CHAPTER 5

ARCHITECTURE OF THE PROPOSED SYSTEM

5.1 INTRODUCTION

Prosodic phrasing is an important and a tedious problem for Tamil language, as Tamil language scripts lags punctuation. This research is a preliminary attempt at data-driven modeling of prosodic phrase boundary prediction for the Tamil language. The Speech Synthesis System shows its naturalness when the corpus is annotated with prosodic information. Prosody modeling is subdivided into modeling the following constituents of prosody - phrasing, duration, intonation and intensity. A well-designed speech corpus, annotated with various levels of prosodic information discussed by Kiruthiga & Krishnamoorthy (2012a) is used.

Based on the performance on test data, the corpus is analyzed automatically to create a prosodic model. A training data set is then synthesized. Based on the performance on the test data, the prosody models are improved. Test data evaluation follows the above synthesis. The syllables have sufficient duration information and thus this phenomenon improves the quality of synthetic speech when used as a duration model. Though diphones lag such prosodic information, they address some other Articulatory issues, which syllables cannot. Thus syllables are identified as the best-suited processing units for (Kiruthiga & Krishnamoorthy 2012a) Tamil language Speech synthesis and diphones are identified as best-suited processing units (Kiruthika & Krishnamoorthy 2015c) for the sentences to be ended naturally.

5.2 CORPUS ENTRY FOR SYLLABLES AND DIPHONES

Syllable extraction is an important issue that has to be dealt with when selecting the text-corpus. The synthesizer covers the most frequent and unique syllables by an optimized prompt-list, which was selected for recording the speech corpus for Tamil. A rule-based parser was used to generate syllables from the UTF-8 text. The two types of speech data are manually labeled at syllable boundaries and using Ergodic Hidden Markov Models (Matt Shannon & William Byrne 2010). The first method resulted in a number of labeling errors, while the latter required a large amount of training data, to mark syllable boundaries accurately. To resolve the issues discussed above, the speech data is labeled using a semi- automatic labeling tool developed at IITM, based on segmentation and identification of syllable units done using group delay function and onset vowel point. A separate database is maintained for Diphones. Diphones are relatively bigger units than phones. With the available data, there are about 1000 to 2000 diphones found in Tamil language. Unlike phones, the allophonic variations exhibited by diphones have separate entries in the database, i.e., each diphone has only one instance of pronunciation. To limit the size of the database, diphone entries are used only for last entries of sentences to end the speech naturally and to pronounce punctuations, scientific notations, email addresses and website links naturally and descriptively.

The Unit selection paradigm uses the recorded speech corpus to extract the syllable units to synthesize speech.

