Chapter 3

SPEECH DATA ACQUISITION AND DATABASE CREATION

3.1 Introduction

Automatic speech recognition (ASR) by machine can yield the most natural and efficient method of communication between human and machine. In recent years accurate speech recognition systems are beginning to emerge from various research laboratories with the affirmation that the formation of speech recognition system into realistic operating environments will require a powerful and accurate speech database which contains an appropriate number of speech samples to model the inherent variability in the speech signal. One widely used and well known speech database is the TIMIT database which contains 630 native speakers of American English. Speech databases of several European languages like: English, French, German, Greek, Italian, Spanish, Finnish, Dutch and Danish [Robinson and Renals(1995)] [Schultz(2002)] [Zheng and Wu(2002)] [Tseng and Huang(2003)] [Muthusamy and Godfrey(1995)] [Langmann and den Os(1996)] and Indian languages like: Tamil, Telugu, Marathi, Kannada and Hindi [Gopalakrishna and S.P.Kishore(2005)] have been reported in
the literature. However for Indian language like Malayalam, there is only a relatively small Malayalam speech database is available for the research purpose [Sunilkumar(2002)] [Prajith(2008)].

Malayalam is one of the language in Indian subcontinent. Among the Dravidian languages Malayalam, Kannada, Tamil, Telungu, Tulu and Kongani. Malayalam is the youngest and most dynamic language. Malayalam is ranked as eight in the list of eighteen popular languages in India. This is the principal language of the South Indian state of Kerala and also of the Lakshadweep Islands of the west coast of India spoken by about 36 million people. This thesis is motivated by the knowledge that only little attempts were rendered for the automatic speech recognition of Vowel/Consonant-Vowel (V/CV) speech unit in Indian languages like Hindi, Tamil, Bengali, Marathi etc., and very less works have been found to be reported in the literature on the recognition of V/CV speech units in Malayalam. Very few research attempts were reported so far in the area of Malayalam vowel recognition. Consequently a standard database is not available in the language. So more basic research works are essential in the area of Malayalam V/CV speech unit recognition. This chapter presents the work carried out to create a reasonably large and representative database for Malayalam V/CV speech units.

This chapter is organized as follows. Section 3.2 presents a brief introduction on phonetics. Section 3.3 contains an overview of Malayalam Vowel/Consonant-Vowel (V/CV) sounds. The V/CV speech data acquisition and the need for antialiasing filter in data acquisition system is explained in section 3.4. Finally section 3.5 concludes this chapter.
3.2 Phonological Description of Speech

The field of phonetics includes the study of speech production and the acoustics of the speech signal, and provides a way to effectively describe speech. It can be broadly classified into articulatory phonetics and acoustic phonetics. Articulatory phonetics deals with the articulatory aspects of speech sounds. That is, articulatory phoneticians are interested in how the different structures of the vocal tract, called the articulators (tongue, lips, jaw, palate, teeth etc.), interact to create the specific sounds. Acoustic phonetics is a subfield of phonetics which deals with acoustic aspects of speech sounds. Acoustic phonetics investigates properties like the mean squared amplitude of a waveform, its duration, its fundamental frequency, or other properties of its frequency spectrum, and the relationship of these properties to other branches of phonetics.

3.2.1 Articulatory phonetics

The process of air being expelled from the lungs and pushing through the vocal tract produces speech signals. The resulting sound pressure wave radiates out from the lips. The various organs involved in speech production process are shown in Figure 3.2.1. According to the permutation and combination of their positioning, a large variety of sounds can be produced. In order to discuss these sounds unambiguously, they are categorised into a series of distinct types like, nasals, plosives, fricatives etc. according to how they are produced. The larynx is at the base of the vocal tract and mainly comprises two bands of muscle and tissue called the vocal cords or folds. All air from the lungs must pass through the vocal folds, and they can obstruct its passage to a greater or lesser extent. In terms
of speech production, the vocal folds can operate in three ways:

- vibrating in a pseudo-periodic manner to create voiced sounds. The frequency of this vibration is called the fundamental frequency, and corresponds to the tone heard by a listener which is called pitch;
- not vibrating for unvoiced sounds;
- stopped or closed to produce a glottal stop, the glottis being the gap between the vocal cords;

The different articulatory organs (e.g. tongue, lips, soft-palate) in the vocal tract can be positioned so as to modulate the flow of air through the tract in different ways viz., close, narrow and open as discussed below.

