ respectively. sil denotes the silence before the unvoiced plosive /k/.

| Table 3.3: Kannada basic acoustic units |
|---------------|-------------|
| a A i I u U e E ai o O au ae |
| clk k       clkh kh   clg g   clgh gh  ng |
| clc c       clch ch   clj j    cljh jh  nj |
| clT T       clTh Th    clD D    clDh Dh  N |
| clt t       clth th    cld d    cldh dh  n |
| clp p       clph ph    clb b    clbh bh  m |
| y r l w sh S s h L |

In lekka and mele, le is written in similar way but pronounced slightly differently since the vowels are changed.

In addition, a word can also be pronounced differently within and across the regions. Therefore, the multiple pronunciation of the same word can be represented with multiple entries in the pronunciation dictionary. An excerpts from
the pronunciation dictionary appears like this,

\[
\begin{align*}
&\ldots \\
&\text{mane} \quad m\ a\ n\ e \\
&\text{mane} \quad m\ a\ n\ ae \\
&\ldots
\end{align*}
\]

We can observe that \textit{mane} is pronounced in two different ways. It ends with vowel \textit{e} and \textit{ae} during two variations of pronunciation. These pronunciation variations has to be incorporated in the pronunciation dictionary for a better modeling of the speech units.

**Bigram grammar**

The language models under speech recognition can be grouped into two broader classes. In case of speech recognition, in a specific test domain, the syntactic constraint can be modeled in the form of a Finite State Automata (FSA) where the nodes are the words and directed arc represent possible word transitions. However, for unconstrained very large speech recognition, statistical grammar are more practical. N-gram is once such grammar where \textit{n}th word can be predicted based on \textit{n} − 1 previous words. When \textit{n} = 2 such a grammar is called bigram grammar. Throughout this work, we use bigram grammar which is defined as

\[
p(i, j) = \begin{cases} 
(N(i, j) - D)/N(i) & \text{if } N(i, j) > t \\
\beta(i)p(j) & \text{otherwise}
\end{cases}
\]

(3.1)

**3.3 Recognition of spoken Kannada sentences**

A Kannada speech recognizer was built using the clean speech corpus discussed in section 3.2.4. This clean speech recognizer is used throughout the experiments in this research for analyzing and comparison of performance with accent dependent speech recognizers. In this section, we introduce the various evaluation methodologies and accent independent Kannada speech recognition system.
Since the recognition task is continuous speech, where the number of words in
an utterance is not known, not only the misclassified words, but also extra words
(insertions) or missing words (deletions) are a source of error. A recognition hy-
pothesis is aligned against a correct transcription using dynamic programming to
minimize the number of misclassified words. Then, the accuracy of word recogni-
tion is computed using the following formula,

\[
Accuracy = \frac{N - D - S}{N} \times 100
\]  

(3.2)

Where \(N\) is the total number of labels in the reference transcriptions, \(D\) deletion
ers, \(S\) substitution errors.

A Kannada speech recognition system was built using Kannada speech data
described in section 3.2.4. Input speech signals are 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. MFCC features
are computed with a frame width of 25 ms and frame shift of 10 ms. A single
Gaussian model is used to model 39 MFCC acoustic features: MFCC and energy,
plus their first and second order temporal derivatives. The system was trained
and tested using entire speech data. The word recognition accuracy of 90.08%
was observed and sentence accuracy was 66.89%.

In reality, while testing we may get speech data which are very different from
trained speech data, resulting in poor performance of speech recognition system.
In order to evaluate the performance of the speech recognition system for unseen
speech data, we divide the speech corpus into training data set and testing data
set. Since the database is small, the train data may be having extremely good
or extremely bad data by chance. Hence, the results are highly influenced by
the way test and train data sets are divided. To avoid bias in choosing data set
for training or testing, we follow round robin or leave-one-out method. Table
3.4 shows results of round robin experiment conducted on Kannada speech data.
Round robin experiment was conducted dividing the Kannada speech data into
Table 3.4: Kannada sentence and word recognition accuracy (%). Three rounds of recognition experiments were conducted on Kannada speech data dividing the speech data into training set and testing set. Minimal variations in recognition accuracies are observed in round robin experiments.