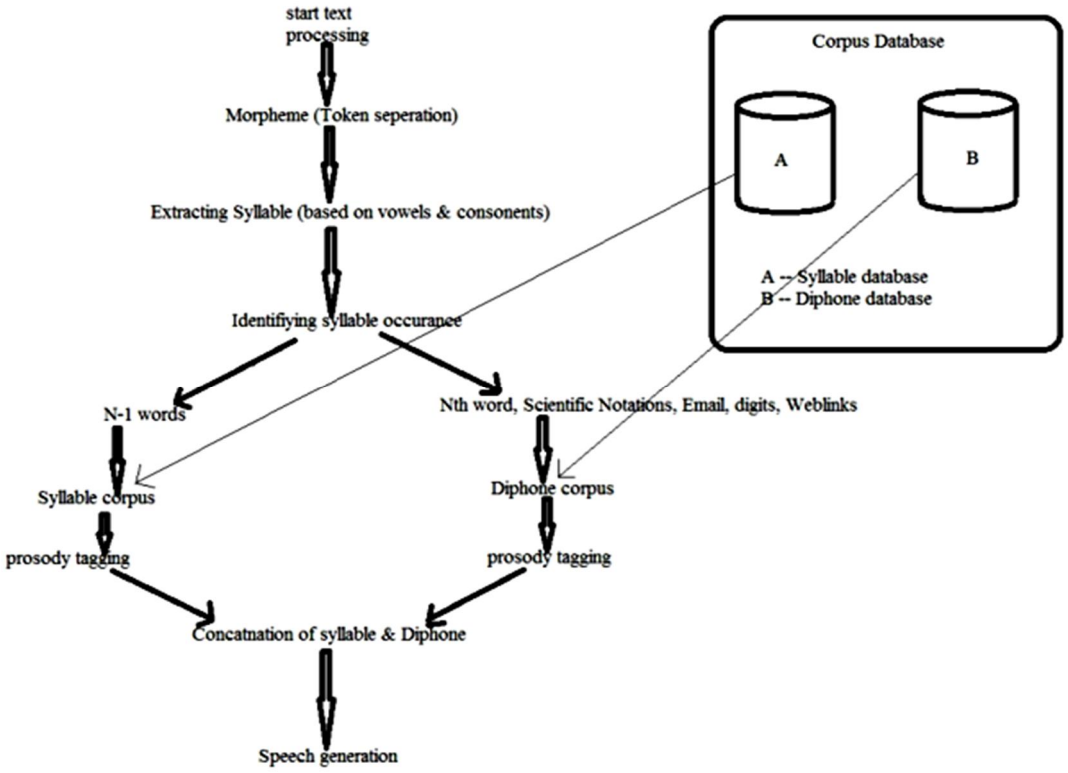

Figure 5.1 Modeling a dual database for Tamil Speech Synthesis.

The synthesizer employs a clustering technique to select the area where diphones are to be used, as there are many realizations of the same syllable being present.

An acoustic join cost is a measure to select an optimal path through these candidate sets. When speech is processed as diphones in the above said areas, the speech would be natural and descriptive. The following criterion has to be taken into account to decide whether to synthesize the given phrase as a monosyllable or a diphone:-

(i) The Identity of neighbouring units.

(ii) The Position of syllable in the word.

(iii) The Syllable at a phrase boundary.

Figure 5.1 shows the model of a dual database for a Tamil Text to Speech Synthesis system. After the morpheme analysis of the given training set, occurrences of syllables are identified. If there are N words, syllables are extracted up to the (N-1) words. Syllables in N^{th} word are also extracted. Utterances of these syllables are derived from the syllable database. Diphones are separately synthesized from the last syllable. The end-syllable of the N^{th} word is divided into diphones. Acoustic representations of the diphone units are taken from the diphone database for the actual speech synthesis.

5.3 MEETING THE FUNCTIONAL COMPONENTS OF PROSODY

The three main functional components of Prosody that have to be modeled are:

1. Phrasing
2. Duration
3. Intonation

Of course, there are some more prosodic features like Tempo, Melody, Pitch and Rhythm. The attributes of the pitch are addressed by intonation itself. Tempo, melody, and rhythm do not play a significant role in our Text to Speech Synthesis System. Therefore, the components that are listed above are to be met by our Text to Speech Synthesis System.

Phrasing is carried out at various levels such as a sentence, syntactic phrase, semantic phrase and polysyllables. Diphones are separately extracted at special places of articulations viz., sentence boundaries, website addresses, email ids, numerals, etc.

Duration modeling is carried out using CART (Grazyna Demenko et al. 2006) model. Intonation is carried out using original English ToBI (Tones and Break Indices) conventions.

5.3.1 Phrase Modeling

Phrasing needs parsing the text into various levels of articulation, viz., sentences, syntactic and semantic patterns. The text is parsed into sentences at the first level. Each sentence is parsed into syntactic patterns. The purpose of POS tagging is to find out the syntactic category of a word in a sentence. Table 5.1 gives the list of POS tags and their usage (Amrita POS Tagset - Proposed by Dhanalakshmi et al. 2009a)

For example, the following sentence can be tagged as below:

inRu vIttil niRaiya vElaikal uLLana.

N NP Adj N VP

There are also four sub tags describing other grammatical details of the words, such as tense, number, person and case of the words. The Sub-tags are:

1.Tense: Prs - present tense, Pst – past tense, Fut – future tense

2.1,2,3 – first, second, and third person respectively,

3.S – singular Pl – plural,

4.Neg – negative, acc – accusative case, dat – dative case

Therefore, the above said tagged sentence is tagged as below:

inRu <N><Prs>

vIttil <NP>

niRaiya <Adj>

vElaikal <N><Pl>

uLLana. <VP><Pl>

Table 5.1 POS Tags and their descriptions

S.No.	Tag	Description
1	N	Noun
2	NP	Noun Phrase
3	NN	Noun + noun
4	NNP	Noun + Noun Phrase
5	IN	Interrogative noun
6	INP	Interrogative noun Phrase
7	PN	Pronominal Noun
8	PNP	Pronominal Noun Phrase
9	VN	Verbal Noun
10	VNP	Verbal Noun Phrase
11	Pn	Pronoun
12	PnP	Pronoun Phrase
13	Nn	Nominal noun
14	NnP	Nominal noun Phrase
15	V	Verb
16	VP	Verbal phrase
17	Vinf	Verb Infinitive
18	Vvp	Verb verbal participle
19	Vrp	Verbal relative participle

