# CHAPTER 1

# INTRODUCTION

This chapter emphasizes the significance and characterization of a Text to Speech Synthesis System. We have presented a brief review of literature about the present work. The chapter gives the terminology, overview and general dataflow of a Text to Speech Synthesis system. It also deals with the syllabification rules of the Tamil language, which throws some light upon the understanding of the thesis to a complete extent.

## 1.1 TEXT TO SPEECH SYNTHESIS

One of the best media of communication until years had been the print media, as the reach and circulation of information is a defined and consistent one. The print media has registered its significance in almost all fields. More business is performed using this media; lots of researches are conducted in this area for the development of such a form of communication. Though there is an ample range of users for the written form of communication, people feel it simple and more convinced when the written form is in audio format. In this case, there arose a necessity for a person or a machine that can read out a given arbitrary text.

Today Artificial Intelligence plays a vital role in almost any field that mankind involves. Computational Linguistics is one of the wide sub-areas in Artificial Intelligence. Computational linguistics is an interdisciplinary field dealing with the statistical or rule-based modeling of natural language from a computational perspective. It deals with the processing of Natural Languages. Natural Language Processing involves the processing of natural language text such as information extraction, translation into another language, providing domain specific transcriptions, etc. There

have arisen needs for the conversion of Natural Language text into Natural Language Speech, which included applications such as assisting blind people, native speakers but nonreaders, people who are non-native to the particular language, etc. This phenomenon led to the evolution of a Text to Speech synthesis System. Speech Synthesis is the artificial production of human speech out of a given text. A computer system used for this purpose is called a speech computer or a speech synthesizer, otherwise the 'Text To Speech Synthesis (TTS) System.' A text-to-speech synthesis system converts the normal language text into speech. There are even other systems which can render symbolic linguistic representations like phonetic transcriptions into speech.

### 1.1.1 Terminology

Morpheme: A Morpheme is the smallest meaningful unit of a language. A morpheme may be a stand-alone one (Generally called as a root morpheme) or an affix. Ex: The word 'successful' - Here 'success' is the root morpheme, and 'ful' is an affix.

Allomorph: An allomorph is one of two or more complementary speech units which manifest a morpheme in its different phonological or morphological environments.

Grapheme: A Grapheme is described as a letter or a string of letters of an alphabet, which represents a phoneme. Ex: For the phoneme /s/, the graphemes are's', 'es', and 'ce'.

Phoneme: A phoneme or a phone is the basic acoustic unit of a language. A phone is always a simple unit of sound, which cannot be subdivided. There are about 31 phones in English and 41 phones in the Tamil language, including the allophones.

Diphones and triphones: A diphone is a concatenation of two phones; a triphone is that of three phones. Diphones are the permutations of each phone with all other phones of the language. Triphones are the permutations of each phone with two other phones. So totally there are ${}_{41}P_2$ = 1640 diphones and ${}_{41}P_3$ = 63960 triphones in the Tamil Language.

Syllable: A syllable is a unit consisting of unbroken sound that can be used to make up words. A syllable is a prearranged unit of sound. It is a composition of phonemes. Moreover, syllables seem to act as a fundamental unit for many aspects of Prosody. Each syllable constitutes two parts: Onset and rhyme. Rhyme consists of two parts – the nucleus and the coda. Figure 1.1 shows the structure of a syllable.

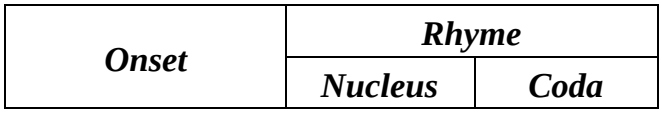


**Figure 1.1 Structure of a syllable**

The nucleus is the essential part of a syllable.

Syllabification of an English word *'/window/'* is as follows:

*'/win/'and '/dow/'*

'Window' is a two-syllable word.

The phonetic representation of the above syllables is given as: *[wIn], [dO]*. Here, in the syllable *[wIn],*

- Onset: *[w]*
- Rhyme: *[In]*
  - Nucleus: *[I]*

- Coda: *[n]*

In the syllable *[dO]*,

- Onset: *[d]*
- Rhyme: *[O]*
  - Nucleus: *[O]*

There is no coda in the second syllable.

