CHAPTER 2

TEXT-TO-SPEECH SYNTHESIS

2.1 INTRODUCTION

This chapter explains the significant issues of text-to-speech synthesis – selection of an appropriate speech unit for synthesis, and prosody modeling. While examining the significance of the choice of a suitable unit for synthesis, there found is a relationship between phrases and syllables. Syllables themselves comprise of a specific tone, which consists itself of their fundamental frequency (f_0). We have discussed various aspects of Prosody and techniques for implementation in detail. The research work has also identified some of the aspects need not be modelled. At the perceptual level, naturalness in speech is attributed to certain properties of the speech signal related to audible changes in pitch, intonation, loudness and syllabic length (duration), collectively called prosody. These are the auditory aspects of Prosody. Those changes correspond to the variations in the fundamental frequency (f_0), amplitude and duration of speech units and are called the acoustic features.

Prosody is very important for a synthesized speech, as it conveys aspects of meaning and structure implied. Generating the right prosody for synthetic speech in TTS systems is a challenging task because the Tamil language text mostly contains little or no explicit information about the expected prosody for the text and it is extremely tricky to deduce prosodic information automatically. The output of the text analysis module is a phonetic or a linguistic transcription of the input text along with the desired pitch, duration and intensity information of the output waveform. The

evolution of a Text to Speech synthesis system begins with the Natural Language Generation paradigm. NLG applications are data-to-text systems. They generate textual summaries of the databases and data sets. These systems usually perform data analysis as well as text generation.

2.2 NATURAL LANGUAGE GENERATION

NLG is a natural language processing task. It generates natural language from a machine representation system such as a knowledge base or a logical form. The simplest examples are systems that produce form letters. Grammars and syntax analysis are not parts of it. They just work on templates. The Text to Speech synthesis is called as 'Spoken Language Generation' in this context. Figure 2.1 describes the Typical Natural Language Generation System. NLG plays a critical role in applications such as text summarization, machine translation, and dialog systems. NLG used two types of techniques called rule-based and Template based methods.

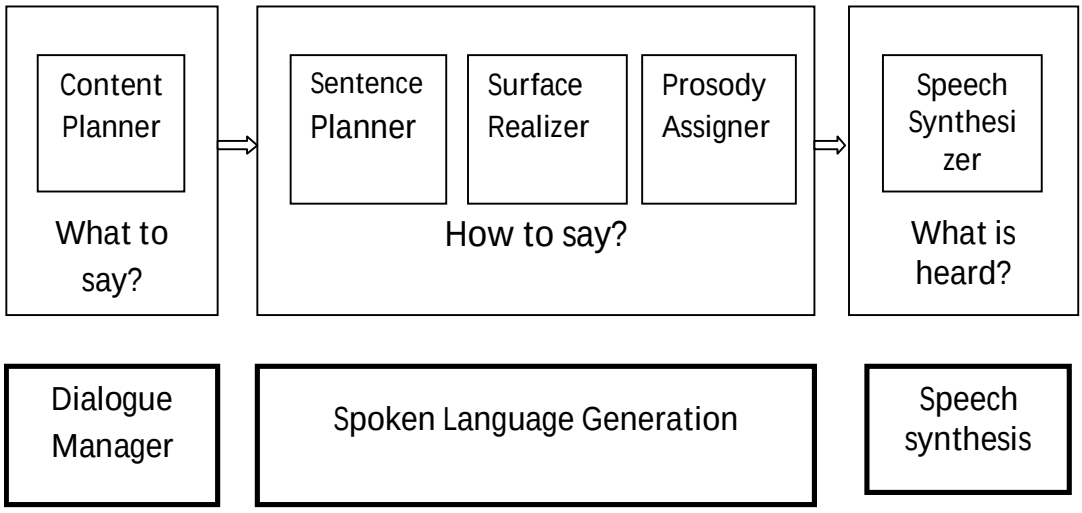

Figure 2.1 Typical Natural Language Generation Systems

The Dialogue Manager takes care of "What to say?"; The Spoken Language Generation subsystem governs "How to say?" and the Speech

synthesizer decides “What is heard”. The corpus is an utterance dictionary; the annotated corpus is a set of ‘tagged utterances’. A generation algorithm produces these utterances. There are more than one utterances generated out of a single sample text. For each randomly generated utterance, we compute a penalty score. The score depends on the heuristics we have empirically selected. A single utterance is assigned various penalty scores. A penalty score determines which utterance has to be chosen during the third phase – speech synthesizer.

