CHAPTER 5
SYLLABLE BASED MODELS

5.1 INTRODUCTION

A CD phone or triphone has been the dominating sub-word unit in speech recognition for the past one decade. Most of the commercial and laboratory LVCSR systems use triphones as the fundamental acoustic unit. The same also holds good even for agglutinative languages like Tamil, as it has been demonstrated by experiments described in chapter 4. A medium vocabulary triphone based continuous speech recognition system for Tamil language was built and a WER of 9.44% was obtained. The triphone based system has been taken as the baseline system for comparison with the proposed systems described in this chapter and the next chapter.

There are two major issues in building LVCSR for Tamil. Firstly, the grammar of Tamil unlike English is agglutinative in nature. Secondly, Tamil also exhibits occurrences of morphophonology. These issues result in a large number of unique words with the same lexical root and new compound words in the vocabulary. The increase in words in the vocabulary results in the increase of models which cannot be sufficiently modeled with the given training set. These aforesaid issues also result in OOV words in the test set.

This problem has been addressed by Saraswathi and Geetha (2007) by employing a morpheme based language model in speech recognition. They have demonstrated that the morpheme based language model outperforms
other statistical language models like trigram, distance based bigram and trigram, dependency based models and class based models.

In this chapter, an alternative method which uses prosodic syllable as a sub-word unit in the acoustic model is suggested. The identification of prosodic syllables is done by an algorithm which leverages linguistic rules of Tamil to segment words into prosodic syllables. Earlier works have been focused only phonological syllables rather than prosodic syllables. In a language like English, it is difficult to segment text into phonological syllabic components due to complex syllable-initial and syllable-final consonant clusters and vowel-type-consonant dependencies. On the contrary, Tamil is syllable-timed in nature and has well-defined linguistic rules to access prosodic syllable components from the text.

5.2 SYLLABLES IN TAMIL LANGUAGE

As already mentioned in section 1.3.1, the phonology of Tamil is characterized by an inventory of 18 consonants and 12 vowels (5 long and 5 short, 2 diphthongs). The basic syllable consonant-vowel phonotactics is characterized by a regular expression shown in equation 5.1.

\[ RE(S) = \{C\} \lor \{C \{C\}\} \]  

(5.1)

There are constraints on which consonants can appear in each of the three consonant positions and in combination with vowels. With no constraints, the maximum number of syllables will be \(18^3 \times 12 = 69,984\). However, because of constraints the actual number of possible syllables is in order of magnitude smaller. The number of lexically attested syllable is smaller still. In addition, there are constraints on stress patterning in Tamil words.
Properties that constitute prosody are fundamental frequency or formant $f_0$ (perceived pitch), duration, intensity (perceived loudness) and to some extent vowel quality. Prosodic properties of speech are also used in detection of word boundaries and in other higher tasks of speech understanding like encoding or decoding pragmatic differences like a statement vs. a question, emotion and so on. At the word level, prosodic properties encode lexical tone, lexical stress and lexical pitch accent.

5.2.1 Justification for using Prosodic Syllable as a Speech Unit

English has complex syntactical grammar but simple morphological structure. On the contrary, Tamil has complex morphological grammar but simple syntactical structure. In fact, in ancient times, Tamil was written continuously without word spaces, punctuations and word divisions. The following sentence in Tamil is composed of two words which exhibit complex inflectional morphology and morphophonology.

