Int J Speech Technol (2009) 12: 47–57
DOI 10.1007/s10772-009-9058-0

Syllable modeling in continuous speech recognition for Tamil language

R. Thangarajan · A.M. Natarajan · M. Selvam

Received: 1 November 2009 / Accepted: 2 November 2009 / Published online: 18 November 2009

Abstract In automatic speech recognition, the phone has probably been a dominating sub-word unit for more than one decade. Context Dependent phone or triphone modeling accounts for contextual variations between adjacent phones and state tying addresses modeling of triphones that are not seen during training. Recently, syllable is gaining momentum as a new sub-word unit. Syllable being a larger unit than a phone addresses the severe contextual variations between phones within it. Therefore, it is more stable than a phone and models pronunciation variability in a systematic way. Tamil language has challenging features like agglutination and morpho-phonology. In this paper, attempts have been made to provide solutions to these issues by using the syllable as a sub-word unit in an acoustic model. Initially, a small vocabulary context independent word models and a medium vocabulary context dependent phone models are developed. Subsequently, an algorithm based on prosodic syllable is proposed and two experiments have been conducted. First, syllable based context independent models have been trained and tested. Despite large number of syllables, this system has performed reasonably well compared to context independent word models in terms of word error rate and out of vocabulary words. Subsequently, in the second experiment, syllable information is integrated in conventional triphone modeling wherein cross-syllable triphones are replaced with monophones and the number of context dependent phone models is reduced by 22.76% in untied units. In spite of reduction in the number of models, the accuracy of the proposed system is comparable to that of the baseline triphone system.

Keywords Context dependent · Context independent · Continuous speech recognition · Hidden Markov model · Syllable · Tamil language · Triphone

Abbreviations

ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
CD	Context Dependent
CI	Context Independent
CIIL	Central Institute of Indian Languages, Mysore
CMU	Carnegie Melon University
HMM	Hidden Markov Model
LVCSR	Large Vocabulary Continuous Speech Recognition
SVM	Support Vector Machine
WER	Word Error Rate

1 Introduction

Automatic Speech Recognition (ASR) deals with automatic conversion of acoustic signals of a speech utterance into text transcription. Speech recognition requires segmentation of speech waveform into fundamental acoustic units. The word is the preferred and natural unit of speech because ultimately it is the word that one is trying to recognize. The word, being the largest unit of speech, has an advantage that its

R. Thangarajan (✉) · M. Selvam
Department of Information Technology, Kongu Engineering College, Perundurai 638 052, Erode, India
e-mail: thangs_68@yahoo.com

M. Selvam
e-mail: amm_selvam@yahoo.co.in

A.M. Natarajan
Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam 638 401, Erode, India
e-mail: amnatarajan2006@yahoo.co.in

acoustic representations are well defined. A word accounts for all contextual effects within it. However, using word as a speech unit in large vocabulary continuous speech recognition (LVCSR) system introduces several problems. Each word has to be trained individually and there cannot be any sharing of parameters among words because a word can appear in any context (Huang et al. 2001). Therefore one has to have a very large training set so that all words in the vocabulary are adequately trained. The second problem lies with a memory requirement which grows linearly with the number of words. Hence, word models are not practical for LVCSR systems. However, word models have been successfully used for limited vocabulary ASR applications (Lippmann et al. 1987; Rabiner et al. 1988). Recently, there is a growing interest in ASR for Tamil and other Indian languages. There are speech recognition applications for Tamil which are targeted towards a low and restricted vocabulary task (Khan and Yegnanarayana 2001). There are some funded research works in spoken digit recognition (Plauche et al. 2006). Others have attempted speech recognition for isolated word recognition in Tamil using an Artificial Neural Network (ANN) (Saraswathi and Geetha 2004). The major hurdle in speech research for Tamil or any Indian language is the deficiency in resources like speech and text corpora.

1.1 Phones as sub-word units

A sub-word unit is the next choice because one has to have sharing of parameters across models in order to save computing resources. The most popular sub-word units are phones. Since there are more or less 50 phones in English and other language phone-sets, phone models can be sufficiently trained with a reasonable size of training corpus. However, there are other problems like context variability and instability in phones. It is a well known fact that the same phone in different words has different pronunciations. This is because the articulators cannot move from one position to another instantaneously. Hence the pronunciation of a phone is strongly affected by its adjacent phones or in other words, phones are highly context dependent. For example, some phones are aspirated when they are word-initial and the same phones are not aspirated when they are word-final. The acoustic variability of basic phonetic units due to context is therefore sufficiently large and not well understood in many languages. It has been shown that word based Dynamic Time Warping (DTW) performs significantly better than phone based HMM (Bahl et al. 1988; Paul and Martin 1988). Hence, it can be observed that phone models do not account for the high contextual variations among phones resulting in over-generalization whereas word models need to be trained individually and lack generality.

