CHAPTER 6

DESIGN OF THE SPEECH DATABASES FOR THE TEXT TO SPEECH SYNTHESIS IN TAMIL

6.1 INTRODUCTION

In this thesis, the previous chapter vivaciously explained the functions of group delay based segmentation algorithm and the Festival framework. Those concepts form the basis for the unit selection synthesizers developed for the Tamil language. In this chapter, we shall discuss the steps followed to build a synthesizer along with the evaluations of the speech units. The results of the assessments show the effectiveness of the speech units and the appropriateness of the synthesis technique.

6.2 GENERATING POLYSYLLABLE UNITS USING GROUP DELAY BASED SEGMENTATION ALGORITHM

As we have made a survey of the design issues for Indian language TTS systems, we were in need of testing the efficiency of various speech units (such as diphones, syllables, words, polysyllables, etc.) to be employed. The group delay based segmentation algorithm proposed by Prasad (2002) helped in evaluation of various speech units in implementation aspects. The algorithm introduces N_c, which is the length of the STE (Short Term Energy) sequence. By varying the values of N_c, the system can produce units of various sizes. When the value of N_c is high, smaller units are extracted, because of the increased resolution of the segmentation process. If N_c is made low, the resolution decreases and larger units result. Therefore, to generate polysyllables, we lower the value of N_c. The efficiency and applications of the

diphone synthesis is a known phenomenon. So we have evaluated various syllable units and tabulated in Table 6.1.

Table 6.1 Various speech units studied and evaluated using Group Delay based Segmentation Algorithm

S.No.	WSF	Monosyllables	Bisyllables	Trisyllables
1	4	103	22	2
2	6	85	24	3
3	8	62	15	10
4	10	22	21	16

For example, various speech units are extracted from the Tamil word */tholkAppiyam/*.

(a) For WSF = 4, boundaries of the monosyllables */thol/*, */kAp/*, */piy/*, and */am/* are identified

(b) For WSF = 8, boundaries of two bisyllables */tholkAp/*, and */piyam/* are identified

(c) For WSF = 15, boundary of one trisyllable */tholkApiy/* and one monosyllable */am/* are identified.

On varying the WSF factor from 4 to 15, mono, bi-, and tri-syllables are automatically generated by the system. The quality of the synthesized speech generated using unit selection based synthesis proportionately depends on the available units in the speech corpus. For good quality output, all the units of the language must be present. Moreover, the units should be generic so that they can result in an unrestricted text-to-speech synthesis. The units should also be correctly labeled. They should belong to one of the syllable groups - CV, VC or CVC or a combination of these

groups. This phenomenon aids the definition and use of well-defined letter-to-sound rules during the pronunciation generation phase of the speech synthesizer. At times, units that do not fall into syllable categories of CV, VC or CVC are also obtained. Since these units are present in speech synthesis, the following steps are performed to extract the occurrence of such units:

- Word level segmentation is performed manually. The automatic segmentation algorithm then runs on a speech repository of words. This post-process ensures that units that span across words are separated.

- Those separated units are then stored in diphone repository.

- The speech repository of words is carefully listened to and correctly labeled with appropriate transliterated English text.

- A text segmentation algorithm then runs on each of the tagged labels. This process is to identify the syllables that are present in the corresponding speech data. Those tiny units (generally diphones in Tamil) that occur during segmentation are branded separately.

For example, if the text segmentation algorithm runs on the text:

pAdalaik kEL

It would identify three syllables - *'/pAd/'*, *'/al/'*, */aik/*, and */kEL/* {CVC, VC, VC, CVC patterns.} Now the unit *'/kEL/'*, is the end-of-sentence unit. The algorithm divides the phrase into C and VC pattern as *'/k/'* and *'/EL/'*. These two units are diphones and are stored in the diphone corpus.

The number of syllable units identified by the text segmentation algorithm ($|n_t|$) is now compared with the number of syllable-like units (which also includes units divided as diphones) generated by the group delay based segmentation algorithm, say $|n_s|$. If the algorithm identifies more units than expected for each word (i.e, if $|n_s| > |n_t|$), a set of generated syllables with cardinality equal to the number of expected syllables ($|n_t|$) is selected. Those candidate syllables are the syllables that occur as the first $|n_t|$ (sorted in descending order) units from the list of $|n_s|$ units, assuming that the units with longer duration would be the suitable syllable units. The algorithm checks the remaining ($|n_s - n_t|$) units for their relative occurrence in the sentence phrase. If any of those units (i.e., ($|n_s - n_t|$)) occur at the end of the sentence or in special words, then they are processed as diphones and are stored in diphone corpus.