**Closure**: As well as the glottal stop, the vocal tract may be closed at other places
such as at the lips, or between the tongue and hard palate. If the velum is lowered, then air can flow out through the nose creating a *nasal* sound. However, if it is raised, there is no way for the air to escape. Therefore the pressure in the vocal tract increases, and when the closure is removed the air bursts out creating a plosive sound.

**Narrowing**: If rather than completely closing the vocal tract, two speech organs are instead brought close together, then the air flow through them becomes turbulent and produces *fricative* sounds. The narrowing can occur at any point in the vocal tract.

**Open**: With the speech organs sufficiently open so that no turbulence is produced in the airflow, *vowel* sounds are generated. These sounds are always voiced, and it is mainly the position of the highest part of the tongue that determines the vowel produced. This leads to a widely used description in which vowels are specified according to which part of the tongue is highest (front, central, back) and how high it is (close, mid, open).

### 3.2.2 Acoustic phonetics

In the production of speech, an acoustic signal is formed when the vocal organs move resulting in a pattern of disturbance to the air molecules that is propagated outwards in all directions eventually reaching the ear of the listener. Acoustic phonetics is concerned with describing the different kinds of acoustic signal that the movement of the vocal organs gives rise to in the production of speech by male and female speakers across all age groups and in all languages, and under different speaking conditions and varieties of speaking style.
3.3 Vowel/Consonant-Vowel (V/CV) sounds in Malayalam

Generally phones are divided into two classes namely vowels and consonants. Vowels are the most interesting class of sounds in any language. The most practical speech recognition systems rely heavily on vowel recognition to achieve high performance [L. R. Rabiner and Wilpon(1979)][Rabiner and Juang(1993)]. Vowels are produced by exciting a fixed vocal tract with quasi-periodic pulses of air caused by vibration of the vocal cords. Conventional methods used to classify vowels are the articulatory configurations required to produce sounds, typical waveform plots, typical spectrogram plots and formant frequency analysis [Gimson.A(1972)] [Rabiner and Schafer(1978)].

A consonant can be defined as a unit sound in spoken language which is described by a constriction or closure at one or more points along the vocal tract. According to Peter Ladefoged, consonants are just ways of beginning or ending of vowels [Ladefoged(2004)]. Consonants are made by restricting or blocking the airflow in some way and each consonant can be distinguished by where this restriction is made [Jurafsky and Martin(2004)]. The point of maximum restriction is called the place of articulation of a consonant. A consonant also can be distinguished by how the restriction is made. For example, where there is a complete stoppage of air or only a partial blockage of it. This feature is called the manner of articulation of a consonant. The combination of place and manner of articulation is sufficient to uniquely identify a consonant.
Indian languages are mainly classified into three language families namely Indo-European languages, Indo-Aryan languages and Dravidian languages. Malayalam is one of the major languages from Dravidian language family. The earlier writing style of the Malayalam is now substituted with a new style from 1981. Compared to Malayalam and all other Indian languages Tamil seems to be different in the sense that Tamil doesn’t have aspirated sounds and thus the pronunciation is different from other Dravidian language structures. Tamil contains only ‘kharam’ and ‘anunasikam’ sounds and thus the script used to represent ‘mridu’ sounds are using ‘kharam’. In Tamil the pronunciation of ‘kharam’ lies in the range between ‘kharam’ and ‘mridu’ compared to Malayalam. For example the word ‘ganapathi’ pronounced and scripted as ‘kanapathi’. In Bengali the pronunciation of the vowel ‘a’ is replaced with ‘au’. Due to lineage of Malayalam to both Sanskrit and Tamil, Malayalam language structure has the largest number of phonemic utterances among the Indian languages. Malayalam script includes letters capable of representing all the phoneme of Sanskrit and all Dravidian languages. A unique property of Malayalam is ‘chillukal’ which is derived from the basic consonant units. Malayalam language now consists of 51 V/CV units which contain 15 long and short vowel sounds and the remaining 36 basic consonant sounds. The vowels in Malayalam language are given in table 3.3.1.
Table 3.3.1: Vowels in Malayalam

