CHAPTER 4

TEXT-TO-SPEECH SYNTHESIS WITH AUTOMATIC GENERATION OF SPEECH UNITS

4.1 INTRODUCTION

In this chapter, we are going to discuss in detail the Festival TTS framework (Black et al. 1998) and the automatic generation of speech units. Festival is a free, language independent speech synthesis engine. It includes modules for text processing, linguistic/prosodic processing, and waveform generation. The Festival software allows any of these modules to be rewritten. In this thesis, we focus on a new automatically generated, syllable units and the existing diphone units.

In this chapter, we describe two synthesis techniques supported by the Festival framework namely, diphone synthesis and unit selection synthesis. We have also explained the architecture of Festival software that supports these two techniques.

4.2 DIPHONE SYNTHESIS

In this section, we study the existing diphone synthesis procedure in Festival concerning a speech synthesizer for Indian languages developed using this paradigm. Diphones provide a trade-off between capturing co-articulation effects, minimizing discontinuities at concatenation points and being relatively small in number. The comparative study table concerning

various speech units is given in Appendix 1. A diphone consists of two connected half phones and captures (Kiruthiga S and Krishnamoorthy K, 2012a) the transition between two phones by starting in the middle of the first phone and ending in the halfway point of the second one. Figure 4.1 gives the structure of a diphone.

Figure 4.1 Transition between two phones in a diphone unit

As there are two important synthesis techniques available with Festival, we briefly describe the architecture of the Festival speech synthesis framework given by Black et al. (1998). Figure 4.2 shows the architecture of the diphone synthesis based TTS system in the Festival software.

Prosodic phrasing, segmental duration generation, and the F0 contour generation modules carry out Prosody Modeling. Specifications are derived from these prosodic models. Units are selected from the database and are modified based on these models. Diphone synthesis involves identifying a non-redundant and consistent set of diphones that cover the language. Sparse entries are cleared out of the database. This consistent inventory also includes information such as whether the diphone contains a vowel or consonant, the place and manner of articulation for consonants and vowel length. The identified diphones are embedded in carrier words to ensure that they are pronounced clearly and consistently. Utterances of carrier words are then recorded, tagged and labeled.

The unit selection module and CART based cluster database replaces these modules in the unit selection based synthesis technique.

