Chapter 1

Introduction

We have not made our language ourselves, we have received it. We are what we are, by what those who came before us, have done for us. Like the coral islands which have been built up by the silent and self sacrificing industry of millions and millions of living beings, our languages have been elaborated by the incessant labors of millions and millions of those who came before us.

Max Muller

The words that we think, speak or write are not really ours. It is coming through us, but not ours. We have received it from our parents, teachers and others. They have received it from their parents and so on ad infinitum. Language is the most valuable single possession of human race. It is passed from generation to generation through cultural transmission. The society, culture, literature, art forms and every thing that human civilization possess, is because of language.

Speech and writing are the two main media through which language exist. Speech is the original medium for which human language evolved.

---

1Inspired by the words of Kahlil Gibran, in The Prophet: “Your children are not your children. They come through you but not from you”

2Another medium is sign language which is used primarily by those who cannot speak
Writing, the preservation of language messages using visible marks is a later development [3]. It is the most natural way of communication. When we become literate in a language, we learn to convert a written message (text) to spoken message (speech) and vice versa. We learn to read (text to speech conversion), understand and write (speech to text conversion).

The artificial techniques for converting text to speech and speech to text comes under the discipline of speech processing. Speech processing denotes the digital processing of speech signals for speech synthesis, speech understanding, speech recognition, speech coding and speech analysis. The text to speech synthesis systems (TTS) are aimed at generating speech, corresponding to the given input text. It is a system, that reads any given text aloud. The greatest challenge in a TTS system is the generation of natural sounding speech.

Speech is a social activity and involves direct communication between persons. Along with the literal meaning speech contains information about the speakers, their gender, emotion, background etc. When we convert the spoken message to text, only the literal meaning could be preserved. Other characteristic features of the speech like intonation, loudness, voice quality and timing will be lost. The variation in these features conveys, a lot more information than just literal meaning. When we convert a written message to speech, all these characteristic features have to be incorporated. Along with variation in linguistic contents, the acoustic parameters like duration and fundamental frequency of speech units are dynamically varied by the speakers. In the absence of these variations, the synthesized speech will sound robotic.

The variation in duration and fundamental frequency is manifested as the
rhythm or prosody of the speech. The prosodic features of a language has to be studied and speech units have to be modified accordingly, to produce natural and intelligible speech in that language. Since prosody is language specific, extensive studies have to be performed for each language. Models have to be developed, to catch the variation of these acoustic parameters, which results in the naturalness of the speech. The duration model predicts the duration of speech segments to be used for synthesis. The duration of speech segments in natural speech is analyzed, to find the factors that affect duration. Duration models are developed based on this analysis.

In this work, the duration analysis and modeling is done for Malayalam language. Malayalam is one of the 22 scheduled languages in India\(^3\), used predominantly in Kerala, Lakshadweep and Mahe (Mayyazhi). The total number of speakers is around 33 million in India (as per 2001 census). The durational patterns of Malayalam speech are analyzed using statistical tools and duration models are developed. The duration analysis and modeling of Malayalam speech has not so far been reported by any previous work.

The statistical analysis of duration values are done, based on which duration models were developed. The investigations were primarily oriented to develop novel methods for duration modeling, which can be adapted for modeling duration patterns of other languages as well. The performance of the duration models are evaluated and compared using objective as well as subjective measurements.

\(^3\)Map on Indian language families compiled by Central Institute of Indian Languages (CIIL) is given in Appendix
1.1 Language and Speech

1.1.1 Language

The essential difference between man and animal is exhibited most clearly by human language, in particular by man’s ability to form new sentences which express new thoughts and which are appropriate to new situations. The ordinary use of language is evidence of tremendous creative potential in every human being.

Noam Chomsky

Language is the means by which human beings communicate with each other and with the society. Many other creatures are also able to communicate with other members of their species. But there are certain unique properties for human language, including displacement, arbitrariness, productivity, cultural transmission and duality [4]. These features make the human language creative and provides infinite possibilities of expression.