Table 5.1 (Continued)

S.No.	Tag	Description
20	AV	Auxiliary verb
21	FV	Finite verb
22	NFV	Negative Finite verb
23	Adv	Adverb
24	SP	Sub-ordinate clause conjunction Phrase
25	SCC	Sub-ordinate clause conjunction
26	Par	Particle
27	Adj	Adjective
28	Iadj	Interrogative Adjective
29	Dadj	Demonstrative Adjective
30	Inter	Intersection
31	Int	Intensifier
32	CNum	Character number
33	Num	Number
34	DT	Date Time
35	PO	Post Position

Followed by POS tagging, syllable extraction is done. Prediction of syllable or diphone units depends on the phrase frequency occurrence in Corpus table. It is a higher risk to find phrase boundaries in Dravidian Tamil language because of non-occurrence of case markers. However, it can be done manually with the help of Morpheme tags. Morpheme tags in Tamil used to annotate the phrase correctly and to predict whether it could be processed by syllable or diphone. The Morpheme tags help us to predict the phrase-boundaries based on the occurrence of phrases, and also it helps us to predict the synthesis mode of each and every phrase.

Table 5.2 Predicting occurrences of morphemes in Tamil

S.No	Morpheme	Words	Occurrences	Percentage
1	Adhu	KorpAdhu	65	0.65%
2	Kal	VaazthukKal	248	2.48%
3	L	PoruL	1648	16.4%
4	Da	AnbuDan	945	9.4%
5	Um	SudUm	644	6.44%
6	Illai	PaarkavIllai	453	4.5%
7	Ven	VaruVen	125	1.2%

Here Table 5.2 gives the correlation between predicted morphemes in actual words and it occurrences by CART model.

5.3.2 Duration Modeling

Duration modeling is carried out by using CART (Classification And Regression Trees) Technique. Classification and Regression Trees are models based on self-learning procedures. They sort the instances in the learning data by binary questions about the attributes that the instances have (Samuel Thomas et al. 2006). CART modeling is an optimal technique for Duration modeling in Indian Languages, where there are multiple duration patterns for a single morpheme.

A Classification And Regression Tree is a self-learning Tree Structure. Like any other binary tree, a CART also has a parent node, which can generate only two child nodes; recursively each child node may act as a parent node to create two different child nodes. Those nodes which do not have children are called leaf nodes. Learning in a CART is done using binary

questions that start at the root node. The answer to the binary question will select a child node; another binary question and its answer will lead to the next level till the search reaches the leaf node. Figure 5.2 represents a CART used for Duration modeling of the proposed Text to Speech Synthesis System. The triangles in the given tree represent the leaf nodes.

Clustering is carried out for capturing the gross acoustic properties of the syllables. The phrase boundaries have a massive role in fluently connected speech. The energy contours within a phrase vary depending on its relative position in the utterance.

Therefore, before entering into the CART, syllables are clustered using the following features (Ashwin Bellur et al. 2011):

- Word length of the adjacent words in the units of the number of syllables constituting the word.
- The Distance of the syllable from the beginning of the phrase in the units of a number of words and number of syllables.
- The Distance of the syllable from the end of the phrase in the units of the number of words and number of syllables.
- The Relative position of the parent phrase in the utterance.
- The Position of the syllable with reference to the phrase boundary.
- The Identity of neighboring syllables.
- Features of previous syllables as defined in the syllable set.