2.3 EXPERTS’ FINDINGS IN PREVIOUS WORKS

The existing works proposed by the experts who have worked in this field have found in the literature (Samuel Thomas et al. 2006), in their research that no explicit prosody modeling is necessary when certain units occupy Tamil Speech Synthesis. All they propose is that using units close to syllables (they call it as ‘syllable-like units’) for Tamil Text to Speech Synthesis systems. As syllables themselves are the carriers of prosody, the experts recommend for no explicit prosody modeling for syllables. Still we propose a prosody model for improving the intelligence of the proposed system so as to augment the naturalness of the synthesized speech. But still, some realizations for each unit in the speech corpus may have to vary while working on the test data. Syllable database has to be appended with the newly realized units.

Work on Speech Recognition, which is carried out by Thangarajan et al. (2008) suggested words and triphones as the fundamental units. Since the research belonged to Speech Recognition scenario, prosody and naturalness played no role, as the input itself is the human speech.

A POS tagged corpus was an idea discussed by Dhanalakshmi et al. (2009) claimed to evolve to a database annotated only with Part Of Speech

components. They used The POS tagged corpus for chunking the developed corpus.

Another idea proposed by Alan W Black and Paul Taylor 1997 is a technique that automatically clusters similar units in unit selection paradigm. A decision tree is constructed for each entity in the database, whose leaves are a candidate set of database units that are best identified by the binary questions that lead to that leaf node. Such a clustering algorithm reduces the target cost, which in turn reduces the total cost.

Ravi Teja Rachakonda & Dipti Misra Sharma (2011) proposes an annotated Tamil speech corpus as a discourse resource. In linguistics, discourse is a unit of connected speech or writing longer than a sentence. A part of the corpus is made as a discourse resource and study was conducted. This paradigm helped in augmenting the information provided by dependency annotations at the sentence-level.

Comparative study of various speech units is given in Appendix 1. On examining the properties of the speech units, the phonetic richness of polysyllables and words is the appreciated ones. But database size trades off with the phonetic richness of these larger speech units. So syllables can be considered as the best-performing units for Indian languages.

2.4 APPROACHES TO PROSODY MODELING

The major constituents of prosody are:

1. Phrasing
2. Duration
3. Intonation and
4. Intensity.

They are the functional aspects of Prosody (i.e., aspects to be modelled). Other than these aspects, the other aspects of prosody are Grammar, Focus, Discourse, Tempo, Melody, Rhythm, and Emotion. For modelling prosody, we may have to go for either the rule-based approach or the corpus-based approach. In the rule-based approach, linguistic experts derive a complicated set of rules to model prosodic variations by observing natural speech. On finding new rules, they can be appended only by the experts, after the proper examination. In the corpus-based approach, a well-designed speech corpus, annotated with various levels of prosodic information is built. A test data runs on the corpus for analyzing automatically to create speech units. Based on the performance on test data, the speech units are then annotated to improve the quality of the synthesized speech. On finding new units, or new prosodic information of the existing units, they can be appended by the database administrator.

2.4.1 Prosodic Phrasing

English sentences usually exhibit some prosodic structures with some words in a given sentence group naturally while others introduce breaks. The same structure does not exist in Tamil sentences. Tamil sentences are well-structured ones. The pattern differs from that of the English sentences. The general format of an English sentence is always SVO (i.e., Subject – Verb – Object). In Tamil sentences, the verb concludes, thus having the SOV structure. The following are the examples:

(i) ‘I took the book.’ The sentence exhibits SVO pattern.