The most remarkable feature of language is the way in which a child acquires first language within the age of 3 or 4 years. By that time almost all children are able to form, the structure of the language being spoken, in

4Reference to past and future time, and to other locations. It allows us to talk about things, events and places not present in the immediate environment and whose existence we cannot even be sure of. This contributes to the ability of human to create fiction.

5Arbitrary relationship of the linguistic signs with the objects referred. For example there is no relation between the linguistic form ‘cat’ and the animal referred by it. Majority of animal signal appear to have a clear connection between the conveyed message and the signal. Because of this nonarbitrariness, the set of signals used in communication is finite.

6Ability to produce new expressions and new sentences. Language users manipulate their linguistic resources to produce new expressions and new sentences. Novel utterances are continually being created and the potential number of utterances in any human language is infinite.

7Process by which language is passed on from one generation to the other. Due to this language, literature and there by culture gets enriched with each generation.

8Language is organized in two levels or layers simultaneously - at one level distinct sounds and at another level distinct meanings (with individual sounds ‘n’, ‘b’ and ‘i’, we can have two meaningful words, nib and bin). Due to this duality of levels, with a limited set of distinct sounds, we can produce a very large number of sound combinations.
the community where they live. Unlike other skills which each of us learn to varying degrees, every person masters their first language and generally can speak in that language with considerable proficiency, without making grammatical mistakes. According to Noam Chomsky, there is a language faculty that is, there is some part of the mind-brain, which is dedicated to the knowledge of and use of language [5]. It is a kind of language organ, roughly analogous to the visual system. The first language is not actually 'learned', it is acquired. Language acquisition is more like a growth process with the right simulation around, similar to the development of visual system. During language acquisition, the structure and grammar of the language grows inside the mind of a child.

Language exist through the media like speech, writing and signs. The patterns formed in these media carry the language. The spoken message and written message carry the same information embodied in different media. Speech carries information as variation in acoustic energy with respect to time, and writing, as patterns in white space. The language has interactional (use of language for interaction) as well as transactional (use of language to communicate knowledge, skills and information) function. Language developed initially through the medium of speech, primarily for interaction. Later, writing developed through cave markings, inscriptions and alphabetic writing, perhaps mostly for the need of transaction of knowledge.
1.1.2 Speech sounds, phonetics and phonology

All natural languages invariably use the medium of speech\(^9\). It is the basis on which all other media are constituted \([6]\). In spoken language, different sounds are formed in succession to convey messages, organize activities, express thoughts, ideas, dreams, emotions, desires and fantasies. Sound is produced by the air stream coming out or into our lungs modified by the conscious movement of respiratory organs and the organs in the mouth. Human vocal apparatus can produce a large variety of sounds \([7]\). Of these only a finite subset is used in speech production. For each language there is a finite set of sound patterns which is linguistically significant. The general study of the characteristics of speech sounds is called phonetics \([4]\).

Phonetics

*The study of the medium of spoken language, in all its aspects and all its varieties constitute the subject of phonetics. It is concerned with the medium as used in speaking all human languages and as used in all styles of speech.*

*David Abercrombie*

Phonetics deals with the analysis of the speech sounds in all human languages. It is not language specific and its principles and concepts are applicable to any language. The central concerns in phonetics are the studies on production or articulation of speech sounds (articulatory phonetics), physical properties of speech as sound waves in the air (acoustic phonetics) and perception of different sounds via the ear (auditory phonetics).

\(^9\)There are many languages around the world which have no written form (eg. Konkani, an Indian language). For those languages, speech is the only medium through which they exist.
In most of the languages, speech sounds are formed utilizing the air stream expelled from the lungs, called pulmonic egressive airstream (pulmonic - lungs, egressive - outgoing). In some languages of southern Africa [7], the sounds produced by incoming breath (ingressive airstream) is also used for speech. The air from the lungs passes through the trachea, larynx and pharynx and moves out through mouth. The air arriving at the pharynx will pass through nasal cavity also, if the velum or soft palate is open. If nasal cavity is also used in speech production, the generated sounds are called nasal sounds.