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>Word %</td>
<td>89.43</td>
<td>89.23</td>
<td>89.47</td>
</tr>
<tr>
<td></td>
<td>sent %</td>
<td>61.31</td>
<td>60.36</td>
<td>61.19</td>
</tr>
<tr>
<td>Testing set</td>
<td>Word %</td>
<td>87.52</td>
<td>88.68</td>
<td>87.90</td>
</tr>
<tr>
<td></td>
<td>sent %</td>
<td>59.64</td>
<td>62.24</td>
<td>53.51</td>
</tr>
</tbody>
</table>

In each round, 2/3 of the speech data was used for training and 1/3 was used for testing. Average word recognition accuracy of 89.37% and sentence recognition accuracy of 60.95% is observed on training speech data. While conducting recognition experiments on test speech data, average word recognition accuracy of 88.03% and sentence recognition accuracy of 58.46% is observed. The recognition accuracy of the speech recognition system decreased while recognizing unseen test speech data. Hence, training the system on good speech data is very important which covers all types of speaker variations such as gender, age, dialect and pronunciation.

We compared the speech recognition accuracy results obtained in this study with Hindi speech recognition system[77]. Hindi speech recognition system was trained with data from 90 Hindi speaking speakers with a total of 900 continuous utterances, under clean/noisy conditions. The speech corpus used for training Kannada speech recognition system was three times more than the Hindi speech corpus. In addition, the size of vocabulary of Kannada speech corpus is 11600 words, while, vocabulary of Hindi corpus is 2700. An average word recognition accuracy of 77% is recorded for both training and test data for clean speech in Hindi speech recognition system. Hence, the word recognition accuracy (89%) of Kannada speech recognition system is comparable to Hindi speech recognition system under controlled conditions.
3.4 Recognition of accented Kannada

In Kannada language, there are three prominent accent groups. As discussed in section 3.2.3, the pronunciation of sound units varies across Northern, Southern and Coastal Kannada regions. Accent is one of the important factors which affects the performance of speech recognition systems. The performance of speech recognition systems degrades when speaker accent is different from that in the training data. In this section, accent-specific Kannada speech recognition experiments are described. The study of accent-specific speech recognition will help to analyze the acoustic difference between the accent regions.

Results of the accent recognition experiments are summarized in table 3.5. The general trend suggests that, in accented speech, acoustic models built on appropriate accent training data give high accuracy when tested on speech data from training accent region, compared to other accent regions. It is found that, recognition accuracy decreases when speech from one accent region is recognized using a model trained with data from other accent regions. When an accent model is matched against other accent speech data, recognition accuracy is decreased by approximately 4-5%. This result supports the argument that speech recognition system fail to perform well when the accent speech data differ from that of the accent speech data trained on. Furthermore, comparisons of the results discussed in table 3.4 and table 3.5 proves that performance of speech recognition deteriorates when the test data is different from the training data set.

3.5 Adaptation of acoustic models to accents

As mentioned in section 3.4, accent is one of the important factors in speech recognition systems. The results confirm the impact of accent in recognition systems. Hence, it is important to adapt an unaccented model to a different accent. One of the adaptation techniques used at present is adapting the unaccented model with lot of unaccented data. However, it may not be required to adapt all the phoneme models, as some phonemes may be similar in both the accents. Hence,
Table 3.5: Kannada sentence and word recognition accuracy (%) for every accent speech data trained and tested against other two accent speech data. For each combination of experiments both sentence and word recognition results are presented.

<table>
<thead>
<tr>
<th>Train data</th>
<th>Test data</th>
<th>North</th>
<th>South</th>
<th>Coastal</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>Word%</td>
<td>88.79</td>
<td>85.73</td>
<td>85.28</td>
</tr>
<tr>
<td></td>
<td>sent%</td>
<td>58.77</td>
<td>56.11</td>
<td>57.75</td>
</tr>
<tr>
<td>South</td>
<td>Word%</td>
<td>83.22</td>
<td>88.32</td>
<td>82.37</td>
</tr>
<tr>
<td></td>
<td>sent%</td>
<td>40.32</td>
<td>53.31</td>
<td>42.18</td>
</tr>
<tr>
<td>Coastal</td>
<td>Word%</td>
<td>84.80</td>
<td>85.28</td>
<td>90.64</td>
</tr>
<tr>
<td></td>
<td>sent%</td>
<td>55.30</td>
<td>55.61</td>
<td>66.67</td>
</tr>
</tbody>
</table>

Selective adaptation will decrease the adaptation data size and adaptation time. A measure of difference in accent of two given phoneme/word models could be useful in selective adaptation of speech data; model adapted with such selective data is likely to result in higher speech recognition accuracy. Further discussions on selective adaptation under an accent quantification context is provided in the next chapter.

### 3.6 Summary

The training and performance evaluation of an ASR system that recognizes continuously spoken Kannada sentences by multiple speakers was described in this chapter. Experiments using accent dependent ASR systems illustrated the need for adapting speech models to a new accent in order to achieve better recognition accuracy. Better methods of selecting adaptation data will be explored in the next chapter.