Polysyllable: A polysyllable is a combination of syllables. A polysyllable is recursively an organized combination of syllables. Each language has its own rules for the combination of syllables.

Speech Corpus: A Speech Corpus is a physical component that is used to store the mapping between text units and their pronunciations. i.e., if it is a phone corpus, then it will consist of the letter to sound rules.

Prosody: Prosody is a set of important elements that provide naturalness to the processed speech. A prosodically good speech is a more human-like one.

Intonation: Intonation is the variation of pitch exhibited for a single word, so as to express the emotion of the speaker.

Stress: Stress is the relative emphasis given to a certain part of a sentence or a word or a syllable.

Phrasing: A phrase is a single word or a group of words that constitutes the syntax and semantics of a sentence.

Intensity: Intensity is the quality of language that indicates the degree to which the speaker's attitude toward a concept deviates from neutrality.

Parsing: Parsing is the process of splitting down a sentence or a phrase into its grammatical components. It is a pre-process of ‘tagging’, otherwise called ‘annotation.’

Lexicon: The lexicon is a catalog of words and terms. It consists of words (root morphemes) and bound morphemes (affixes).

### 1.1.2 An Overview of a Text to Speech Synthesis System

A Text to Speech Synthesis system or a ‘Speech Synthesizer’ is a system that converts a given text into its audible form. An arbitrary text is converted into continuous words, which is called a ‘running’ speech. In simple words, Text to Speech Synthesis system is the convergence of Linguistics and Signal Processing, as shown in Figure 1.2. The given text has to be processed using the concepts of Linguistics. Appropriate acoustic signals are produced using Signal Processing paradigm.

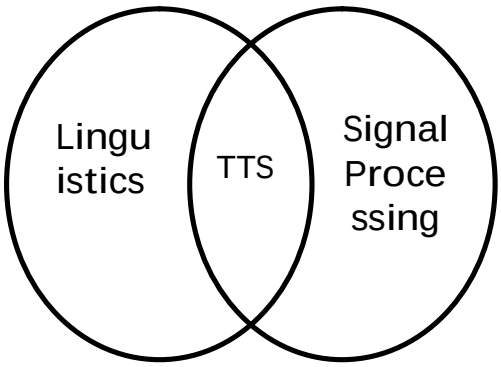


**Figure 1.2 Two main components of TTS**

As shown in the Fig. 1.3, the given text is analyzed in the linguistic aspects, regarding syntax and semantics of the language. Text processing is

carried out wherein the analyzed text is parsed into separate grammatical units, such as sentences, phrases, words, and morphemes. Some initial level tagging is performed in these units, and they are sent to Digital Signal Processing phase. In the Digital Signal Processing stage, appropriate sound signals are produced for each stream of characters (viz., words, and morphemes). The design of a typical TTS system involved in including a pronunciation dictionary in the Digital Signal Processing phase.

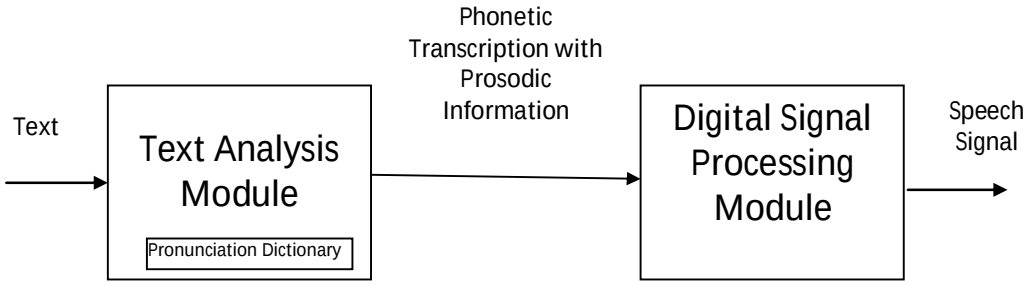


**Figure 1.3 Block diagram of a TTS system**

A pronunciation dictionary includes all possible pronunciations to a single morpheme, including allomorphs. In later course, a Text to Speech Synthesis system had to maintain a dedicated pronunciation dictionary called a Speech Corpus. This speech corpus was of various types such as a word corpus, phone corpus, diphone corpus, syllable corpus, etc.