<table>
<thead>
<tr>
<th>Malayalam Vowels</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
</table>
| /a/                     | /
| /a:/                    | /
| /i/                     | /
| /i:/                    | /
| /u/                     | /
| /u:/                    | /
| /˚rt/                   | /
| /e/                     | /
| /e:/                    | /
| /ai ˆ/                  | /
| /o/                     | /
| /o:/                    | /
| /au ˆ /                 | /
| /am/                    | /
| /ah ˆ /                 | /

Analyzing carefully, we will realize that, there are only about 8 basic sounds in this list of 15. In other words, half of them are just modifications of these eight. The 8 basic vowels are: /a/, /i/, /u/, /˚rt/, /e/, /o/, /am/ and /ah ˆ/.

Among these basic units, the unit /ah ˆ/ cannot be included in the vowel list, since it mostly occur only in words accepted from Sanskrit. Thus for the present study the recognition experiments are carried out for 14 vowel units in Malayalam.

The International Phonetic Association (IPA) has produced a set of phonetic symbols to define all of the individual speech sounds (called phonemes) in terms of their place of articulation. The IPA vowel chart for Malayalam is shown in Figure 3.3.1.

In language terminology, in terms of tongue and hump position, tongue and hump height and typical spectrogram studies, the vowels, /i/ and /e/ are classified as front vowels, /u/ and /o/ are classified as back vowels and /a/ is classified as mid vowel. The vowel /i/ is a front vowel articulated by raising the rear part of front of the tongue (i.e. the part of the tongue nearer to center than to the front) in the direction of hard palate, just above the half-close
The vowel /e/ is also a front vowel. During the articulation of this vowel, the front of the tongue is raised in the direction of the hard palate to a height between half-close and half-open. The vowel /o/ is a back vowel and for the articulation of this vowel, the back of the tongue is just above the fully open position. Vowel /u/ is also back vowel. During the articulation of this vowel the front part of the back of the tongue (i.e. the part nearer to the center than the back of the tongue) is raised in the direction of the soft palate to a height just above half-close position. The vowel /a/ is a mid vowel. During the articulation of this vowel the center of the tongue is raised in the direction of the roof of the mouth.

In general a basic consonant unit in Malayalam can be represented by riding each consonant unit with the vowel /a/ and thus we interpret it as Consonant - Vowel (CV) unit. Conventionally the recognition of CV units in any language is a very difficult task as there are more number of CV classes compared to vowel
classes. In Malayalam language if we consider other remaining 14 vowels riding with each consonant unit a total of 504 classes have to be considered. This thesis reports the recognition experiments conducted for 36 basic CV units and these are tabulated in table 3.3.2.

Table 3.3.2: Consonant-Vowel units in Malayalam

<table>
<thead>
<tr>
<th>Malayalam Consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 /ka/</td>
</tr>
<tr>
<td>2 /kʰa/</td>
</tr>
<tr>
<td>3 /ga/</td>
</tr>
<tr>
<td>4 /gʰa/</td>
</tr>
<tr>
<td>5 /ŋa/</td>
</tr>
<tr>
<td>6 /tʃa/</td>
</tr>
<tr>
<td>7 /tʃʰa/</td>
</tr>
<tr>
<td>8 /dʒa/</td>
</tr>
<tr>
<td>9 /dʒʰa/</td>
</tr>
<tr>
<td>10 /na/</td>
</tr>
<tr>
<td>11 /tʰa/</td>
</tr>
<tr>
<td>12 /tʰʰa/</td>
</tr>
<tr>
<td>13 /dʰa/</td>
</tr>
<tr>
<td>14 /dʰʰa/</td>
</tr>
<tr>
<td>15 /θa/</td>
</tr>
<tr>
<td>16 /θʰa/</td>
</tr>
<tr>
<td>17 /θ̌a/</td>
</tr>
<tr>
<td>18 /θ̌ʰa/</td>
</tr>
<tr>
<td>19 /θ̌θ̌a/</td>
</tr>
<tr>
<td>20 /θ̌orage/</td>
</tr>
<tr>
<td>21 /pʰa/</td>
</tr>
<tr>
<td>22 /pʰʰa/</td>
</tr>
<tr>
<td>23 /bʰa/</td>
</tr>
<tr>
<td>24 /bʰʰa/</td>
</tr>
<tr>
<td>25 /va/</td>
</tr>
<tr>
<td>26 /ja/</td>
</tr>
<tr>
<td>27 /jʰa/</td>
</tr>
<tr>
<td>28 /i̯a/</td>
</tr>
<tr>
<td>29 /u̯a/</td>
</tr>
<tr>
<td>30 /ca/</td>
</tr>
<tr>
<td>31 /řa/</td>
</tr>
<tr>
<td>32 /řʰa/</td>
</tr>
<tr>
<td>33 /l̄a/</td>
</tr>
<tr>
<td>34 /l̄ʰa/</td>
</tr>
<tr>
<td>35 /a̯a/</td>
</tr>
<tr>
<td>36 /ra/</td>
</tr>
</tbody>
</table>