அங்கு சதுஸ்வ முயுலகரிடும் நந்து சுருங்க வசா குணன் கட்ட குர்ந்து தான் மற்றும்

The same sentence can be re-written in a simpler form without changing the meaning after resolving some inflectional morphology and morphophonology as follows:

அங்கு நந்து சதுஸ்வ முயுலகரிடும் நந்து சுருங்க வசா வாயில் குணன் கட்ட குர்ந்து தான் மற்றும்

(My brother realized that toddy is injurious)

From the example, it can be inferred that a word as a building block of a sentence loses its meaning owing to agglutination. Moreover, Tamil does not feature stress and pitch accent at word level (Remijsen, 2003). If these are true, then which element of Tamil grammar governs the prosody of a sentence? It is the prosodic syllable that plays a vital role in the production and perception of a phrase or a sentence.
There are two types of prosodic syllables namely *Ner-acai* and *Nirai-acai*. *Ner-acai* is monosyllabic. It may consist of either one short vowel or one long vowel, either of which may be open or closed, i.e. ending in a vowel or consonant(s) respectively. *Nirai-acai* is always disyllabic with an obligatorily short vowel at first position, while the second phoneme is unrestricted. Like *Ner-acai*, *Nirai-acai* may also be of open or closed type. The prosodic syllable representation can take any of the following eight patterns as shown in Table 5.1. An uninflected Tamil word may comprise one to four prosodic syllables.

### Table 5.1 The Linguistic Rules of Tamil Prosodic Syllables

<table>
<thead>
<tr>
<th>Description</th>
<th>Pattern</th>
<th>Example (with Romanized Tamil and meaning)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short vowel, long vowel followed by consonant(s)*, (<em>Nirai</em>)</td>
<td>SV + LV + C(s)</td>
<td>ழருநா (pulal) (meat)</td>
</tr>
<tr>
<td>Short vowel followed by a long vowel, (<em>Nirai</em>)</td>
<td>SV + LV</td>
<td>யின் (vizha) (function)</td>
</tr>
<tr>
<td>Two short vowels followed by consonant(s)*, (<em>Nirai</em>)</td>
<td>SV + SV + C(s)</td>
<td>யரு (kaLam) (field)</td>
</tr>
<tr>
<td>Two short vowels, (<em>Nirai</em>)</td>
<td>SV + SV</td>
<td>யரு (kala) (echo sound)</td>
</tr>
<tr>
<td>Short vowel followed by consonant(s)*, (<em>Ner</em>)</td>
<td>SV + C(s)</td>
<td>யரு (kal) (stone)</td>
</tr>
<tr>
<td>Long vowel followed by consonant(s)*, (<em>Ner</em>)</td>
<td>LV + C(s)</td>
<td>யரு (vaL) (sword)</td>
</tr>
<tr>
<td>Long vowel, (<em>Ner</em>)</td>
<td>LV</td>
<td>யரு (va) (come)</td>
</tr>
<tr>
<td>Short vowel, (<em>Ner</em>)</td>
<td>SV</td>
<td>ய (ka)</td>
</tr>
</tbody>
</table>

* At the maximum, two consonants can occur

### 5.3 FORMAL REPRESENTATIONS OF PROSODIC SYLLABLES

The FSA and FST have been used successfully in the implementation of phonological rules. Johnson (1972) and Kaplan and Kay (1981) had based their work on finite state models for phonological rules. The
two-level morphology introduced by Koskenniemi (1988) can be readily applied to implement phonological rules with an FST. Bird and Ellison (1994) worked on multi-tier auto-segmental phonology wherein each phonological tier is represented by an FSA. Falling in the same line of research, this section focuses on formal representation of Tamil prosodic syllables using FSA.

5.3.1 Regular Expression for Prosodic Syllable Patterns

Prosodic syllables are composed of phonemes. Based on the linguistic rules tabulated in Table 5.1, a regular expression can be formulated as shown in equation 5.2.

\[ RE(PS) = [SV](SV | LV)(C[C]) \]  \hspace{1cm} (5.2)

where \( SV \) is short vowel.
\( LV \) is long vowel.
and \( C \) is consonant.

This expression describes all the possible patterns of prosodic syllables. In other words, an optional short vowel is followed obligatorily by either a short vowel or a long vowel, and zero or one or two consonants.