As shown in the Fig. 1.5, Natural language processing (NLP) is a field concerned with the interactions between computers and human languages. It is one of the sub areas of Computational Linguistics. It has two main categories of research, namely Natural Language Understanding (NLU) and Natural Language Generation (NLG). Fig 1.4 describes the flow of data in a TTS system.

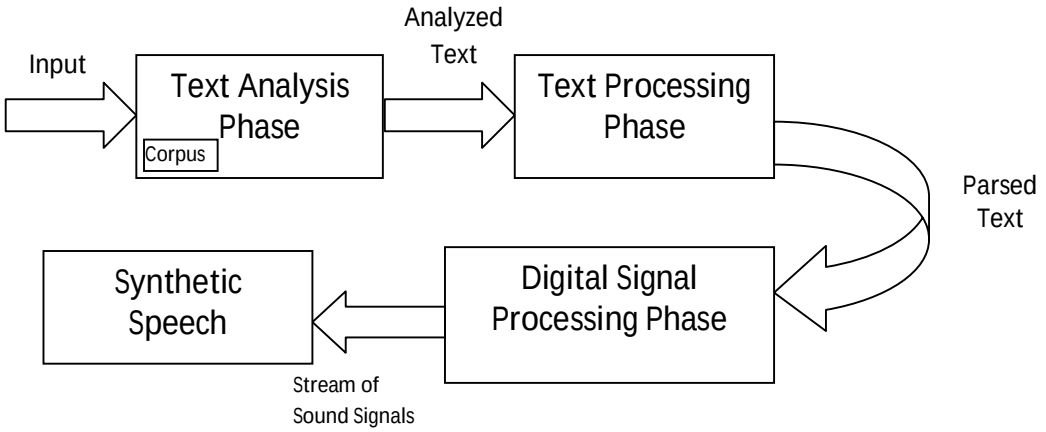


**Figure 1.4 A Typical Text to Speech Synthesis System**

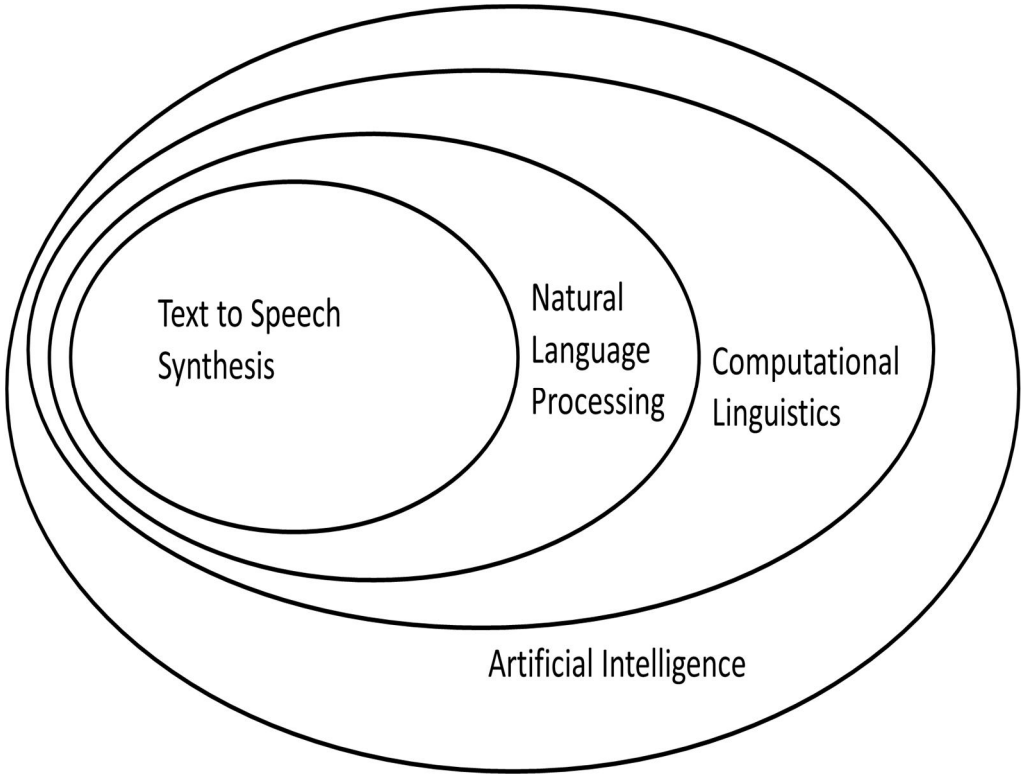


**Figure 1.5 Evolution of Text to Speech Synthesis system**

Natural Language Generation is the artificial production of human speech. When text is converted into speech, then it is called the Text to Speech Synthesis. Various other forms of Natural Language Understanding are the Speech Recognition and Speech to Text conversion.

## 1.2 OBJECTIVE OF THE RESEARCH

Upon reviewing the framework of a Text to Speech Synthesis System, it was found that there was a scope for developing a Text to Speech Synthesis System for the Tamil language. A Text to Speech Synthesis System may help people who cannot read a language but still can understand. The beneficiaries included blind people and native speakers who cannot read the language. There were also situations where vision communication is infeasible for a large crew, and there is a need for speech as a communication medium. Ex.: Automatic announcement systems in Railway Stations, Bus stations, airports and other public gatherings. There was a need for a Text