Among all these basic CV units according to *place of articulation* of sounds,
each consonants are divided into different categories. The sounds /ka/, /kha/, /ka/, /ga/, /gha/, /nga/ are grouped into the class velar, in the manner the sounds in this class are made by pressing the back of the tongue up against the velum (a movable muscular flap at the very back of the roof of the mouth). The sounds /cha/, /chcha/, /ja/, /jha/, /na/ are grouped into the palatal or palato-alveolar class. Here the palate rises sharply from the back of the alveolar ridge. The above sounds are produced with the blade of the tongue against this rising back of the alveolar ridge. The sounds /ta/, /tta/, /da/, /dda/, /na/ are in the retroflex class, which are produced by touching the underside of the tip of the tongue to the front part of the hard palate. /tha/, /tha/, /da/, /dha/, /na/ sounds are form the dental class in which these sounds are made by placing the tongue against the teeth. In sounds /pa/, /fa/, /ba/, /bha/, /ma/ the main restriction is formed by the two lips coming together and are named as labial sounds. Above discussed sounds are known as varga consonants in Malayalam, and other basic CV units grouped into other classes depending upon their context [Jurafsky and Martin(2004)].

According to manner of articulation of consonant units in Malayalam, Malayalam CV speech database can be divided into five different phonetic classes as unaspirated, aspirated, nasals, approximants and fricatives. In the aspirated class, the manner of articulation is the pronunciation of a stop with an audible and forceful release of breath, while in unaspirated class the sounds are produced without air inhalation. The unaspirated sounds include /ka/, /cha/, /ta/, /tha/, /pa/, /ga/, /ja/, /da/, /ba/ and aspirated sounds are /kha/, /g/gha/, /chcha/, /jha/, /tta/, /dda/, /ththa/, /dha/, /fa/,
The nasal sounds are produced by lowering the velum and allowing air to pass into the nasal cavity. The nasal sounds in Malayalam are /nga/, /na/, /nna/, /ma/. The sounds /ya/, /zha/, /va/, /la/, /lha/, /ra/, /v/ /zha/ are from the approximants class and in this class the two articulators are close together but not close enough to close turbulent airflow. In the class fricatives air flow is constricted but not cut off completely. The turbulent airflow that results from the constriction produces a hissing sound. The sounds in this class are /sha/, /shsha/, /sa/, /ha/ etc [Jurafsky and Martin(2004)].

3.4 Speech Data Acquisition

The inbuilt sound card coming with the personal computer when applied to external audio interfaces and necessary audio recording software generates digitized sound. To remove the phenomenon of aliasing in speech processing, the sound card is to be integrated with an antialiasing presampling Low-Pass Filter(LPF) appropriately. This thesis reports the data acquisition using the sound card incorporated with an 8th order butterworth LPF at the input. The performance of the system with such a modified sound card is comparable to the available high cost systems in the market [Prajith(2008)].