The larynx or voice box has a pair muscular tissue called vocal folds, which can be moved to different positions. The vocal folds remain wide open while we are breathing and not talking. When the vocal folds are close together, the air from the lungs repeatedly pushes them apart as it passes through, creating a vibration effect. These sounds are called voiced sounds. When the gap between the vocal folds are narrow, air passes with a rushing noise producing sounds which are called voiceless sounds. Since the voiced sounds involve vibration of vocal folds, the speech signal produced by them will be periodic/quasi periodic while the waveform of unvoiced or voiceless sounds will be like random noise.

The passageway from larynx to lips is the vocal tract where the air stream gets modified due to the movement of the articulators like tongue, lower jaw and lips (active articulators). The active articulators along with the passive articulators like hard palate, upper teeth, gumridge and upper lips changes the shape of the oral tract, producing the various sounds. If the articulators

\footnote{‘Click’ sounds produced by ingressive air stream are used as speech sounds in the African languages Zulu and Xhosa.}
obstruct or close the vocal tract, the sounds produced are called consonants, while sounds produced with a relatively free flow of air is called vowels.

The different vowels are produced by changing the shape of the vocal cavity using lips and tongue. The vowel sounds are all voiced. The tongue height (high/close, mid, low/open), tongue advancement (front, central, back) and lip rounding can be used to describe the vowels. In producing vowels, the vocal organs remain relatively steady. Diphthongs are combined vowel sounds, wherein the vocal organs are moved from one vocalic position to another (eg. [ay] as in ‘buy’; [aw] as in ‘cow’). Consonants can be classified according to manner of articulation (manner in which air stream is interfered with), place or point of articulation (point in the vocal tract at which the obstruction is made) and voicing (presence or absence of the vibration of vocal folds).

According to manner of articulation, the consonants are grouped into:

- **Stops**: Produced by complete stopping of the air stream and letting it go abruptly, eg: [b] as in ‘bag’

- **Fricatives**: Produced by blocking the airstream and pushing the air through narrow opening, eg: [s] as in ‘sing’

- **Nasals**: Produced with airstream flowing through both oral and nasal cavity, eg: [m] as in ‘morning’

- **Approximants**: Also called as glides or semivowels are vowel like sounds and are strongly influenced by the following vowel, eg: [w],[y]

The different consonants are produced according to the place inside the mouth, where the air stream is obstructed due to the movement of differ-
ent articulators. According to the place of articulation, the consonants are divided into:

- **Bilabials**: Formed by using both (bi) upper and lower lips (labia) for constriction, eg: [p] as in ‘pat’

- **Labiodentals**: Formed with obstruction caused using upper teeth and the lower lip, eg: [f] as in ‘father’

- **Dentals**: Formed with the tongue tip touching behind the upper front teeth, eg: [θ] as in ‘thin’

- **Alveolars**: Formed with the front part of the tongue on the alveolar ridge, eg: [t] as in ‘top’

- **Alveopalatals**: Produced with the tongue at the very front of the palate, near the alveolar ridge eg:[s] as in ‘she’

- **Velars**: Produced with back of the tongue against the velum, eg: [k] as in ‘kid’

- **Glottals**: Produced without active use of tongue and other parts of the mouth, eg: [h] as in ‘house’

The consonants can be voiced or unvoiced, depending on whether the vocal folds vibrate or not during speech production.

The airstream coming out of the lungs is not in a continuous flow, but in a pulse-like manner resulting from the alternate contractions and relaxations of the breathing muscles. The muscular contraction and the consequent rise in air pressure result in a chest pulse and the speech segment emerging out of each chest pulse is called a syllable [7]. Words are formed with one or more
syllables. For example the word /signal/ is formed of two syllables /sig/ and /nal/. A vowel alone or vowel in combination with one or more consonants constitute a syllable.