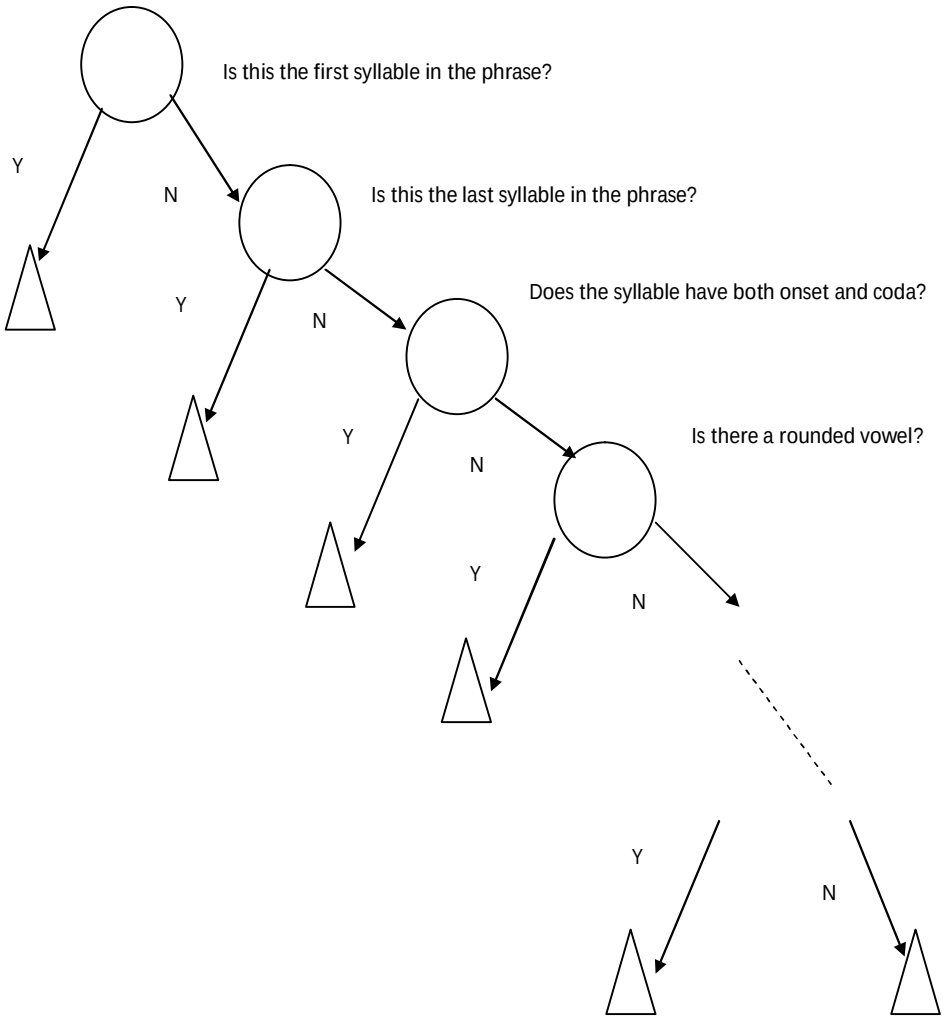

Figure 5.2 An example CART tree

5.3.3 Intonation Modeling

Intonation modeling is carried out by following the ToBI (Tones and Break Indices) standard prescribed for original English. There are no dedicated Tones and Break Indices tabulated for the Tamil Language. So English ToBI guidelines are followed. There are two simple tone indices H for a High tone; and L for a low tone. An asterisk (*) indicates which tone aligns with the stressed syllable of the word. L+H indicate the Rising tone, and the H+L implies the falling tone. With these basic tone indices, the

following pitch patterns are used for modeling. Table 5.3 gives the standard pitch accent patterns that are used in our system.

Table 5.3 Standard pitch accent patterns

S.No.	Pitch Pattern	Description
1	H*	Pitch peak
2	L*	Pitch trough
3	L+H*	Rising peak accent
4	L*+H	Scooped accent
5	H+!H*	High pitch unaccented

H and L indicate relative high pitch and low pitch in the intonation contour. Their actual phonetic realization contributes to the factors such as pitch range and the sequential pitch accents in the phrase. Since Tamil is a kind of post-lexical language, pitch accents occur on stressed syllables. They form their unique characteristic patterns upon the pitch contour. The fundamental syntactic information of POS of words in a sentence contributes to framing rules for pause insertion (Youngim Jung & Hyuk-Chul Kwon 2011). The different levels of break markers in the model represent the prosodic structure of a sentence.

5.4 SYSTEM ARCHITECTURE

The system architecture deals majorly with the formation of speech database. Figure 5.3 gives the system Architecture. The classical method of construction of formal speech corpora needs the following necessary steps:

- Selection of appropriate textual content
- Recording of Textual content

- Parsing the recorded speech content at various levels
- Annotating features to the speech units
- Storing those annotated units into the corpus

The proposed synthesizer is designed to use a dual corpus. One of the corpora is a repository of annotated syllable units; while the other one is the repository of annotated diphone units. Therefore, we need to distinguish orthographic text and special words viz., website addresses, email ids, numerals, etc. The selected textual content converted into spoken content. The morpheme analysis phase takes care of distinguishing normal text and specific words.