to Speech Synthesis System in regional languages as an assistive technique for the public. Apart from public aiding systems, the Text to Speech Synthesis Systems finds their applications in Machine Translation systems as a post-processing paradigm. The processed speech is needed to be more natural, so as to assist people who are non-native to the language also. The core objective of the Research work is to develop a Tamil Text to Speech Synthesis System, which can convert the given Tamil text into Tamil Speech that is equivalent to a natural human speech.

## 1.3 SCOPE FOR A TEXT TO SPEECH SYNTHESIS SYSTEM

As Natural Language Processing evolved as a stand-alone area of research, it led to the allied fields such as Automatic Summarization, Machine Translation, Information extraction and retrieval, Speech Processing, etc. Speech processing included acquisition, manipulation, storage, transfer and output of speech signals.

Figure 1.6 shows the text-to-speech synthesis cycle. The text is preprocessed into an intermediate form. The process includes breaking down the text into tokens and expanding the numerals. The prosodic phrasing phase

groups the tokens into meaningful chunks. Pronunciation generation phase identifies the appropriate utterances from the Lexicon. The Segmental Duration phase assigns the duration value of each utterance. The Intonation Generation phase assigns proper tones and breaks. Appropriate waveforms originate from the available waveform repositories, so that the speech is synthesized.