In the scientific literature it is already reported that additional analog elements are necessary prior to sampling to guarantee correct digital signal processing [Feuer and Goodwin(1996)] [B. Wittenmark and Arzen(2002)]. Anti aliasing plays a considerable part to make certain reliable and accurate frequency measurement in speech signal. According to sampling theorem, any bandlimited signal can
be accurately reconstructed exactly from values sampled at uniform intervals if it is sampled at a rate at least twice the highest frequency component present in the signal. Anti aliasing avoids higher frequency components from being sampled. An ideal anti alias low-pass filter passes all the appropriate input frequencies below the bandlimited frequency and cut off all the undesired frequencies above the bandlimited frequency. Such an anti aliasing filters are commonly built into the analog interface chips and codecs, which convert analog input signals into digital form for processing, by a digital signal processor.

Most of the sound blaster cards available are general purpose cards which has to handle both speech and music. This card consists generally of one or more A/D and D/A converters. They do not have front-end anti aliasing filter circuits to band limit the speech data. Thus these cards cannot be directly used for speech data acquisition. These cards can be converted into one suitable for speech processing by incorporating an anti aliasing filter before the analog to digital converter. We have constructed an 8th order Butterworth low-pass filter that can convert our multimedia system to low cost data acquisition system. We have constructed an 8th order low-pass anti aliasing filter by cascading four 2nd order filters, the circuit diagram of which is shown figure 3.4.1.
3.4.1 Speech processing system

The block diagram of a typical speech processing system is shown in figure 3.4.2. Before feeding to the analog-to-digital converter, the input speech signal is filtered with an 8th order Butterworth lowpass filter with a cut off frequency of 4 kHz, to remove all frequencies above the Nyquist frequency (since sampling frequency is 8 kHz). This is done to prevent aliasing during sampling, and is correspondingly called an anti alias filter. The system equipped with anti aliasing filter is now suitable for acquisition of speech data. We used the data acquisition system supported by Windows based multimedia system using GoldWave. GoldWave is a digital audio editor for Microsoft Windows. It has the advantage of sound editing, playing, recording and converting. The recording part of this software consists of option for selecting the sampling frequency and bits per sample for quantization. The sound files are recorded in the software as Windows wave format (.wav).
3.4.2 Speech database creation

The input speech signal is low-pass filtered to 4 kHz and is sampled at 8 kHz sampling rate and quantized with 16 bit A/D converter. The V/CV speech unit database is created using the above data acquisition system. We have used 50 vowel/Consonant-Vowels (V/CV) units as discussed earlier. The database consist of 14 long and short vowel units and 36 basic consonant-vowel units in Malayalam. Each V/CV speech unit is uttered by both male and female young native 96 Malayalam speakers of age between 21-35 years, is digitized and stored in separate data files. The file name of the data reveals the speech unit identity, speaker identity and repetition. The structure of the file name is as shown below.

MLXXZZZ

The first ML indicate language code and here ML is used for Malayalam. The second part XX indicates the V/CV speech unit identity number. For example, for the vowel æ u [u], this part is 05 and for œ ka[ka] this is 15. The third part, ZZZ indicates the speaker identity. For example, M01, F01 represent the first male speaker and first female speaker respectively.
As discussed above the developed database consist of 50 V/CV speech units including 14 long and short vowels and 36 consonants. Another recognition experiment is reported in this thesis using the same V/CV speech database by grouping into six different phonetic classes including vowels according to their manner of articulation of the consonants and is given in table 3.4.1. Generally the classification of phonetics depend upon two mode of characteristics namely place of articulation and manner of articulation. In the analysis, physical characteristics are crucial in manner of articulation rather than the place of articulation and this made the grouping for the present study. The wave form representation of the speech sound from six different phonetic classes in the created database are given in 3.4.3 to 3.4.8.