(ii) ‘nAn puththakaththai eduththEn.’ The sentence parades SOV pattern.

As an example, in the phrase ‘I saw you, and I called you’, there are two main prosodic phrases. Prosodic phrasing involves finding meaningful phrases, which may or may not be explicit.

Prosodic phrasing is important because it increases the naturalness of the synthesized speech. It also helps to ascribe meaning to certain parts of the synthesized speech, in the same way as humans do, by varying the prosody. This process deals with creating prosodic boundaries at explicit identifiers. Tamil language sentences rely greatly on phrasing than the punctuation marks. Another approach is to use statistical models, with probabilistic phrase predictors like CART decision trees, to predict prosodic phrases based on some features. The features include the parts of speech of the surrounding words, the length of an utterance, the relative position of the word/sub word, the relative distance of a potential boundary, etc.

Even though the rules based on punctuation are the better predictors of prosodic phrases, there are many cases where explicit punctuation marks are not present to indicate the phrase boundaries. This problem is prominent in the case of Indian languages where there is little or no use of punctuation marks. POS tags lend their hands in these cases majorly.

2.4.2 Pitch Modeling

The prosodic boundaries are identified, and the speech synthesizer applies the prosody elements (namely, duration, intonation, and intensity) on each of the phrases. The primary factors of intonation are the context of words and the intended meaning of sentences. Consider the utterance ‘adhaip pArththEn’, which is a simple Tamil sentence which exhibits the pattern SV {S – ‘adhai’ and V – ‘pArththEn’}. There exists a single prosodic phrase. Still, there are many prosodic variants of this single sentence. For example, we might stress on the subject ‘adhai’ (in the concept of ‘seen which’).

Sometimes the verb ‘pArththEn’ needs a stress (in the concept of ‘whether seen or not’). Here we observe different pitch contours for each of the sentences, namely declarative, interrogative, imperative and exclamatory. Intonation is based on the gender, physical state, emotional state and attitude of the speaker.

There are two approaches for the automatic generation of pitch patterns:

1. Superpositional Approach
2. Linear Approach

The fundamental frequency of the speech unit is called as the f0 contour. It is the acoustic representation of the minimal utterance of any speech unit. The superpositional approach considers an f0 contour as consisting of two or more superimposed components. Modestly representing f0 alone is highly insufficient in describing the utterance of a speech unit. In this approach, the f0 contour is the sum of a global component g(f0) that represents the intonation of the whole utterance and the local components $l(f0_1,f0_2....f0_n)$ that model the change of f0 over the accented syllables.

Therefore, the following equation represents the f0:

$$f0 = g(f0) + l(f0_1) + l(f0_2) + ... + l(f0_n)$$

The linear approach considers an f0 contour as a linear succession of tones. An example of the linear approach to pitch modeling (Silverman et al. 1992) is the Pierrehumbert or ToBI (Tones and Break Indices) model that describes a pitch contour regarding the pitch accents. Pitch accents occur on the stressed syllables. They form characteristic patterns in the pitch contour. Pitch contours differ from each other in the dialects of the same language.

Table 2.1 ToBI Transcription for 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

The ToBI model for Original English uses five-pitch accents obtained by combining two simple tones, the high (H) and the low (L) in different ways. The model uses an H+L pattern that indicates a fall, an L+H pattern that describes a rise and an asterisk (*) to indicate which tone falls on a stressed syllable. The five-pitch accents are shown in Table 2.1.

2.4.3 Duration Modeling

The duration of speech units in continuous speech can sometimes become as short as half their time period when spoken in isolation. It is usually indicated in milliseconds. The Tamil language has its own, exclusive unit of time called ‘maathirai.’ A maathirai is the time taken for the blink of a human eye. Duration of a speech unit depends on factors such as:

(i) The characteristics peculiar to the speech unit,

(ii) The influence of the adjacent units and

(iii) The number of speech units.