Phonology

*Human beings perceive reality by fitting the external stimuli into patterns that exist in the mind, which involves ignoring some of the actual differences and inventing differences which do not exist.*

*K.P. Mohanan*

Each language uses a subset of the speech sounds that can be produced by human vocal apparatus. The different languages select different sets of speech sounds and organize them into various patterns for expressing meanings. The selection of speech sounds and their organization in a particular language constitute the phonology of that language. Phonology is the description of systems and patterns of speech sounds in a language [4]. It is the study of how sounds are organized and used in each language. The phonological system of a language includes the inventory of sounds and the rules which specify how the sounds interact with each other.

When we speak, our vocal apparatus functions normally continuously in motion moving from one sound to next without stopping in between. Hence the acoustical signal generated is continuous in time. But we conceive it as a sequence of discrete entities. Each utterance can be considered as a sequence of discrete segments and can be represented using a sequence of discrete symbols. This is one of the fundamental assumptions in phonology. The smallest identifiable unit found in a stream of speech is called a phone. The phones can be represented using discrete symbols. The internationally
accepted way of representing phones is using IPA symbols (Figure 1.1.2). IPA is an evolving standard originally developed by the International phonetic association in 1888 for representing or transcribing the sounds of all human languages [7].

Every individual has a physically different vocal apparatus and hence every individual produce sounds differently. In addition, each individual will not be producing sound physically identical on all occasions [4]. All these different physical sounds, corresponding to a given phonetic symbol are conceived as the same phone by the listener. Phonetic representation is the abstraction of the physical sounds produced. The speech sounds or phones used in each language is further grouped into a set of distinctive functional units called phonemes. The function of these sounds is to distinguish meaning of one word from another. For example in the words /tar/, /far/ and /bar/ the meaning is distinguished by the phonemes /t/, /f/ and /b/.

In continuous speech the articulators are moved from one sound to next without stopping. The phonemes undergo changes due to this process of making one sound, before the stopping of the previous sound. This process called co-articulation is mainly manifested as assimilation and elison [4]. Assimilation is the process by which some aspect of one phoneme is taken or ‘copied’ by the other. The omission of a sound segment which would be present in the deliberate pronunciation of a word in isolation is called elision. The phonological rules specify how the phonemes get altered in different phonetic environments.

The phonemes are abstraction of the phones produced in speech. The phoneme is more abstract than phone and is actually the mental picture of the actual sounds produced. The set of phones that are versions of one
THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

<table>
<thead>
<tr>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Postalveolar</th>
<th>Retrareflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>c</td>
<td>j</td>
<td>k</td>
<td>g</td>
<td>g</td>
<td>?</td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>m</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td>r</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Tap or Flap</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
<td>V'</td>
</tr>
<tr>
<td>Fricative</td>
<td>f</td>
<td>v</td>
<td>θ</td>
<td>θ</td>
<td>s</td>
<td>z</td>
<td>s</td>
<td>z</td>
<td>z</td>
<td>z</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>Approximant</td>
<td>u</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
</tbody>
</table>

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

<table>
<thead>
<tr>
<th>Clicks</th>
<th>Voiced implosives</th>
<th>Explosives</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>Bilabial</td>
<td>B</td>
</tr>
<tr>
<td>L</td>
<td>Dental</td>
<td>D</td>
</tr>
<tr>
<td>(Postalveolar)</td>
<td>F</td>
<td>Palatal</td>
</tr>
<tr>
<td>#</td>
<td>Alveolar lateral</td>
<td>G</td>
</tr>
</tbody>
</table>

OTHER SYMBOLS

<table>
<thead>
<tr>
<th>Voiced labial-velar fricative</th>
<th>Voiced alveolar-velar fricative</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Z</td>
</tr>
</tbody>
</table>

Voiced labial-velar approximant | Voiced alveolar-velar approximant |
| W                             | H                               |

Voiced epiglottal fricative | Voiced epiglottal fricative |
| C                             | ?                               |

Diacritics may be placed above a symbol with a descender, e.g. ɪ.