5.3.2 Finite State Automata for Prosodic Syllable Patterns

In the next step, an \( \varepsilon \)-NFA can be deduced directly from the regular expression. Using structural induction technique, it is possible to prove that \( RE(PS) \) in equation 5.2 is equivalent to the \( \varepsilon \)-NFA shown in Figure 5.1. For the basis, an automaton corresponding to the regular expression of single symbol is created with an initial and an accepting state. In the inductive step, each primitive operation of regular expression – concatenation, union and closure – can be imitated by an automaton with \( \varepsilon \) transitions.
Figure 5.1 The $\varepsilon$-NFA for Tamil Prosodic Syllables

In Figure 5.1, it can be seen that the automaton is constructed by individual automaton for $[SV]$, $(SV \cup LV)$ and $[C[C]]$ joined by $\varepsilon$ transitions. With the help of formal methods like eliminating $\varepsilon$ transitions and subset construction, the $\varepsilon$-NFA is converted into an equivalent DFA. The DFA for the prosodic syllable patterns of Tamil is shown in Figure 5.2.

Figure 5.2 The DFA for Tamil Prosodic Syllables
The DFA of prosodic syllable patterns includes a dead state. A state is a dead state if it is not an accepting state and has no out-going transitions except to itself.

Theoretically, the number of prosodic syllables will be quite larger (of the order of 3,674,160), since there are 90 (18 times 5) short vowels, 126 (18 times 7) long vowels and 18 consonants. But, the actual number will be smaller due to constraints like phonotactics and morphotactics. Hence, it is essential to estimate the number of prosodic syllables with the help of a corpus.

5.4 ANALYSIS OF TAMIL TEXT CORPUS

In this section, a Tamil text corpus provided by CIIL with 2.6 million words is taken and useful statistics about prosodic syllables is collected. This corpus is a collection of Tamil text documents collected from various domains, viz. agriculture, biographies, cooking tips and news articles. A simple algorithm to segment prosodic syllables from a word is proposed whose pseudo code is given below.

```plaintext
function Syllabify (Word[0..n-1])

k ← 0 // Index of current letter in the WORD
m ← 0 // Index of syllable array
// for each letter in the 'Word' categorize it as
// short vowel, long vowel or consonant
for k ← 0 to n-1
    if (Word[k] is a short vowel)
        CharCategory[k] ← 0
    else if (Word[k] is a long vowel)
        CharCategory[k] ← 1
    else
        CharCategory[k] ← 2 // it is a consonant
end for

for k ← 0 to n-1
    if ((k+2) <= n and CharCategory[k] = 0 and
        CharCategory[k + 1] = 1 or
        CharCategory[k + 1] = 0))
```
copy(Syllable[m], Word[k], Word[k+1]);
k = k + 2;
else if (CharCategory[k] = 1 || CharCategory[k] = 0)
copy(Syllable[m], Word[k]);
k = k + 1;
end if

while(k < n && CharCategory[k] = 2)
copy(Syllable[m], Word[k]);
k = k + 1;
end while
m = m + 1;
end for
return m; // returns the no. of syllables; syllables
// are stored Syllable[]
end function

The algorithm works in two stages. Initially, grapheme to phoneme conversion (phonetisation) is done by scanning all the letters of a word and categorizing them as vowels and consonants. The next step of the algorithm combines the letters into syllables with the help of linguistic rules which are presented in Table 5.1. This step is called syllabification. Syllable patterns are checked from the biggest syllable to the smallest one. The algorithm stores the syllables in an array and returns their count.