<table>
<thead>
<tr>
<th>Class</th>
<th>Sounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>/a/,</td>
</tr>
<tr>
<td></td>
<td>/aa/,</td>
</tr>
<tr>
<td></td>
<td>/e/,</td>
</tr>
<tr>
<td></td>
<td>/ee/,</td>
</tr>
<tr>
<td></td>
<td>/u/,</td>
</tr>
<tr>
<td></td>
<td>/uu/,</td>
</tr>
<tr>
<td></td>
<td>/r/,</td>
</tr>
<tr>
<td></td>
<td>/ae/,</td>
</tr>
<tr>
<td></td>
<td>/ai/,</td>
</tr>
<tr>
<td></td>
<td>/o/,</td>
</tr>
<tr>
<td></td>
<td>/oo/,</td>
</tr>
<tr>
<td></td>
<td>/au/,</td>
</tr>
<tr>
<td></td>
<td>/am/</td>
</tr>
<tr>
<td>Unaspirated</td>
<td>/ka/,</td>
</tr>
<tr>
<td></td>
<td>/ga/,</td>
</tr>
<tr>
<td></td>
<td>/cha/,</td>
</tr>
<tr>
<td></td>
<td>/ja/,</td>
</tr>
<tr>
<td></td>
<td>/ta/,</td>
</tr>
<tr>
<td></td>
<td>/da/,</td>
</tr>
<tr>
<td></td>
<td>/th/,</td>
</tr>
<tr>
<td></td>
<td>/da/,</td>
</tr>
<tr>
<td></td>
<td>/ba/</td>
</tr>
<tr>
<td>Aspirated</td>
<td>/kha/,</td>
</tr>
<tr>
<td></td>
<td>/gha/,</td>
</tr>
<tr>
<td></td>
<td>/chcha/,</td>
</tr>
<tr>
<td></td>
<td>/jha/,</td>
</tr>
<tr>
<td></td>
<td>/tta/,</td>
</tr>
<tr>
<td></td>
<td>/dda/,</td>
</tr>
<tr>
<td></td>
<td>/ththa/,</td>
</tr>
<tr>
<td></td>
<td>/dha/,</td>
</tr>
<tr>
<td></td>
<td>/pha/,</td>
</tr>
<tr>
<td></td>
<td>/bha/</td>
</tr>
<tr>
<td>Nasals</td>
<td>/nga/,</td>
</tr>
<tr>
<td></td>
<td>/na/,</td>
</tr>
<tr>
<td></td>
<td>/nna/,</td>
</tr>
<tr>
<td></td>
<td>/na/,</td>
</tr>
<tr>
<td></td>
<td>/ma/</td>
</tr>
<tr>
<td>Approximants</td>
<td>/ya/,</td>
</tr>
<tr>
<td></td>
<td>/zha/,</td>
</tr>
<tr>
<td></td>
<td>/va/,</td>
</tr>
<tr>
<td></td>
<td>/lha/,</td>
</tr>
<tr>
<td></td>
<td>/la/,</td>
</tr>
<tr>
<td></td>
<td>/ra/,</td>
</tr>
<tr>
<td></td>
<td>/rha/</td>
</tr>
<tr>
<td>Fricatives</td>
<td>/sha/,</td>
</tr>
<tr>
<td></td>
<td>/shha/,</td>
</tr>
<tr>
<td></td>
<td>/sa/,</td>
</tr>
<tr>
<td></td>
<td>/ha/</td>
</tr>
</tbody>
</table>
Figure 3.4.3: Speech Waveform of Vowel sound /aa/

Figure 3.4.4: Speech Waveform of Consonant sound /ka/ from the class Unaspirated
Figure 3.4.5: Speech Waveform of Consonant sound /kha/ from the class Aspirated

Figure 3.4.6: Speech Waveform of Consonant sound /nga/ from the class Nasals
Figure 3.4.7: Speech Waveform of Consonant sound /ya/ from the class Approximant

Figure 3.4.8: Speech Waveform of Consonant sound /sha/ from the class Fricative
3.5 Conclusion

This chapter focuses on an effort made to collect moderately large and representative samples of Malayalam V/CV speech units. A well structured and standard database of isolated Malayalam V/CV speech units, which facilitate the research in V/CV speech unit recognition is developed. Each speech unit is low-pass filtered to 4 kHz and sampled at 8 kHz sampling rate using 8th order Butterworth anti aliasing presampling filter is collected. The constructed database has the capability to be used as a bench marking resource for the Malayalam V/CV speech unit recognition research and this work is the first of its kind reported in Malayalam. Hence the developed database will help other researchers also to explore more recognition experiments for V/CV speech units.