VOWELS

<table>
<thead>
<tr>
<th>Diphthongs</th>
<th>Monophthongs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unitones</td>
<td>Multitones</td>
</tr>
</tbody>
</table>

TONES AND WORD ACCENTS

LEVEL

<table>
<thead>
<tr>
<th>TONE</th>
<th>CONTOUR</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra high</td>
<td>Falling</td>
<td>e or A</td>
</tr>
<tr>
<td>High</td>
<td>Falling</td>
<td>e or A</td>
</tr>
<tr>
<td>Mid</td>
<td>Falling</td>
<td>e or A</td>
</tr>
<tr>
<td>Low</td>
<td>Rising</td>
<td>e</td>
</tr>
<tr>
<td>Extra low</td>
<td>Falling</td>
<td>e</td>
</tr>
</tbody>
</table>

SUPRASEGMENTALS

<table>
<thead>
<tr>
<th>PRIMARY STRESS</th>
<th>SECONDARY STRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long e</td>
<td>e</td>
</tr>
<tr>
<td>Half-long e</td>
<td>e</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SYLLABLE BREAK</th>
<th>INTONATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linking (absence of a break)</td>
<td>e</td>
</tr>
</tbody>
</table>

Figure 1.1: Table of the phonemic alphabets
phoneme is referred to as allophones of that phoneme. For example, the phoneme /i/ is nasalized in 'seen' due to the effect of the nasal /n/. The nasalized sound [ɪ] and [i] sound in 'seed' are allophones of the phoneme /i/.

Phonemes are language specific. Phonology analyzes the sound patterns of a particular language by determining which phonetic sounds are significant and by explaining how these sounds are interpreted by the native speaker. It is based on a theory that every speaker of a language knows unconsciously about the sound patterns of that language [4]. It is concerned with the abstract or mental aspect of the sounds in language rather than with the actual physical articulation of speech sounds.

1.2 Malayalam Language

_Speech is a human activity that varies without assignable limit as we pass from social group to social group, because it is a purely historic heritage of the group, the product of long continued social usage._

Edward Sapir

Malayalam belongs to Dravidian family of languages along with Tamil, Kannada and Telugu. Malayalam has been considered as an offshoot of classical Tamil. Another theory regarding the origin of the language assumes that the split occurred in prehistoric period with the two languages, developing out of 'Proto-Tamil-Dravidian' - the common stock of Tamil and Malayalam. Tamil being similar in grammar and vocabulary is considered as the closest relative of Malayalam. Malayalam has heavily borrowed from Sanskrit in terms of vocabulary, syntax and phonology and the borrowings have be-
come a major part of the core of the language. This has resulted in a close blend of Dravidian and Sanskrit elements in the language. Malayalam has retained the Dravidian phonological and grammatical features. Through the borrowed lexical items from Sanskrit, it expanded its phonology by incorporating sounds like aspirated plosives and fricatives (not present in Dravidian languages like Tamil). The dual identity of Dravidian (phonological and grammatical) and Sanskrit (lexical) created a peculiar place for Malayalam among Dravidian languages.

Most of the Indian languages share a common phonetic base. There is one-to-one correspondence between orthographic symbols and phonemes, unlike languages like English. The basic linguistic sound called Varnam, correspond to the basic sound unit, phoneme. Varnams are divided into Swarams (corresponding to vowels) and Vyanjanams (corresponding to consonants). The Swarams can be uttered on their own. Vyanjanams are rendered with the help of vowel phonemes. The Swaram either with one or two Vyanjanams or by itself forms a syllable.