After applying the algorithm to the text corpus, the frequency counts of various prosodic syllable patterns are gathered. The algorithm segmented 26,153 numbers of unique prosodic syllables in the corpus. Since the text corpus used here is not clean, it contains a lot of abbreviations, digits and other foreign characters. Therefore, the prosodic syllable patterns with frequency less than 10 are eliminated. Then, the rest of the patterns are validated with the help of the DFA deduced in section 5.3.2. Then, it is found that there are only 10,015 numbers of unique prosodic syllables.
Table 5.2 Prosodic Syllables in CIIL Corpus

<table>
<thead>
<tr>
<th>Details</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>686</td>
</tr>
<tr>
<td>Sentences</td>
<td>455,504</td>
</tr>
<tr>
<td>Words</td>
<td>2,652,370</td>
</tr>
<tr>
<td>No. of unique prosodic syllables segmented by the algorithm</td>
<td>26,153</td>
</tr>
<tr>
<td>No. of unique prosodic syllables validated by the DFA</td>
<td>10,015</td>
</tr>
</tbody>
</table>

Since the number of models is small compared to triphone models, it is motivating to use prosodic syllable models in speech recognition.

5.5 PROPOSED SYLLABLE BASED RECOGNITION

The novelty of the approach lies in the fact that a suprasegmental feature i.e. prosodic syllable is used as a sub-word unit. The proposed algorithm works on the very basic linguistic rules of the language to form prosodic syllables in Tamil. Prosodic syllables are the natural units of pronunciation in Tamil. It is different from the phonologically defined syllables used by Lakshmi and Hema Murthy (2006) on transliterated or Romanized Tamil text. The phonologically defined syllables (VC, CVC, CVCC etc.) yield more number of syllable units which increase the number of units to be modeled.

The following example illustrates the difference. When the word ஹொருப்பரசர்ஸ் (reporters) is segmented, the proposed algorithm segments into 4 units whereas the method based on phonological syllable segments into 5 units.
Proposed Method:

Consonant-vowel cluster

A disadvantage associated with phonological syllable is ambi-
syllabic consonants. Ambi-syllabic consonants present hurdle in segmenting
syllable (Ganapathiraju et al 2001) since they occur at syllable boundaries and
belongs to both the preceding and following syllables. The proposed method
avoids the problem of ambi-syllabic consonants.

In the example shown above, ambi-syllabic consonant occurs in $\tilde{\gamma}u$ and
$u_{LT}$ units. On expanding $\tilde{\gamma}u_{LT}$ as $\tilde{\gamma} (C) + \hat{\theta} (V) + \ddot{u} (C) + \underline{\varphi}_h (V)$, it is
found that consonant $\ddot{u}$ is common to both units namely $\tilde{\gamma} (C) + \hat{\theta} (V) + \ddot{u} (C)$ and
$\ddot{u} (C) + \underline{\varphi}_h (V)$. On the other hand, the proposed method naturally
segments $\tilde{\gamma}u_{LT}$ (short + long vowel) as a single prosody syllabic unit
(disyllable - CVCV) thus avoiding the problem of ambi-syllabic consonants.

5.6 CREATING CI SYLLABLE MODELS

A lexicon based on prosodic syllables has been created with the aid
of the proposed algorithm where every word in the dictionary is segmented
into its constituent prosodic syllables. Along with the dictionary, a list of
prosodic syllable models, and continuous speech with sentence aligned
transcription are given as input to the training program. The transcription are
force-aligned with Baum-Welch training followed by Viterbi alignment.

In order to keep the complexity low, it is preferable to model CI
syllable units with single Gaussian continuous density HMM. The continuous
speech is transformed into a sequence of feature vectors. This sequence is
matched with the optimal/best concatenated HMM sequence found using
Viterbi algorithm. The time stamps of segmented syllable boundaries are obtained as a by-product of Viterbi decoding. The duration of the prosodic syllables is found to vary from 290 ms to 315 ms. Even though a prosodic syllable is either monosyllabic or disyllabic, its duration is more or less equal to 300 ms on average. This is due to vowel duration reduction which occurs in non-initial syllables as reported by Asher and Keane (2005).

Based on these considerations, it is decided to have eight states per HMM. Figure 5.3 shows the schematic block diagram of a syllable based recognizer.