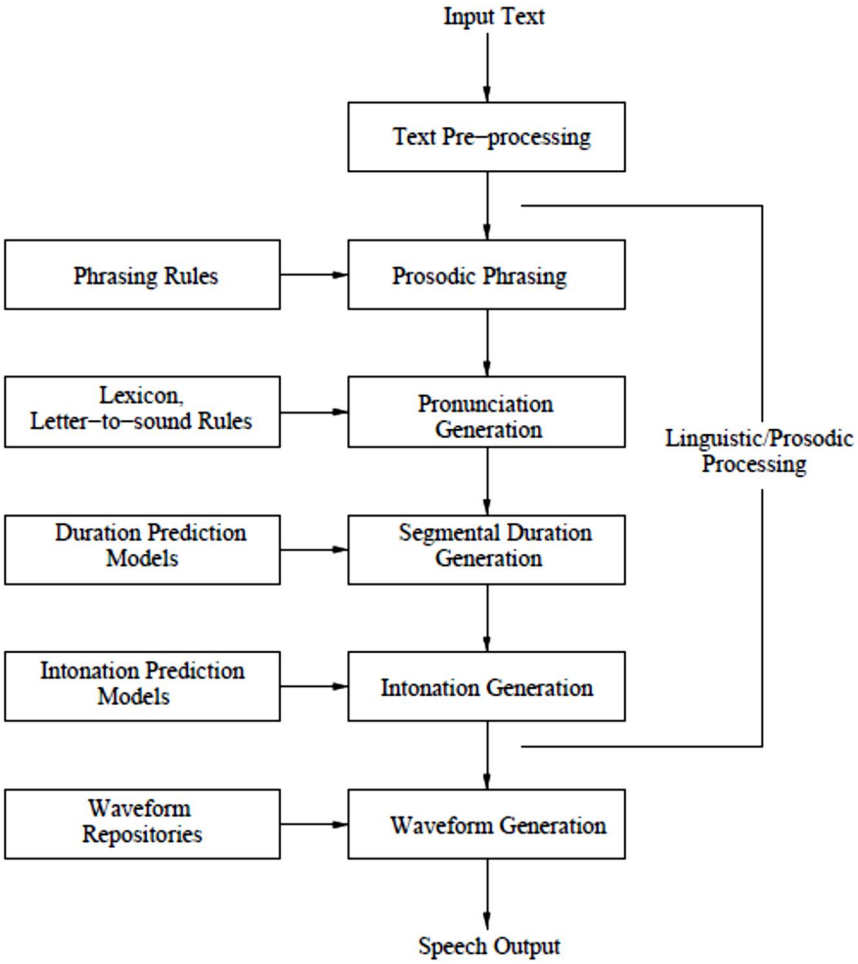


**Figure 1.6 Text-to-speech synthesis cycle**

The waveform-generation component takes as input the phonetic and prosodic information generated by the various components described above and produces the speech output. This cycle runs for each phrase of the parsed text.

## 1.4 TEXT-TO-SPEECH SYNTHESIZERS FOR INDIAN LANGUAGES

Low-density languages are those for which few language resources are available. Work relating to low-density languages is becoming a focus of increasing attention within language engineering, as discussed (McEnery Tony et al. 2000). MILLE is designed to address a range of issues to enable language engineering research on Indic languages. When Jayavardhana Rama et al. (2001) developed the first TTS system, more work on intonation and duration modeling was proposed and carried out, than the aspect of reducing the effective synthesis cost. Anumanchipalli Gopalakrishna et al. (2005) proposed a simple methodology for database creation in Indian languages.

The Dravidian language Tamil has over 65 million speakers worldwide. It is the official language of Tamil Nadu, and also of Singapore, Sri Lanka, and Mauritius. Tamil speakers widen their spectrum in countries like Malaysia, Qatar, Thailand, United Arab Emirates, Canada and the United Kingdom. It will be a boon to all Tamilians, if there is a user interface with the computer in Tamil, that too in the form of speech. The synthesizer can be used as an automatic text reader for the blind. It has applications to read e-mails, web pages and also as talking newspaper.

## 1.5 ISSUES FOUND IN DIPHONES

A diphone is defined as two connected half phones and describes the transition between two phones (Kiruthiga & Krishnamoorthy 2012a) by starting in the middle of the first phone and ending in the middle of the second phone. It describes the co articulation effects and minimizes the discontinuities at the concatenation points. Diphones are comparatively bigger units than phones. There are about thousand and seven hundred diphones found in the Tamil language. Unlike phones, they do not have allophonic

variations. i.e., each diphone has only one instance of pronunciation. Diphone concatenation can generate a reasonable quality speech. A single example of each diphone is not enough to produce good quality speech (Kiruthiga & Krishnamoorthy 2012a). Moreover, diphone-based synthesizers need elaborate prosody rules to produce natural speech. Diphones cannot capture co-articulation better than recent methodologies. As analyzed by Kiruthiga & Krishnamoorthy (2012b), concatenation points required by them are comparatively more, so it needs large size of the database to store corpus data.

While selecting and synthesizing the individual speech units, a Target cost has to be calculated, as proposed by Hunt et al. (1996). Clustering techniques were used by Alan Black & Paul Taylor (1997), to reduce the selection cost. But for a Dravidian language like Tamil, clustering was not specifically required due to the structure of the language.