The parsing stage divides the content into various levels such as paragraphs, sentences, phrases, syllables and, of course, diphones at required instances. Each phrase goes into POS Tagging. Syllables are segmented from the phrases for further processing. The general format of an Indian language syllable is C*VC*, where C is a consonant, V is a vowel and C* (Kiruthiga & Krishnamoorthy 2012b) indicates the presence of 0 or more consonants.

Generally when grouping Indian languages, there are about 35 consonants and 18 vowels. There is a defined set of syllabification rules formed by researchers, to produce computationally reasonable syllables. A rule-based morpheme to syllable converter syllabifies the given parsed phrases. Scientific notations, website link, email address, end of sentences, stress notes are to be processed by diphone units and concatenated with the already processed syllable unit.

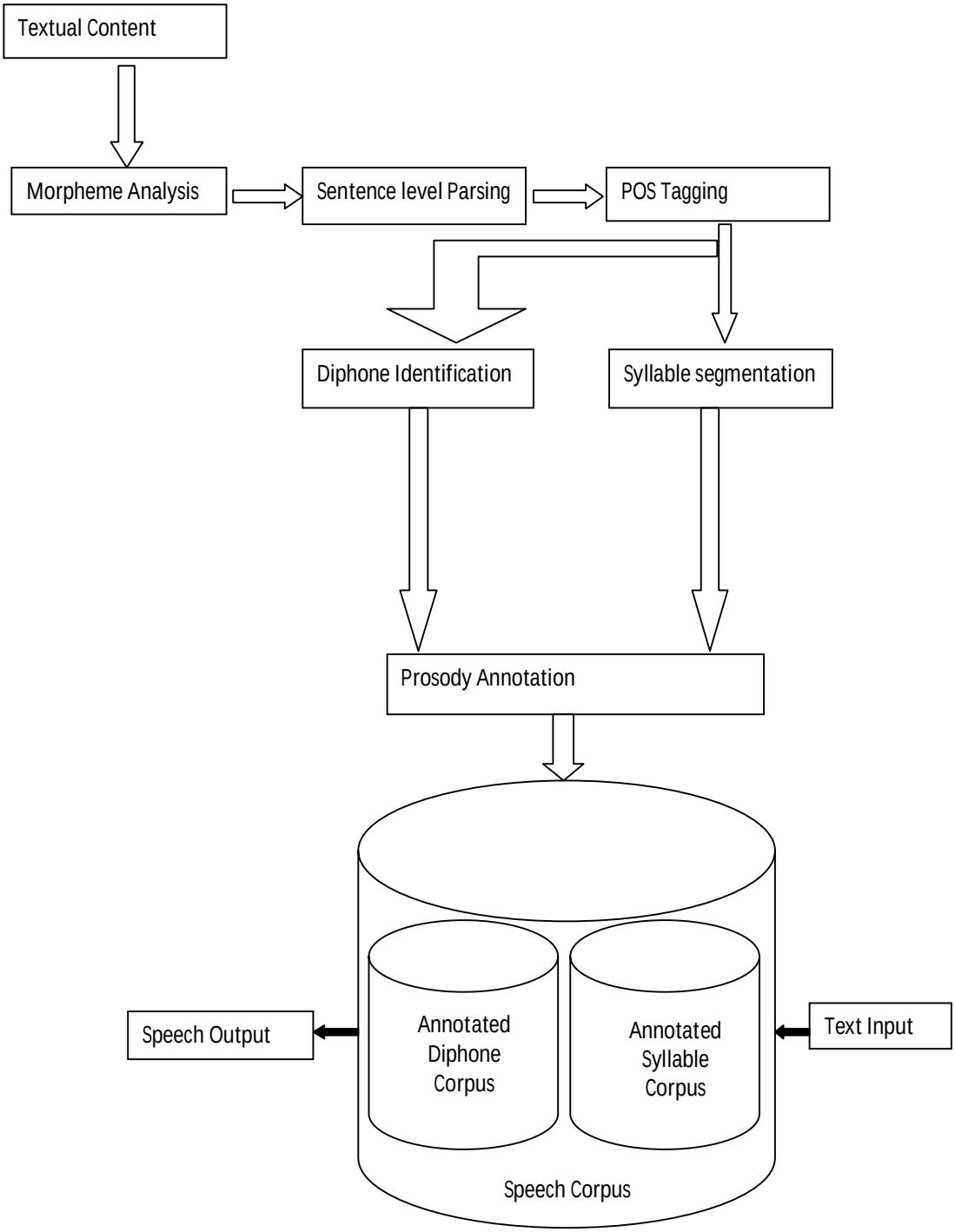

Figure 5.3 Proposed System Architecture

The Prosody Annotation phase deals with modeling the various components of prosody such as Phrasing, Duration, and Intonation. Phrasing starts with POS tagging, where there are 35 main tags and about 5 sub tags. Furthermore, phrase boundary detection may also be carried out for the complete phrasing scenario. CART technique carries out duration modeling of the phrased units, which is associated with an appropriate clustering algorithm. Tones and Break Indices guidelines for original English are followed in Intonation modeling since there is no ToBI model exists especially for Tamil. Figure 5.3 explains the Architecture of the system.

5.5 SYLLABLE AND DIPHONE CLUSTERING

The present clustering process proposed by Matt Shannon & William Byrne (2010) was unable to capture the gross acoustic properties of the units using the default feature set. A better linguistic feature vector set can extract the complex acoustic properties of a syllable and diphone. Therefore, such a vector is employed. On analyzing the speech corpora, it was seen that phrase boundaries had a vital role in fluently connected speech. Tamil languages lack punctuations in general. Phrase boundaries and intra phrase prosodic patterns using in present scenario help in understanding an utterance. To overcome the