Like most other Indian languages, Malayalam is syllabic in nature and syllables are read as units. There is direct correspondence between written and spoken syllable. Each syllable has a unique orthographic symbol. Malayalam script has letters corresponding to all the sounds in Sanskrit as well as letters for Dravidian specific sounds. The script consists of 51 letters

11 Malayalam has also taken vocabulary from languages like English, Persian, Hindi, Arabic, Syriac, Urdu, Pali, Prakrit, Portuguese and Dutch.
12 Varna, Swara and Vyanjana in Sanskrit
13 In languages like English, written syllable need not always invariantly correspond to a spoken syllable. For example the syllable represented by the letter combination 'cha', may correspond to [ka] sound or [cha] sound according to the context in which it occur.
14 One exception is the the common orthographic symbol used for dental nasal and alveolar nasal.
Malayalam alphabet

Vowels (svaram)

Vowel diacritics with ka

Consonants (vyanjanam)

A selection of conjunct consonants

Figure 1.2: Malayalam Alphabet
including 13 vowels and 38 consonants (Figure 1.2). Each vowel is represented by separate letter. When combined with consonants to form syllables, diacritics are used to modify the orthographic symbols of the corresponding consonant. All consonants are represented by letters which has an inherent vowel /a/. The vowel /a/ can be suppressed using the diacritic 'v', called chandrakkala. When diacritics corresponding to other vowels are added, /a/ gets suppressed and the orthographic symbol represent the syllable corresponding to the consonant combined with that vowel (Examples are given in Malayalam phonemic chart).

Consonant clusters are formed when two consonants of same type or different combine together. If the consonants are of the same type the combination is called geminate of the corresponding consonant. All consonant clusters have separate orthographic symbols which combines the features of the constituent phonemes. Malayalam words do not end in consonants, except for [m] and [n] in casual speech and [m,n,ŋ,l,r] in careful or literary speech.

1.3 Text to speech synthesis

There was geometry in the world before Newton, and philosophy before Decartes, but before language there was nothing but bodies and their images, because language is the necessary instrument of every intellectual operation - nay the means of every moral existence.

De Bonald
The artificial speech generation systems have a long history, starting from the mechanical synthesizers in 18th and 19th century\textsuperscript{15} to electrical synthesizers in early 20th century\textsuperscript{16}. The earlier systems can be generally considered as speaking machines and the intelligibility and naturalness was very poor. The full text to speech synthesis systems (generation of speech for any given input) were developed in the later half of 20th century\textsuperscript{17}.

The text to speech synthesis system operates in two phases, the first one is text analysis, where the input text is transcribed into a phonetic representation, and the second one is the generation of speech waveforms, where the acoustic output is produced from this. The input text may be data from a word processor, standard ASCII from e-mail, a mobile text-message, or scanned text from a newspaper. The character string is then preprocessed and analyzed to obtain the phonetic representation which is usually a string of phonemes. The text analysis module is a natural language processing module and is developed based on linguistic and phonological knowledge of the language. The speech generation module takes in the string of phonemes and generates the corresponding speech. The speech is generated, after applying the necessary signal processing to the acoustic parameters, for the manipulation of prosody, so as to result in natural sounding speech. The methods to generate speech can be generally classified into articulatory synthesis, formant synthesis and concatenative synthesis.

\textsuperscript{15}In 1779 Russian Professor Christian Kratzenstein constructed acoustic resonators similar to the human vocal tract to produce five long vowels.
\textsuperscript{16}VODER (Voice Operating Demonstrator) introduced by Homer Dudley in New York World’s Fair 1939 consisted of wrist bar for selecting a voicing or noise source and a foot pedal to control the fundamental frequency. The source signal was routed through ten bandpass filters whose output levels were controlled by fingers.
\textsuperscript{17}The first full text-to-speech system for English was developed in the Electrotechnical Laboratory, Japan 1968 by Noriko Umeda. In 1979 Allen, Hunnicutt, and Klatt demonstrated the MITalk laboratory text-to-speech system developed at M.I.T.
• **Articulatory synthesis:** Articulatory synthesis typically involves models of the human articulatory organs and vocal folds. The articulators are usually modeled with a set of area functions between glottis and mouth [22]. The first articulatory model was based on a table of vocal tract area functions from larynx to lips for each phonetic segment [8]. The articulatory control parameters may be for example lip aperture, lip protrusion, tongue tip height, tongue tip position, tongue height, tongue position and velic aperture.