## 1.6 ISSUES FOUND IN SYLLABLES

The Tamil language is syllable oriented, where pronunciations are mainly based on syllables. A Syllable can be the best unit in Tamil language (Kiruthika & Krishnamoorthy 2012b) Speech synthesis systems. Intelligible speech synthesis is possible for the Tamil language with the basic unit as the syllable. Though the number of syllables is larger in comparison to phones or diphones, it can describe co-articulation better than phones. When syllables are used, the concatenation points relatively decrease. Low Energy regions characterize syllable boundaries. The following are the syllabification rules followed for the Tamil language:

### 1.6.1 Syllabification Rules

1. Nucleus can be a Vowel (V) or a Consonant ( C )

Examples:

(i) *‘/A/.’* There is no onset or coda. The nucleus is a vowel.

(ii) *‘/sangam/’*

*‘/sang/’, ‘/ga/’, and ‘/m/’*. Here, the last syllable ‘/m/’ is a consonant. There is no onset or coda in this syllable.

2. If onset is C, then nucleus is V to yield a syllable of type CV

Example:

*‘/kErai/’*

*‘/kE/’, and ‘/rai/.’* Here both the syllables have the format CV. (For instance, *‘/k/’* is a consonant and *‘/E/’* is a vowel, thus forming the pattern CV)

3. Coda can be empty or C

Example:

*‘/Am/’* There is no onset. The coda is a consonant *‘/m./’*, and the nucleus ic *‘/A/.’*

4. If characters after CV pattern are of type CV, then the syllables are split as CV and CV.

Example:

*‘/puli/’*

*‘/pu/’and ‘/li/.’* Here both the syllables have the format CV. So those syllables are split as CV and CV.

5. If the CV pattern is followed by CCV, then syllables are divided as CVC and CV.

Example:

*‘/selvam/’*

*‘/se/’and ‘/lvam/.’* Here the syllable *‘/se/’* is of the format CV. The succeeding term *‘/lva/’* is of the form CCV. Therefore, these two terms are split into CVC *(‘/sel/’)* and CV *(‘/vam/’*).

6. If the CCCV pattern follows CV, then the syllables are divided as CVCC and CV

Example:

*‘/pArththAn/’*

*‘/pArth/’,’/thA/’ and ‘/n/.’* Here both the syllable *‘/pArth/’* has the format CVCC, and the syllable *‘/thA/’* has the format CV. The first term *‘/pA/’* is of the form CV. The succeeding term *‘/rththA/’* is of the format CCCV. Therefore, these two terms are split into CVCC *(‘/pArth/’)* and CV *(‘/thA/’).*

7. If the vowel V follows the pattern VC, then the syllables are split as V and CV.

Example:

*‘/athu/’*

*‘/a/’and ‘/thu/’*. Here both the syllable *‘/a/’* is a vowel V. *‘/thu/’* has the format CV. When the VC pattern *‘/ath/’* is followed by V *(‘/u/’),* the terms were split into V and CV.

8. If the pattern CVC follows VC, then the syllables are split as VC and CVC

Example:

*‘/anJal/’*

*‘/anj/’and ‘/Jal/’*. Here, the syllable ‘/anj/’ has the format VC. The next syllable ‘/Jal/’ has the format CVC. This word exhibited the pattern VC with CVC, it is split into the same.

### 1.6.2 Sample Syllabification

Syllabification of a Tamil word *‘/thamizhagam/’* is illustrated as follows:

*‘/tha/’ ‘/mi/’, ‘/zha/’, ‘/gam/’*

CV CV CV CVC

The phonetic representation of the above syllables is given as: *[tha], [mi], [La], [gha], and [mh]*. Here, in the syllable *[tha],*

- Onset: *[th]*
- Rhyme: *[a]*
  - Nucleus: *[a]*

There is no coda in this syllable.

In the syllable *[mh],*

- Onset: *[]*
- Rhyme: *[mh]*

- Nucleus: *[mh]*

Here the nucleus is a consonant. This is a nucleus-alone syllable.

The general format of a Tamil language syllable is C*VC*, where C is a consonant and V is a vowel. C* indicates the presence of 0 or more consonants. This model alone may not bring a proper natural language speech synthesis since syllable concentrates only in vowels and consonants.