above issue, a separate corpus database was used to process diphone units. The processing of an arbitrary text in a Text to Speech Synthesis system is shown in the Figure 5.4. Initially, the text should be morphologically divided into valid tokens. This process helped us to identify the syllable occurrence and sentence completion before processing. The tokens will be separated and entered into the appropriate fields in the symbol table. Initially, the database has been designed with two individual blocks - one for storing syllable and the other for storing diphones. Figure 5.5 shows the functionalities of the front-end and back-end of the TTS system.

First syllable occurrence in every sentence has been monitored and reported. The (N-1) words syllables are taken and stored in syllable database. The diphone methodology processes the N^{th} words, and the units get stored in diphone database. The need for selecting the phonetically and prosodically best units for synthesis requires clustering the units from both the databases. The synthesizer evaluates the factors concerning prosodic and phonetic context to form clusters within a unit type. A decision tree is built based on questions regarding the phonetic and prosodic aspects of the morpheme. Eventually, the leaves of the decision tree are the candidate set of database units that best suit the required features. At the time of synthesis, to generate each target unit, the appropriate decision tree is used. It finds the best cluster of candidate units. A Viterbi search is then made to find the best path through the candidate sets.

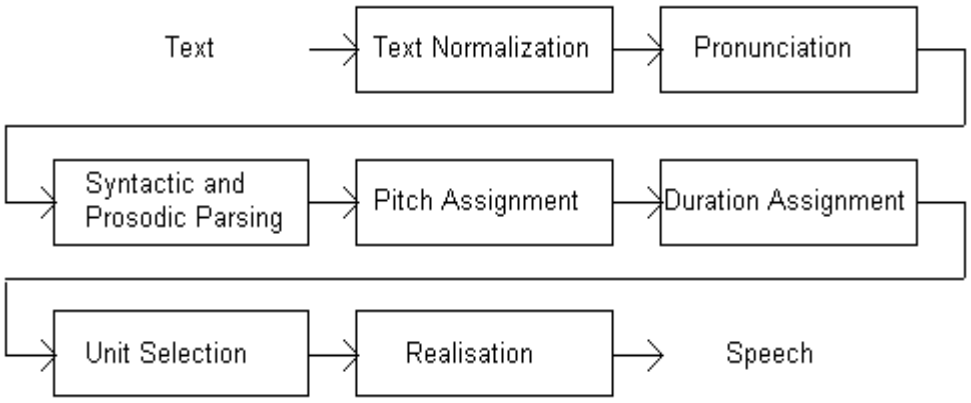

Figure 5.4 Text to Speech Synthesis System

Pruning is performed to remove spurious unusual phrases which may be the results of mislabeling or a poor articulation in the original recording. The synthesizer also identifies and removes units which are so common to each other. I.e., there is no significant distinction between the candidates. Keeping these units in the database may lead to confusions in calculating the Target cost C_t. Databases were tested with this clustering method in hand. The method produced both extremely high-quality examples

and extremely low-quality ones. Minimizing these bad examples is the significant target.