• **Formant synthesis:** Models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model [22]. A cascade formant synthesizer consists of band-pass resonators connected in series and the output of each formant resonator is applied to the input of the following one. A parallel formant synthesizer consists of resonators connected in parallel. Sometimes extra resonators for nasals are used. The excitation signal is applied to all formants simultaneously and their outputs are summed.

• **Concatenative synthesis:** Uses different length prerecorded samples derived from natural speech. Connecting prerecorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech [22]. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade-off between longer and shorter units. With longer units high naturalness, less concatenation points and good
control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. In present systems units used are usually words, syllables, demisyllables\textsuperscript{18}, phonemes, diphones\textsuperscript{19}, and triphones\textsuperscript{20}.

The formant and concatenative methods are most commonly used in the present synthesis systems. The formant synthesis was dominant for long time, but today the concatenative method is most commonly used [22]. The articulatory method is still too complicated for high quality implementations, but may arise as a potential method in the future.

1.4 Duration analysis and modelling

_{Language is peculiarly human and it is found deep inside the mind_}

_{Stanford A. Schane_}

In concatenative synthesis which produces most natural speech, the speech segments corresponding to the given text, are taken from the recorded data base and concatenated together after necessary modification. It is essential to conduct detailed analysis and modelling of duration patterns of speech units in continuous speech, inorder to generate highly intelligible and natural

\textsuperscript{18}Demisyllables represents the initial and final parts of syllables

\textsuperscript{19}Diphones (or dyads) are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain the transitions between adjacent phones.

\textsuperscript{20}Triphones are like diphones, but contains one phoneme between steady-state points (half phoneme - phoneme - half phoneme). In other words, a triphone is a phoneme with a specific left and right context.
sounding speech. The duration models predict the duration of each segment and the duration of the units are modified accordingly before concatenation.

The duration modeling can be broadly classified into rule based approaches and statistical approaches. Rule based approach is based on the results of experimental studies on segment durations. The most prevalent rule based duration model is a sequential rule based system proposed by Klatt [9]. Klatts model is usually presented in the form of rules that are applied successively, starting with an initial segment duration.

Statistical approaches comprise parametric (sum-of-products model) and non parametric regression (Classification and Regression Trees (CART), neural network) models. In parametric regression model, the structure of processing the input parameters is determined a priori. In sum-of-products model the segment duration is represented as a sum of factors and their product terms that affect the segment duration. Non parametric regression models are developed by unsupervised training, and the structure of the model is determined automatically.

In duration models the speech segment is represented by an n dimensional feature vector. The feature vector is formed of elements which correspond to the different factors that affect duration. The feature vector is computed from text. The task of the duration component of a TTS is to predict the duration of every segment depending on its feature vector. It is essential to perform detailed analysis on duration patterns of speech units in continuous speech, in order to find factors which affect this duration.

The duration models are developed, based on detailed duration analysis on natural speech. The duration models infact maps the string of phonemes

\[\text{Detailed description in chapter 2}\]
and syllables to duration values. This mapping is generally considered as a complex nonlinear task [133].

Duration analysis and modelling has been reported for many languages including English [8, 9], French, German, Japanese [26], Czech [39], Korean [18], Spanish and Turkish [23]. In Indian Languages, significant work has been done for duration analysis and modeling in Hindi, Telugu and Tamil [10, 11, 135, 113]. K. Sreenivasa Rao and B.Yegnanarayana have modeled the duration of Tamil, Hindi and Telugu using artificial neural networks [133]. Another work on duration modeling for Hindi TTS is reported by N. Sridhar Krishna et al [136]. The paper by Sridhar Krishna and Hema A. Murthy presents a preliminary attempt on data-driven modeling of segmental (phoneme) duration using Classification and Regression Tree (CART) for Hindi and Telugu [13].