![Diagram of Syllable Modeling and Recognition System]

**Figure 5.3 The Syllable Modeling and Recognition System**

### 5.7 RESULTS AND DISCUSSION

For simplicity, an acoustic model has been trained with 1,398 unique prosodic syllables drawn from agriculture domain. These prosodic syllables almost cover the agriculture data and the test set completely. In this experiment, the number of models to be trained has been significantly reduced compared to the triphone models. The baseline triphone model has 3,171 numbers of unique triphones extracted from the transcript. These models are called CD untied models. Figure 5.4 represents the comparison.
Figure 5.4 A Comparison of the number of Syllable and Triphone Models

The experiment is carried out using syllable based continuous speech recognition for Tamil. The dictionary and the transcripts are segmented into prosodic syllables with the proposed algorithm and models have been trained. The models are trained in CI mode like word models described in section 4.4.2. The syllable based acoustic model is deployed on a conventional continuous speech recognizer and tested with the same test set comprising 400 sentences. The result for test sentences is arrayed in Table 5.3.

Table 5.3  A Comparison of the Performance of Word, Triphone and Prosodic Syllable Models

<table>
<thead>
<tr>
<th>Details</th>
<th>Word Models</th>
<th>Triphone Models</th>
<th>Syllable Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences in the Test set</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Words in Test Set</td>
<td>3,085</td>
<td>3,085</td>
<td>3,085</td>
</tr>
<tr>
<td>Words correctly recognized</td>
<td>2,162</td>
<td>2794</td>
<td>2,488</td>
</tr>
<tr>
<td>No. of Errors and type</td>
<td>923 (Sub: 295, Ins: 10, Del: 618)</td>
<td>291 (Sub: 229, Ins: 10, Del: 52)</td>
<td>597 (Sub: 410, Ins: 30, Del: 157)</td>
</tr>
<tr>
<td>WER</td>
<td>29.92%</td>
<td>9.44 %</td>
<td>19.35 %</td>
</tr>
<tr>
<td>Speed</td>
<td>2.46 × Real time</td>
<td>0.65 × Real time</td>
<td>3.52 × Real Time</td>
</tr>
</tbody>
</table>
The comparison of WER of word, syllable and triphone models is illustrated in Figure 5.5. On comparing the results, it is found that WER of syllable models have been considerably reduced compared to word models by 10%.

Figure 5.5 The WER in Syllable Models vs. Triphone Models

It can also be observed that in the prosodic syllable models, there is larger number of substitution errors than that of insertions and deletions whereas in the case of word models, there is a majority of deletion errors. This comparison is shown in Figure 5.6. The majority of deletion errors in word models signify OOV rate due to morphological inflections. The OOV words in syllable models significantly got reduced. This proves the fact that syllables are effective as sub-word units.

Figure 5.6 The Types of Word Errors in Word Models and Syllable Models
However, compared to triphone models, there is an increase in WER by 10% approximately in syllable models. The increase in WER can be attributed to the large number of syllables to be modeled with the available limited training set. This also indicates the presence of a little contextual effect between syllables. The speed of recognition is also slower.

5.8 SUMMARY

The proposed method of CI prosodic syllable based models has addressed the shortcomings of word models on aspects of large inflected vocabulary. The number of models to be trained has been substantially reduced. Compared to word models, there is a significant improvement in the proposed system.

When compared to triphone models, the proposed system has a few shortcomings. Even though there is an increase in WER, the performance is still reasonable. This indicates that the syllable is a promising sub-word unit for speech recognition. The increase in WER and slower recognition speed are quite expected because of the larger size of the prosodic syllables compared to phones and CI modeling of the models.

There are other shortcomings in proposed system. There is no sharing and tying of states of the models. The proposed system can be further enhanced by modeling the context between syllables. Since the number of prosodic syllables is of the order of 10,000, CD modeling of syllables will introduce thousands of new models.

Another approach to improve the accuracy of syllable based method lies in the integration of triphone and syllable models. This approach is dealt with in Chapter six.