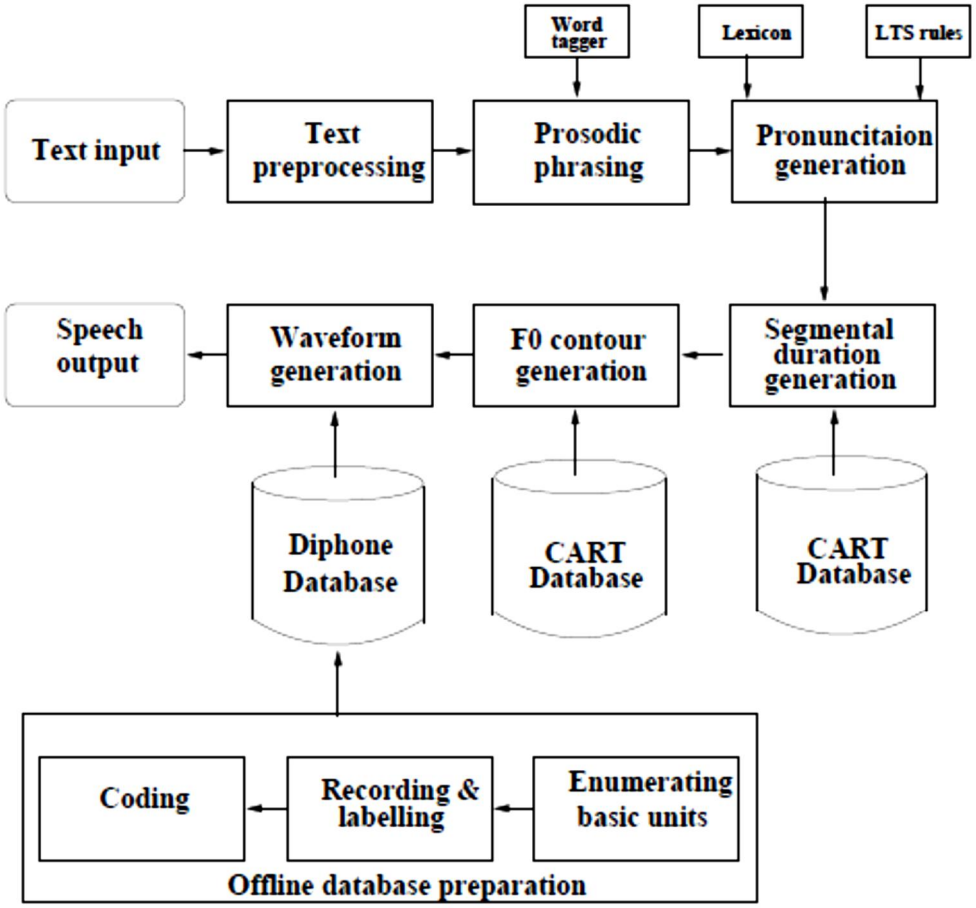

Figure 4.2 Block diagram of TTS system using diphone synthesis in the Festival speech synthesis framework

The system design has also tagged the distance information for phrase break prediction. The concept of 'morpheme tag' is also discussed for Prosody modeling in Indian Languages (Samuel Thomas et al. 2007). The classification and regression tree (CART) based duration models were used for segmental duration prediction for Dravidian Languages. In diphone-based synthesis, signal processing techniques model duration and pitch. Festival employs residual Excited Linear Predictive Coding (LPC) based re-synthesis. These algorithms for modification are pitch-synchronous techniques and hence require information about pitch periods. Pitch marks are extracted using the 'pitch mark' program. After obtaining the pitch marks, 'sig2fv' and 'sigfilter' programs extract pitch-synchronous LPC parameters and residuals.

The diphone database is finally coded into the intermediate form required by the waveform synthesizers. The UniSyn synthesizer and the OGI diphone synthesizer are those waveform generation artefacts available with Festival.

TTS systems based on diphone synthesis need prosodic models to produce good speech output. The prosodic analysis for these models requires a database of speech tagged with linguistic and prosodic information. Tools are also needed to generate appropriate syntactic information essential to predict prosody from the text. Automatic prediction of Prosody is an important task and feed forward methodology is adapted to carry out the same.

4.3 UNIT SELECTION SYNTHESIS

Unit selection synthesis technique selects the best string of speech units from a speech corpus and concatenates them to generate speech. These selected speech units should satisfy the following two constraints (Vikram et al. 2010).

(i) They should best match the target specification given by the linguistic components of the text analysis module and

(ii) They must be the best units that join smoothly when concatenated.

The cost associated with the first constraint is called the target cost. The cost associated with the second constraint is known as the concatenation or join cost. The target specification is a sequence of speech units along with features related to the phonetic and prosodic context for each syntactic unit (Kishore & Black 2003). The phonetic context features include the identity of a particular syntactic unit, the relative position of the speech unit in the given

word and the phonetic parameters of the adjacent speech units. The prosodic features include the pitch, duration, and stress of the individual syntactic term and the prosodic parameters of the preceding and following units. Similar information associates with each unit of the speech database.

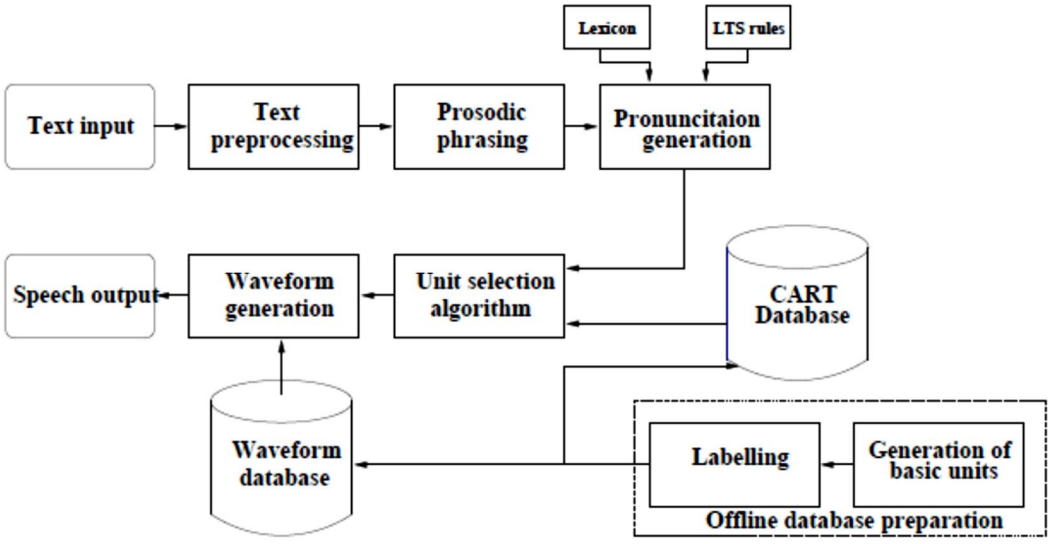

Figure 4.3 Architecture of TTS system using unit selection based synthesis in Festival