The duration can also be a function of the sentence context. Duration modeling is essential in a TTS system. Because there is a need to generate the speech units with appropriate time spans to produce natural sounding synthetic speech.

Several methods are available for duration modeling. O'Shaughnessy, 1984 proposed an approach, where the intrinsic duration of a speech unit is modified by successively applying rules derived from analysis of the speech data.

In another approach, huge speech corpora are first analyzed by varying some possible control factors simultaneously to obtain duration models, such as an additive duration model by Kaiki et al (1990), CARTs by Riley (1992). The CARTs (classification and regression trees) proposed by Riley are data-driven models. The CART algorithm uses binary yes/no questions about the attributes that the instances have. Starting at the root node, the CART algorithm builds a tree structure, selecting the best element and binary queries asked at each node. The selection depends on the binary answer for the question. Figure 2.1 shows a sample CART.

Riley uses a set of the following factors in building the tree: segment context (three segments to left, segment to predict, three segments to right), stress, lexical position and phrasing position (word count from start and end of a phrase).

2.4.4 Intensity Modeling

Intensity modeling is not done intrinsically. It is either neglected or modeled along with intonation. Because intensity features are implied in intonation itself. Lee et al. (1998) uses Artificial Neural Networks (ANNs) to predict syllable energy. Sreenivasa Rao & Yegnanarayana (2009) proposed an

Intonation model which included Intensity modeling concerns. Deriving the prosody automatically from the given text is a tedious task. It cannot be performed using any precise algorithm. Prosody modelling is investigated over the years, but still found difficult. Some manual works like phrasing, labelling, and tagging have to be carried out by linguists.