Clustering includes pre-clustering works that tag the syllables as begin, middle and end, depending on the occurrence of the syllable in the word. Tagging the syllables based on the type of the syllable (V, C*V, VC*, C*VC*) and nature of the participating vowels and consonants comprises further clustering. Syllables of the same pattern were clustered using features like word length of the phrase, the relative position of the syllable in the phrase, the relative position of the parent phrase and the acoustic parameters of the preceding and the following syllables in the phrase.

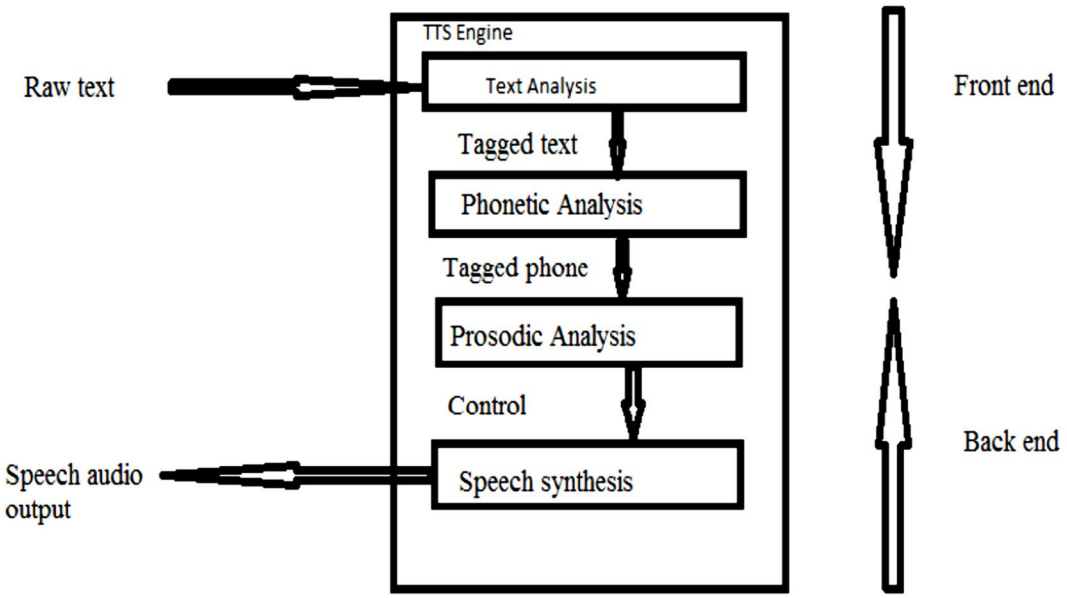

Figure 5.5 Various phases of a Text to Speech Synthesis System

Using the feature set and the acoustic distance measure, the decision tree is built for each of the individual syllable entries in the database whereas; the diphone database gives naturalness and descriptive ending while reading. A proper punctuation notes will be given spotlessly by the diphone units while reading digits, email addresses, and scientific notations. Questions

are used at the nodes to find the best set of candidate syllables. Morpheme tagging helps in phrase boundary prediction. The above paradigm is used to differentiate the selection between syllable and diphone databases.

5.6 CONTEXT SWITCHING BETWEEN DATABASES

When the dual database concept is introduced in the Text to Speech Synthesis systems, overall performance increases. But the time taken for actual speech synthesis also increases. This phenomenon is because of the time taken for the switching between databases. Figure 5.6 explains the Context-Switching between the databases. The synthesizer has to decide upon which database to access, at a particular instance of time. As we have already discussed, at normal text, up to (N-1) words, the syllable database is used. At the N^{th} word, diphones come to the picture. Therefore, there is a constant switching of databases at every sentence termination i.e., if there are 'n' number of sentences, the context switching takes place 'n' number of times.