1.4.1 Duration analysis and modelling for Malayalam

Since the duration patterns are language specific, detailed durational analysis on natural speech has to be performed for each language, based on which suitable models can be built. The objective of the thesis is to study the duration patterns of speech segments in Malayalam and to derive models to predict duration.

Duration analysis is performed on recorded database taken from Malayalam news of Doordarshan TV Channel22. The durational patterns of the phonemes in Malayalam news is first analyzed using statistical tools like

---

22Doordarshan is the public television broadcaster of India and is a division of Prasar Bharati, a public service broadcaster nominated by the Government of India. It is one of the largest broadcasting organizations in the world in terms of the infrastructure of studios and transmitters.

---
boxplots, distribution fitting, Q-Q plots, multiple comparison, confidence interval and ANOVA. Statistical analysis is carried out to find the different factors affecting duration, as well as to study the effect of the different factors on the variation of duration values. The feature vector is formed based on the analysis, combining the different factors.

The feature vector framed is used for training the duration models developed. The first part of the duration modeling presents two models, i) probabilistic model ii) hybrid model combining CART and HMM, which are new in the literature. These models follow the conventional approach of considering each speech unit separately, analyzing the factors affecting duration and framing models based on this.

The second part of modeling, introduces a new model named memory based duration model, which differs from the traditional models, in terms of the basic approach in modeling. The basic premise on which memory based model is built, is the observation that the sentences, phrases and words have rhythm as a whole. The rhythm is not imposed on each phoneme or syllable.

In traditional models, it is assumed that, speech is a linear combination of segments occurring one after other and that the duration patterns can be predicted using factors determined from text. But in continuous speech, the speech segments interact with each other and hence the prediction of duration patterns is a complex nonlinear task. The durational patterns of each language is different and even in one language, the patterns vary for different styles, which means that textual information is not sufficient to produce natural sounding speech. Hence it is very difficult for the duration model to capture the duration patterns of a language for a particular style, beyond a certain limit. At the same time, human brain is able to generate
any speech construct in a particular style, if we have sufficient exposure to it.

The durational patterns can be captured, if we mimic the way human brain acquires different speaking styles. The new duration model is developed, using the basic concepts of memory prediction framework and exemplar theory. The memory prediction framework explains the hierarchal structure of human brain and cognition. The human neocortex has a hierarchal structure where each layer stores different speech constructs. The memory prediction framework is expected to be the direction in which speech processing research will proceed in the coming decades [152]. The exemplar theory is a psychological model of perception and categorization. According to exemplar theory, the human brain perceives and reproduces speech constructs, by retrieving the stored exemplars of previously perceived examples. A new speech construct, which is not already stored, will be produced analogous to the most similar exemplar stored in memory.

The model is further supported by analogical theory and Zipf's law from phonology. The analogical theory explains how we create new utterances, based on what we have already perceived and cognized [153, 154]. Zipf's law deals with the nonuniform distribution of words and phrases in natural languages [156, 157]. These two theories, suggest that the rhythm of any particular style can be captured, by storing the prosodic parameters of most commonly occurring phrases and words.

The new duration model captures the duration patterns by storing the durational variation of most commonly occurring syllables, morphemes, words and phrases. The model predicts duration by retrieving the stored pattern for each speech construct. The memory prediction framework or exemplar
theory has not been previously used for duration modeling.

The new model named as memory based duration model differs from the traditional methods, in terms of the basic approach in modeling. In traditional methods, each speech segment is analyzed separately, factors affecting duration are determined and models are framed based on this analysis. The basic premise on which memory based model is built, is the observation that the sentences, phrases and words have rhythm as a whole. The rhythm is not imposed on each phoneme or syllable.

The performance of the duration models are evaluated objectively and perceptually. The different models are compared objectively by calculating the root mean squared error (RMSE) and correlation. The mean opinion score is used for perceptual evaluation of the model.