6.3 CREATING SAMPLE SYLLABLE AND DIPHONE DATABASES

Since the entire procedure of enumerating syllable units and using them to create a unit selection synthesizer using the Festival framework exists, sample syllable and diphone databases establish the panorama. The sample databases are created using speech recordings of an existing bulletin news database, called DBIL. The bulletin news is recorded by a native Tamil speaker in the laboratory environment.

The following steps are performed to create the sample database:

- Create the audio file of the bulletin news sentences with a native Tamil reader

- Phrase the audio file into sentences, phrases, and words.

- Generate all possible diphones (including allophones); check for sparse usage and prune unwanted diphones.

- Run the group delay based segmentation algorithm on the phrased audio file.

- With WSF = 4, more number of syllables are extracted from the phrased audio file.

- Verify that the extracted units are syllables.

- Run the group delay based segmentation algorithm on the larger units till the syllables and possible diphones are extracted.

- Integrate the syllables and diphones generated for the sample (individual syllable and diphone) databases into the Festival framework.

Speech database in the Festival framework consists of three components (A. W. Black et al. 1998):

1. a dictionary file,
2. a set of waveform files, and
3. a set of pitch mark files.

The dictionary file consists of one entry per line. Each line has five fields:

(i) Text form of the speech unit in the pattern P_1-P_2..P_n. Where P is the individual phoneme in the speech unit. The index ‘n’ extends to the

number of phones. For example, n=2 for diphones. The value of ‘n’ cannot take a defined value for syllables and polysyllables.

(ii) Name of the Audio file

(iii) Start position of the speech unit in milliseconds

(iv) Mid position of the speech unit in milliseconds

(v) End position of the speech unit in milliseconds

Ex: *‘/tha/’* is stored in the database as:

```
th-a t21 612.034 663.008 690.124
```

So, the database takes the following schema.

```
t-a t32 823.034 846.008 890.234

ch-a s21 412.035 463.009 518.23

p-a p21 612.034 663.008 690.124

m-i m48 356.814 403.54 437.522

r-i t21 612.034 663.008 690.124
```

Waveform files are in any of the audio file forms, the speech tool or the hardware supports. They may be standard linear PCM waveform files in the case of PSOLA. Pitch mark files consist of a simple list of positions in milliseconds in the order of the switching of intensity.

6.4 EVALUATION OF THE SYLLABLE AND DIPHONE UNITS

The syllable-like speech units generated for the sample database are then integrated with the Festival framework. As described in Chapter 3, Festival uses a selection criterion with two costs, namely

(a) Target cost C_t

(b) Concatenation cost C_c

Festival selects the unit that minimizes both the costs.

The sample speech synthesizer for Tamil evaluates and verifies whether combining the syllable and diphone units is indeed a good technique for use in concatenative speech synthesis. Based on the conclusions drawn from this evaluation intelligible speech synthesizers are designed for Tamil.

6.4.1 General Evaluation of the Speech Units

In order to test the improvement of synthesized speech quality using syllables and diphones, a perceptual evaluation of 5 sets of synthesized Tamil sentences is conducted using 5 native Tamil readers. Each set had 4 sentences synthesized using different methods: the first sentence in each set is synthesized using a diphone synthesizer; the second sentence was synthesized using syllables only; the third sentence was synthesized using both syllables and diphones. The sample sentence used was:

'/vAzhga vaLamudan/'

a) Diphones: {/v-A/ */zh-g/ /a/*} { */v-a/ /L-a/ /m-u/ /d-a/ /n/*}

b) Syllables: */vAzh/ /ga/ /va/ /La/ /mu/ /dan/*

CVC CV CV CV CV CVC

c) Syllables and diphones: */vAzh/ /ga/ /va/ /La/ /mud/ {/a-n/}*

CVC CV CV CV CVC Diphone

The speakers were asked to score the naturalness of each output on a scale from 1 to 5 (1=Bad, 2=Poor, 3=Fair, 4=Good 5=Excellent) (M. Nageshwara Rao et al. 2005). Table 6.2 gives the Mean Opinion Score (MOS) for each of the synthesized speech.