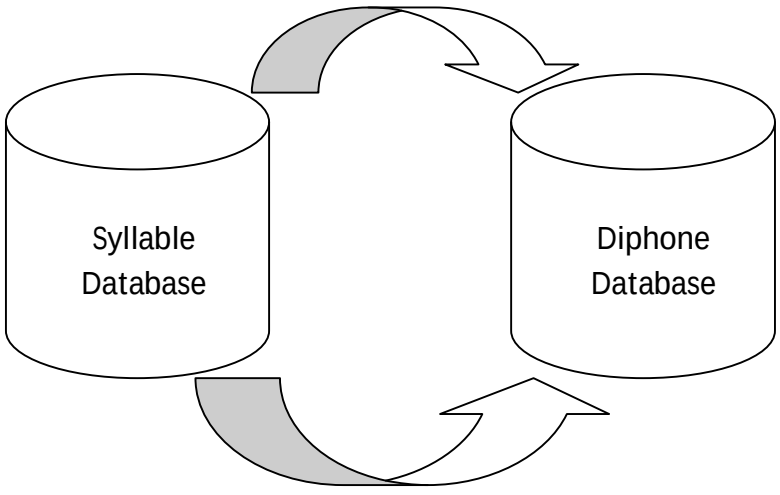

Figure 5.6 Context switching between databases

Apart from this, the special places of articulations need diphones for synthesis. At every special place of articulation, databases must be

switched twice. (Even though it is ‘twice’, the term ‘context switching’ considers this ‘to and fro’ process as a single operation.) Therefore, there are more than ‘n’ numbers of context switches for ‘n’ sentences. Particularly for texts which have more scientific notations, e-mail addresses, dates, price-values, etc., furthermore transitions are needed. Hence there is an increase in processing time.

While evaluating the increase in synthesis time, the synthesizer becomes ‘idle’ for a few milliseconds, which is unlikely. Figure 5.7 shows the Duration of speech synthesis in Dual Database.

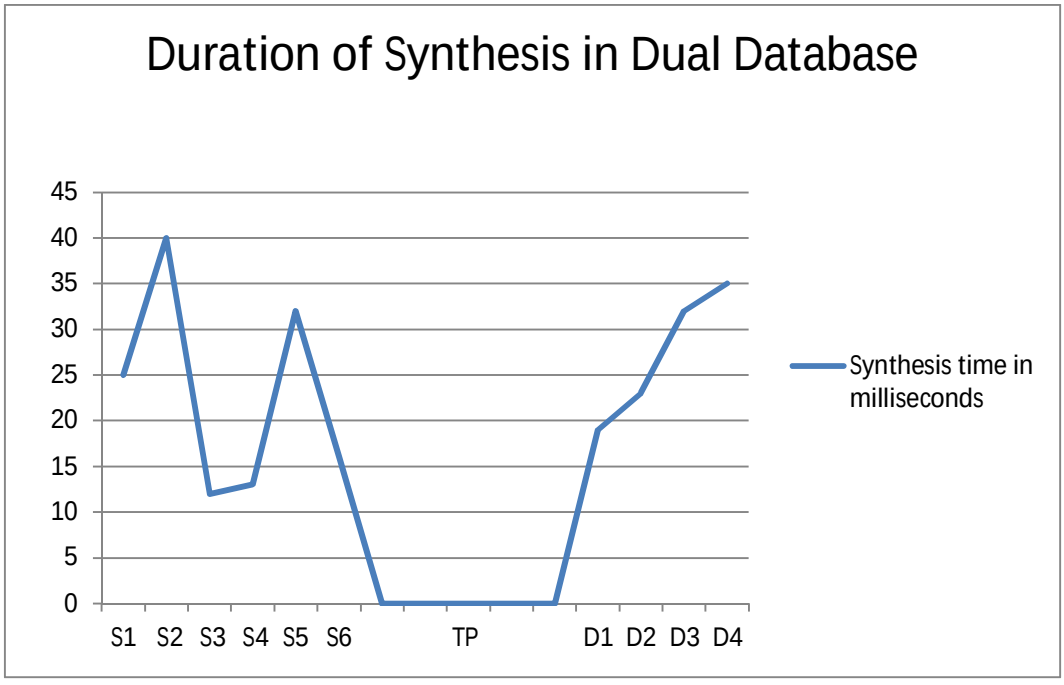

Figure 5.7 Effect of Context Switching between databases in time taken for overall synthesis

The trough in the graph shows the Transition Period (TP). It is the time taken for deciding and selecting between databases. The overall

processing time increases, as there is a time overhead due to context switching.

5.7 SUMMARY

This chapter has presented the architecture of the proposed system. It clearly explains various aspects that led to the design of a dual database for the Tamil Text to Speech Synthesis system. We have reviewed all the aspects of Prosody and have chosen those aspects which need extensive modeling. We have also examined the concept of Context Switching, which is the factor concerned with the overall performance of the system. In the next chapter, we shall discuss the techniques used to build a synthesizer using the Festival framework.