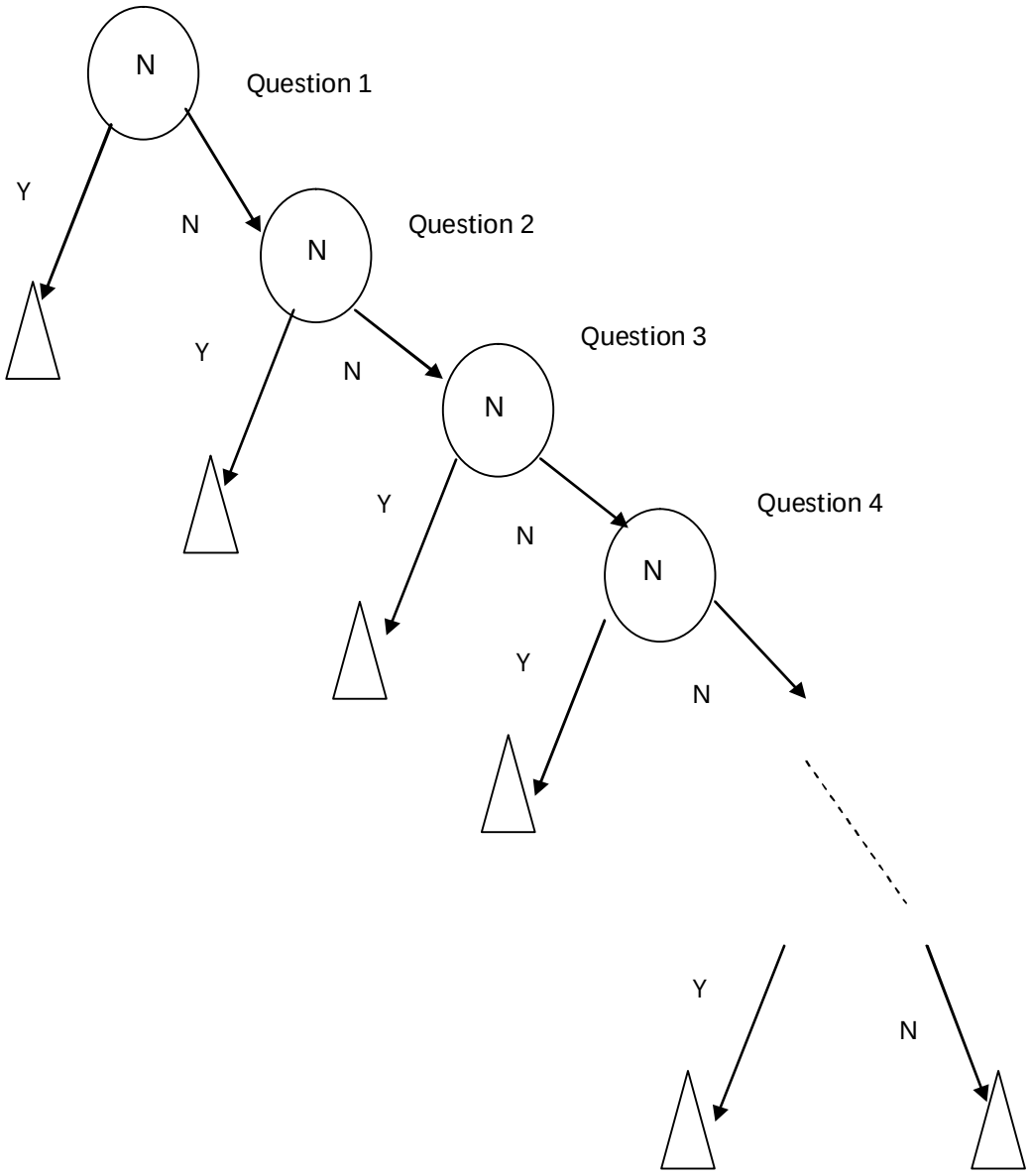

Figure 2.1 An example CART tree

It requires the analysis of large amounts of speech data and expert knowledge for prosody modeling.

2.5 CONCATENATIVE SPEECH SYNTHESIS

Concatenative speech synthesis is one of the most trusted techniques for Dravidian language TTS systems, thanks to its ability to produce human-like, natural sounding speech. In concatenative speech synthesis, a waveform is generated by selecting and concatenating the appropriate units from a database.

The quality of continuity of acoustic features at the concatenation points and the availability of units with appropriate prosodic features in the database plays a vital role in delivering a good quality synthesized speech. For content – specific or application – specific speech synthesis, large units such as phrases or sentences are stored and used. The quality (intelligibility and naturalness) of synthesized speech is better in this case. There is always a trade-off between the sizes of the speech units with the restriction of the text. If small units such as phonemes find their place, a limitless text can be synthesized, but with a poor speech quality. Two major problems exist in concatenating the speech units to produce a sentence. With the duration of speech units in continuous speech being as short as half their period of articulation when spoken in isolation, simple concatenation of speech units makes synthesized speech sound slow.

The second problem is that the sentence stress pattern, rhythm, and intonation are artificial if speech units derived from the inappropriate contexts are used. For example a */t/* before an */a/* sounds very different from a */t/* before an */s/*. Jurafsky & Martin (2000) have explained this paradigm at length.

To resolve such problems, synthesis methods using speech units like a diphone are employed. Diphones provide a balance between context dependency and size (typically 1000-2000 units in a language). Even though the diphone-based synthesizers produce appreciable quality speech, the pitch and duration of each phone in the concatenated waveform does not correspond to the desired prosody. Several signal processing techniques are developed to date, for improving the prosody in concatenative speech synthesis. While signal processing and diphone concatenation can produce appreciable quality speech, the results are often not ideal - the primary reason being that a single example of each diphone or speech unit is not enough. Unit selection based concatenative synthesis is an attempt to address these issues by collecting several examples of each speech unit at different pitches, durations, and linguistic situations, so that each speech unit is close to the target in the first place, hence requires less signal processing.