The employment of appropriate acoustic unit was carried out with units such as syllables and polysyllables by Samuel Thomas et al. (2006), and Words and Triphones by Thangarajan et al. (2008). Individual letter parsing done in phoneme level needs to be implemented at the end of every sentence to pronounce the sentence termination naturally. Scientific notations, website link, email address, stress notes are processed by diphone units (Kiruthika & Krishnamoorthy 2015a) and need to be concatenated with the already processed syllable unit.

The idea of combining the usage of two different speech units evolved and that was also not sufficient for the naturalness of synthesized speech. Therefore annotating the database with necessary prosodic information had to be carried out for bringing naturalness in the synthesized speech had been proposed by Grazyna Demenko et al. (2006). Various Language Modeling techniques suggeted by Ashwin Bellur et al. (2011) were surveyed, and the necessary labeling and annotations carried out brought more naturalness to the synthesized speech.

## 1.7 SCOPE OF THE THESIS

The commonly used speech units for the concatenative speech synthesis now-a-days are diphones, triphones, and syllables. This

phenomenon is primary because the concatenation points in each speech unit are well-addressed. Diphone-based synthesis requires a significant amount of prosody modeling for the duration, intonation, and energy. It, in turn, needs analysis of a voluminous amount of data and deduction of proper rules from the data. These efforts are both time-consuming and laborious.

It is a great advantage if a speech unit can produce a synthetic speech with a few number of concatenation points and intrinsically has sufficient prosodic information in it. Identifying such components and taking them into implementation is the chief task in any speech synthesis technique. This process reduces the works for prosody analysis and tagging. It results in faster development of TTS systems for Indian languages. Phonemes (the stand-alone elements) are not suitable for speech synthesis because they fail to model the dynamics of speech. Therefore, it is necessary to look for larger units. The syllables are of longer duration. Observations state that syllables are less dependent on the speaking rate variations than phonemes. The human auditory system integrates a time span of 200 milliseconds of speech that approximately corresponds to the duration of syllables (Greenberg 1996).

Syllables consist of co-articulation information between sound units better than the phonemes. For Indian languages, it is also seen that the syllable is a better choice than units like a diphone or a phone. Manual segmentation and labeling of speech units are tedious, time-consuming and error-prone. The scope of this thesis is to design a dual database model to improve the naturalness of the synthesized speech. We try to identify units suitable for synthesis and to automatically extract them from continuous speech.

## 1.8 ORGANIZATION OF THE THESIS

The thesis is organized as follows: Chapter 2 describes the various approaches for Text to Speech Synthesis. It analyzes a range of techniques used for Text to Speech Synthesis. This chapter explains the need for an annotated corpus for prosody modeling. Chapter 3 describes the architecture of the Distributed Speech Synthesis Software and the Festival speech synthesis framework. Chapter 4 illustrates the automatic generation of speech units. It explains techniques to automatically produce the speech units - syllables and diphones. We have also dealt with the generation and evaluation of syllable and diphone databases. This chapter also shows how the generated speech units are integrated into the Festival speech synthesis framework. It clearly explains the co articulation duration between various sound units. We have analyzed two different speech synthesis techniques the framework supports. This chapter explains the manual extraction of syllables and group delay based segmentation algorithm used to generate the speech units. Chapter 5 gives the big picture of the proposed System. It gives the detailed view of the works done in building an efficient Tamil Text to Speech Synthesizer that produces a natural, human-like speech. Chapter 6 gives the elaborate study of the design of two different databases for the proposed system. Chapter 7 explains the Prosody Modeling in detail. We have detailed the evaluation of the proposed synthesizer, with respect to the Prosody aspects. Chapter 8 presents a summary of the thesis and presents the directions for future work.

## 1.9 MAJOR CONTRIBUTIONS OF THE RESEARCH WORK

The major contributions of the research work presented in this thesis are as follows:

i. Identification of speech units suitable for concatenative speech synthesis.

ii. Demonstration of an automatic segmentation algorithm for generating the speech units.

iii. Development of a dual database one for the syllables and the other for diphones and integrating them into the Festival Framework.

iv. Proposing a suitable technique to develop distributed speech synthesis (DSS) systems for mobile devices.