Festival uses a clustering technique designed by Black & Taylor (1997) to organize the units in the speech database. The organization is according to the phonetic and prosodic context of individual elements. Figure 4.3 clearly shows the Architecture of TTS system using Unit Selection based synthesis. It gets a Text input, processes the input in various phases and gives out a Speech output. The system handles the speech as simple units. For example, if there is a speech unit */bi/*, the algorithm clusters all the instances of the speech unit */bi/* with different phonetic and prosodic contexts into the same class. Each category organizes itself as a decision tree. The leaves of the binary tree are the various instances of the same speech unit. The branches of the tree are queries based on various features that describe those syntactic units. Figure 4.4 shows the decision tree that clusters the various entries for

the entity *'/bi/.'* During synthesis phase, for each unit to be synthesized, its decision tree is identified from the speech database. The synthesizer performs a search starting from the root node of the decision tree to attain the leaf node. Queries are thrown based on the target specification of the speech unit at each node. A set of candidate units that best match the target specification is derived at the leaf node.

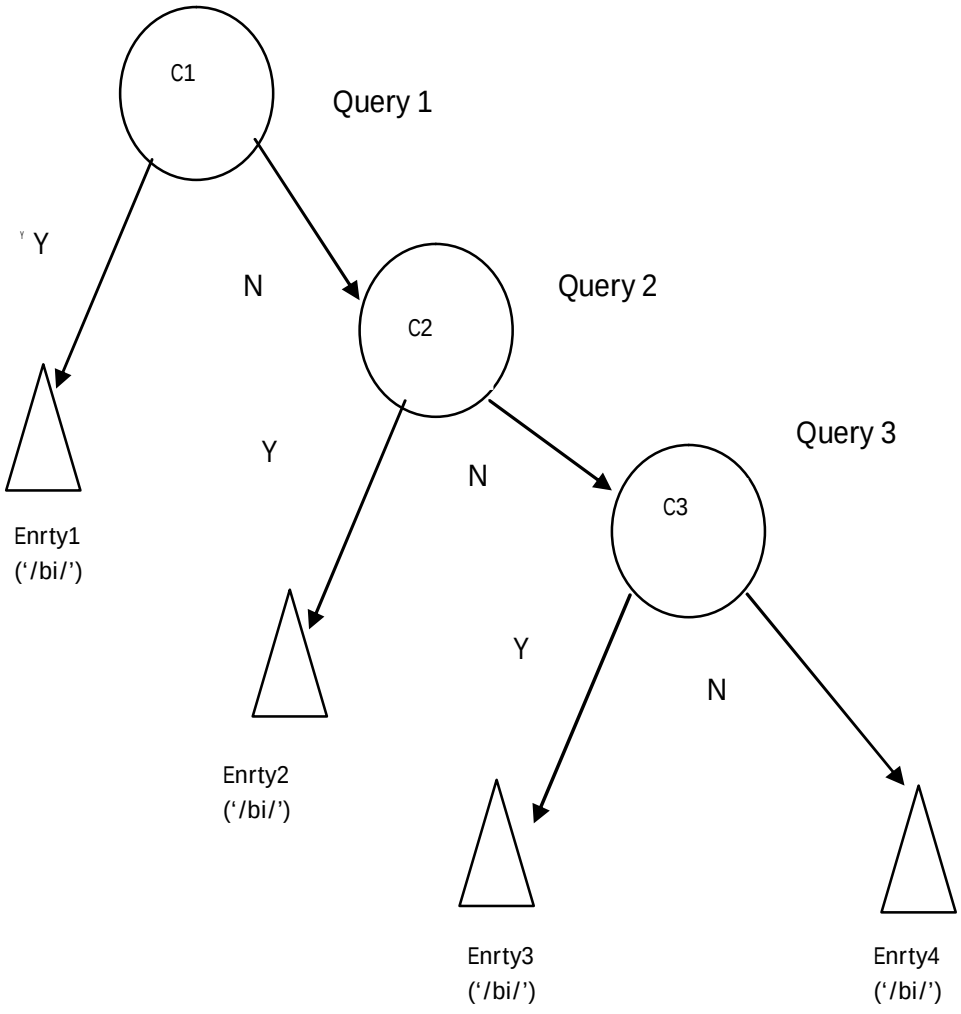

Figure 4.4 Decision tree that clusters various entries of the speech unit *'/bi/'*