2.5.1 Pitch and Duration Modification

Even though diphone synthesizers produce an appreciable quality speech waveform, in many cases the pitch and duration of the speech units from database need to be modified to the pitch and duration required for proper sounding synthetic speech. Two known signal processing techniques used for concatenative speech synthesis to improve pitch and intensity of synthesized speech are the Time Domain Pitch Synchronous Overlap-add (TD-PSOLA) method and the Harmonic plus Noise Model (HNM) method.

2.5.2 Unit Selection Synthesis

Even though speech synthesis by concatenation of sub-word units like diphones produces clear speech, it does not have naturalness mainly because each diphone has only a single example. First of all, signal processing

inevitably incurs distortion, and the quality of speech gets worse when large amounts of the pitch and duration stretched.

Furthermore, there are many other subtle effects which are out of the scope of most signal processing algorithms. For instance, the amount of vocal effort decreases over time at the spoken utterance, producing weaker speech at the end of the speech phrase. Diphones taken from near the start of an utterance sounds unnatural in phrase-final positions.

Unit selection synthesis is an attempt to address this problem by using a huge corpus with a variable number of units for a particular class. The aspiration of this method is to pick up the best string of units from all the possibilities in the tagged/ untagged corpus and concatenate them to produce the ultimate speech. By selecting units closest to the target, the extent of signal processing required to produce prosodic characteristics is reduced and thus minimize distortion of the synthesized waveforms.

The unit selection depends on two cost functions. Roughly the Total Cost C_{Total} is given by the following equation:

$$C_{Total} = C_t + C_c$$

The target cost, c_t (u_i, t_i), is an estimate of the difference between a database unit, u_i, and the target, t_i, which it is supposed to represent. The concatenation cost, $c_c(u_{i-1}, u_i)$, is an estimate of the quality of a join between successive units (u_{i-1}) and (u_i). The unit that minimizes the total cost is selected. Hunt & Black (1996) has suggested this paradigm. There are two different unit selection techniques used:

(i) Unit selection used in the CHATR synthesis system

The first stages of synthesis of the CHATR system transform the input text into a target specification. These transformed acoustic signals then append with prosodic features. The speech database containing the candidate units is a state transition network with each unit in the database being represented by a separate state. The task of picking the best set of units is performed using the Viterbi algorithm in a similar way to HMM speech recognition. Here the target cost is the observation probability, and the concatenation cost is the transition probability.

(ii) Unit selection used in the Festival synthesis system

The Festival synthesis system is designed by Black Taylor & Caley (1998). It uses a cluster unit selection technique for selecting speech units from a speech database. In this method, the speech inventory is divided into clusters. Each cluster possesses units of the same grammatic class, based on their lexical, syntactic, and prosodic contexts. During synthesis, the appropriate cluster is selected for a target speech unit, offering a small set of candidate units. This process is equivalent to finding the target speech units with lowest target cost (C_t) as described in the previous section. Some systems then use signal processing to make sure the prosody matches the target while others simply concatenate the units.

2.6 SUMMARY

In this chapter, we have seen the basic Natural Language Generation and Natural Language Understanding constructs. We have discussed briefly the works done by the predecessor research fellows. A survey that has shown the current works done in the field is also presented. Here we have seen various approaches to prosody modeling. These techniques are primarily divided into rule-based approaches or data-driven

approaches. Data-driven approaches produce better results than their counter-parts.

We have attempted to build a TTS system that requires minimal prosody Modeling. If finding a speech unit that intrinsically has excellent prosody features is successful, then a minimal Prosody modeling becomes feasible. The unit selection technique is adapted. Nowadays it is affordable to have speech repositories with a huge number of feasible speech units of a language in different prosodic contexts. Therefore, memory constraints are not bigger issues in the design of such a synthesizer.

The Festival speech synthesis framework allows the development of such a system using its incorporated unit selection algorithm. In the next chapter, issues in building speech synthesizers for embedded devices (Distributed Speech Synthesis) are discussed.