Table 6.2 MOS for sentences synthesized using Diphones, Syllables, and dual database

S.No.	Speech unit used for Synthesis	Mean Opinion Score
1	Diphones	1.34
2	Syllables	1.47
3	Syllables and Diphones	3.97

The results of the first MOS test show that speech synthesis with syllables is better than other syntheses. The MOS for using syllables is marginally better than that for diphone synthesis. Though diphone synthesis gives a lesser MOS score, the database size may give it a winning edge. When the objective narrows down to the naturalness of synthesized speech, Syllable synthesis wins. Finally, it is vivacious from the scores that combining syllable and diphone units, when used, achieve good results, but without a significant increase in the number of units needed. Table 6.3 clearly shows that:

Table 6.3 Comparative number of units synthesized using Diphones, Syllables, and Dual database

No. of units synthesized	Diphones	Syllables	Dual database
1	*/v-A/*	*/vAzh/*	*vAzh*
2	*/z-h/*	*/ga/*	*ga*
3	*/g-a/*	*/va/*	*va*
4	*/v-a/*	*/La/*	*La*
5	*/L-a/*	*/mu/*	*Mud*
6	*/m-u/*	*/dan/*	*a-n*
7	*/d-a/*		
8	*/n/*		

6.4.2 Prosodic Evaluation of Syllables and Diphones

The naturalness of synthetic speech generated by the synthesizers which can use a dual database needs assessments. The first prosodic evaluation test was to find the amount of useful duration information intrinsically present within the syllables and the diphone units. The test is done by first synthesizing an arbitrary text with the sample synthesizer. The tested sentences are used to build a CART tree. A second CART tree is generated using recorded speech.

The CART trees built from both the synthesized and natural speech databases are then separately used as duration models for a diphone synthesizer. Then, the speech was synthesized using the different duration models and the subjects were asked to score the naturalness with the MOS. The first set of sentences is those synthesized with no duration modeling. The sentences of the second set are CART duration model generated using synthetic speech from the sample synthesizer. This prosodic evaluation shows

that the syllables intrinsically have sufficient duration information. All syllables in a syllable database have their own occurrences.

A second assessment evaluates the quality of the synthesized speech using the diphones. It is same as the test conducted for syllables. We have built two CART trees, one with the duration-modeled sentences, and the other with a recorded speech.

Another evaluation test estimates the overall perceptual quality of the synthesized speech using the dual database. Native Tamil speakers appraised the overall perceptual quality of the synthesized speech, based on the following aspects:–

(i) Intelligibility

(ii) Naturalness

(iii) Distortion

A natural sounding speech should score more values for the first two aspects and less value for the third aspect. The result of this test shown that around 80% of the speakers felt the synthesized speech was intelligible, natural sounding, and had low distortion.

This evaluation the speech synthesizers with dual databases perform well even with simple prosodic models. However, because some of the concatenated speech units were not appropriate, distortions may appear. There are no dedicated Tones and Break Indices tabulated for the Tamil Language. So English ToBI guidelines (Silverman et al. 1992) are employed for intonation modeling. This will yield a more natural speech. Intensity modeling is not necessary because articulation points which may need an intensity deviation are addressed using diphones.

A good quality of the synthesis using the dual-database technique results because syllables have more prosodic and acoustic information and fewer discontinuities when compared to other synthesis procedures using phones or polysyllables. The diphones can also cater for the prosodic necessities wherever they are employed. The boundaries of the syllables correspond to low energy regions, which lead to minimum coarticulation points. Therefore they are preferable for concatenative waveform synthesis. Figure 6.1 shows the speech waveform for the Tamil phrase */vAzhga vaLamudan./*

The spectrogram clearly shows that the syllable corpus, when used, gives out an unnatural ending. But, when combined with the diphone units, the output is natural, as diphones replace some distortions in the end-syllables. Moreover, spectral changes are uniform across the other syllable boundaries. It vividly shows that the dual database technique used for concatenative speech synthesis ends in the built of a good Speech Synthesizer. The number of concatenation points for syllables is also comparatively very less for syllables and diphones. When duration modeling is carried out using CART technique and an appropriate ToBI standard is followed, the synthesizer ends with good results.