A function T_{dist} (U) is the distance of a unit U to its cluster center (equivalent to finding the target cost). Another function J_{dist} (U_i,U_{i-1}) is also

defined to find the join cost between a candidate unit U_i and the previous candidate unit U_{i-1}.

4.4 MANUAL EXTRACTION OF SYLLABLES FROM TAMIL WORDS

Manual syllabification of a Tamil word *'/arasAngam/'* is illustrated as follows:

'/ar/' '/as/' '/Ang/'and '/gam/'

VC VC VC CVC

Table 4.1 Details of the manual syllabification of the word 'arasangam'

S.No.	Syllable	Onset	Rhyme
1	*'/tha/'*	*'/th/'*	*'/a/'*
2	*'/mi/'*	*'/m/'*	*'/i/'*
3	*'/zha/'*	*'/zh/'*	*'/a/'*
4	*'/ga/'*	*'/g/'*	*'/a/'*
5	*'/m/'*	-	*'/m/'*

Table 4.1 gives the details of the manual syllabification process. Each syllable is divided into its respective codas and rhymes.

4.5 SEGMENTATION USING MINIMUM PHASE GROUP DELAY FUNCTIONS

The Short-term Energy (STE) of a speech signal is a simple time domain method of processing and characterizing important features of it (Samuel Thomas 2007). Syllables are typically of the form C_VC_ (C: consonant, V: vowel). Regarding the STE function, this is characterized by

regions with high energy at the centre and energy reduces at both the ends. Based on this phenomenon, a group delay based segmentation algorithm is given by Prasad (2002). Using this algorithm, segmenting the speech signals at their minimum energy points generates units that have the syllable structure. The group delay based algorithm can locate polysyllables by adjusting the so-called 'window scale factor (WSF)'.

As an example, consider the segmentation of the speech signal for an utterance of the Tamil phrase */peNgaLE nAttin kaNgaL AvArgaL/*

The boundaries are for the following units: */pen/, /ga/, /LE/, /nAt/, /tin/, /kaN/, /gaL/, /Av/, /Ar/, /gaL/,* where

(i) */pen/, /nAt/, /kaN/, /kaL/,* and */tin/* are CVC units.

(ii) */Av/, /Ar/* are VC units.

(iii) */ga/, /LE/* are CV units.

Labels are assigned to these units based on the syllabification rules. The above illustration vivaciously shows that the Tamil language has more C_VC_ class syllables.

The group delay based segmentation algorithm sometimes may identify boundaries at semivowel regions, which are otherwise perilous to extract automatically. An example of this can be seen as in the case of segmentation of speech unit */ga/* and */LE/* from the word */peNgaLE/*.

Tamil is a syllabic language (Vinodh Vishwanath et al. 2010) which contains 12 vowels, 18 consonants and a special morpheme called 'Ayutha Ezhuthu' in its script. The language has certain well-defined rules which introduce seven other phones depending on the presence of consonants

on the vowels or the other consonants. Hence, there are 39 phones in the language. The language is said to be agglutinative in nature. The language has morphemes affixed to the roots of the individual words. This is a common property exhibited by some Dravidian languages and languages like Turkish, Estonian and Japanese.

4.6 SUMMARY

In this chapter, we have studied the various speech synthesis techniques available with the Festival software. We have also discussed the group delay based segmentation algorithm used for generating speech units for concatenative speech synthesis. In diphone synthesis, the quality of synthesized speech is dependent on the intrinsic prosodic rules that are present in the synthesizer. If we can rewrite these rules, we can generate appropriate prosodic specifications. As we employ diphones for only special places of articulation, the intrinsic prosody model is sufficient. In the unit selection method, the quality of synthetic speech relies on the number and quality of the available units present in the database. For good quality synthesis, all possible units of the language should be present in the corpus. The group delay based segmentation algorithm is not only consistent but also produces decent results that are close to manual segmentation. In the next chapter, let us look at the Architecture of the proposed system.