Phrase boundaries grant key contributions in fluently connected speech. Tamil languages lack punctuations in general. Phrase boundaries and intra phrase prosodic patterns used in current scenario help in understanding an utterance. A set of 21,072 unique words was synthesized to find out the most frequently occurring realizations. The sparse entries were then deleted from the speech repository. This process reduces the size and redundancy of the speech repository. Still, there are some instances where the speech quality deteriorates. Tamil has an average 12 vowels and 23 consonants (18 consonants + 1 Special vowel = 19 consonants. When ha, sha, jha and shree

includes to the list, there are 23 consonants in the language). These morphemes can be combined to form close to 3400 syllable units of $C^{*}VC^{*}$ pattern. Thus, even though the syllable is a suitable candidate for unrestricted speech synthesis, a good synthetic speech will not be possible until a diphone database assists the syllable database. The next section discusses how syllable and diphone units can be used for unrestricted speech synthesis.

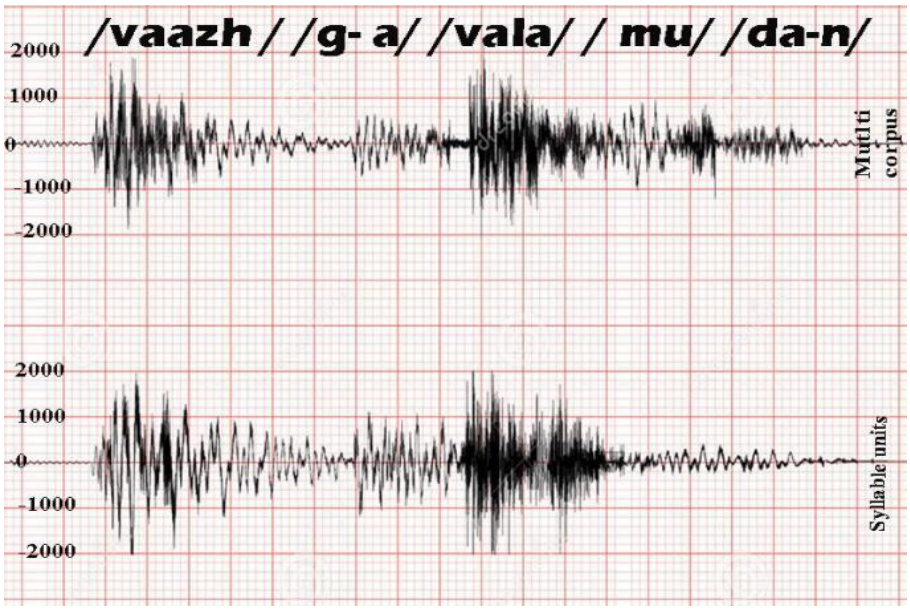

Figure 6.1 The speech waveform for the phrase *'/vAzhga vaLamudan/.'*

6.5 UNRESTRICTED SPEECH SYNTHESIS

Figure 6.1 shows the speech waveform for the Tamil phrase */vAzhga vaLamudan./'* Three syllables which make the up the sentence - *'/vAzh/'*, *'/ga/'*, *'/va/'*, *'/la/'*, *'/mu/'* and *'/dan/'*. This example shows two aspects that need to be addressed while using syllable units as the basic speech units for synthesis. They are -

1. Each syllable unit needs to have a minimum number of realizations in the speech database.

2. There is a predetermined pause between each syllable.

3. The end-syllable of a sentence needs to be adjusted with its previous syllable so that it gives space to the generation of diphones at the sentence termination. The adjustment should be made such that the nucleus of the previous syllable is not disturbed.

To restrict the number of irrelevant syllables present in the database, we need to find out the number of realizations of each syllable. A set of 500 Tamil sentences was labeled at the syllable level to analyze the fundamental frequencies of different realizations of syllable and diphone units.

Table 6.4 Predicting realizations 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%

Table 6.4 shows the morphemes occurring in different word positions (beginning, middle and end) in the Tamil database. With reference to the position of the syllables, there are some realizations concerns to the same syllable.

This phenomenon is because the formant frequencies of the syllables start decreasing as the position shifts from the middle to the end. These characteristics arise in most of the monosyllable units. It is hence evident that a minimum of three realizations - one realization occurring at the word beginning position, one at the middle of a word and another at the end of the word.

6.6 SUMMARY

In this chapter, we have discussed the issues concerning the design of the speech databases for the text to speech synthesis in Tamil. We have discussed the manual segmentation of syllables and the algorithm for automatic generation of syllables using Group Delay based Segmentation Algorithm. We have also thrown some light upon the unrestricted Speech access. In the next chapter, let us discuss modeling and evaluation of prosody features.