The problem of over-generalization in phone models has been overcome by employing a context dependent (CD) phone. Context here refers to immediate left and/or right neighboring phones. Context can also include phones beyond immediate left or right phones (± 2) which can substantially increases the computational load. Phone-in-context can be either a left-context or a right-context or both. The third category is also known as triphone which includes both the left and right contexts. If two phones have the same identity but different left or right contexts then they are considered as different triphones. Triphone models are powerful sub-word models because they account for the left and right phonetic contexts. In some studies (Bahl et al. 1980; Schwartz et al. 1984), the use of triphone models reduced the word error rate (WER) by more than 50% as compared with word models and monophone models. However there were problems with triphones also. For any language, there are large numbers of triphones in the training set. This leads to more demand on memory since a model is created for a triphone even if that triphone is observed only once in the training set. Triphone modeling also ignores the similarity between triphones i.e. allophones. This problem has been overcome by the use of parameter sharing techniques. The technique proposed by Lee (1990) clusters similar models while another much sophisticated technique proposed by Hwang and Huang (1993) clusters similar states of models, which leads to finer granularity of sharing and improvement in performance. This model is called a senone model. A senone represents a set of similar Markov states.

1.2 Significance of syllables

The syllable is also a promising unit of speech segmentation. The importance of a syllable as a unit in ASR has been felt in early researches starting with the work by Fujimura (1975) where irregularities in phonemes are discussed, and it has been claimed that a syllable will serve as the viable minimal unit of speech in the time domain. A syllable is generally composed of three parts: the onset, the nucleus and the coda or rhyme. A syllable is usually a larger unit than a phone since it may encompass two or more phonemes. There are few cases where a syllable may consist of a single phoneme only. Hence the problem of severe contextual effects which is prevalent in phones is relatively reduced in syllables. Greenberg (1998) showed that pronunciation variation in Switchboard corpus is more systematic in the level of a syllable. It has been emphasized that the onset and nucleus of a syllable do not show much contextual dependencies while the coda may still be susceptible to some contextual effects with the following syllable. Therefore, syllables are frequently realized in their standard or canonical form whereas in the case of phones canonical realization is mostly unusual. Moreover, a syllable is an intuitive unit for representation of speech because of its structural integrity based firmly on both production and perception of speech. This is what sets the syllable apart from CD phones.

1.3 Syllables as sub-word units

Ganapathiraju et al. (2001) produced the first successful robust LVCSR system that used a syllable level acoustic unit in telephone bandwidth spontaneous speech. The paper begins with a conjecture that a syllable based system would perform better than an existing triphone system and concludes with experimental verification after comparing a syllable based system performance to that of a word-internal and a cross-word triphone system on publically available databases viz. Switchboard and Alphadigits. A number of syllable based experiments involving syllables and context independent (CI) phones, syllables and CD phones, syllables, mono-syllabic words and CD phones have been reported in that paper. However this system is deficient especially in the integration of syllable and phone models as mixed-word entry. It is because mixing models of different lengths and context might result only in marginal improvements.

Several leading researchers in India have also focused their work on the syllable as a speech unit. In a couple of papers by Nagarajan et al. (2001, 2003), an approach is proposed for automatically segmenting and annotating continuous speech into syllable-like units without the use of manually annotated speech corpora. This has been achieved by segmenting the continuous speech signal into syllable-like units using the short-term energy as the magnitude spectrum. Subsequently similar syllables models are clustered using an unsupervised incremental clustering technique and the syllables are labeled manually. Models are then created for the syllable clusters which are trained and used to transcribe continuous speech. A similar approach has been used by Lakshmi and Hema (2006) where both continuous speech utterance and the transcription text are segmented into syllable-like units and models are created.

1.4 Issues in LVCSR of Tamil language

Since Tamil is agglutinative in nature, LVCSR systems for Tamil pose greater challenges. There are many occurrences of the same root word with different prefixes and suffixes. Morpho-phonological processes (also known as *sandhi*) wherein two consecutive words combine by deletion, insertion or substitution of phones at the word boundaries to form a new word are very common in Tamil. Hence in LVCSR for Tamil, models cannot be sufficiently trained from a reasonably sized training data. And moreover, the performance of the speech recognizer degrades as out of vocabulary (OOV) words occur due to inflections and morpho-phonology in the test set. The problems in ASR for Tamil due to inflectional morphology and OOV words have been addressed by Saraswathi and Geetha (2007) who used a morpheme based language model wherein Tamil words are first decomposed into stems and their associated morphemes by a rule based morphological analyzer. The probability patterns among stem and affixes are subsequently modeled. Even though it is true that the morpheme based language model outperforms other language models like trigram, distance based bigram and trigram, dependency based models and class based models, its performance is still limited by the accuracy of the morphological analyzer. While most of the inflectional morphology and *sandhi* are rule based in Tamil, there are still lots of exceptions in these rules which cannot be accounted by a simple rule based morphological analyzer. Therefore we attempt to address this problem from an acoustic model point of view by transcribing speech utterances into syllable stream.

2 The Tamil language

Tamil is a Dravidian language spoken predominantly in the state of Tamilnadu in India and Sri Lanka. It is the official language of the Indian state of Tamilnadu and also has official status in Sri Lanka and Singapore. With more than 77 million speakers, Tamil is one of the widely spoken languages of the world.

2.1 Tamil alphabet

In Tamil, vowels are classified into short (*Kuril*) and long (*Netil*). There are five short vowels and seven long vowels which include two diphthongs. Consonants are classified into three categories with six in each category: hard, soft, and medium with a total of 18 consonants. The soft consonants are also known as nasals. The classification of consonants is based on the place of articulation. The vowels and consonants combine to form 216 compound characters. The compound characters are formed by placing dependent vowel markers on either one side or both sides of the consonant. There is one more special letter *aytham* (ஃ) used in classical Tamil and rarely found in modern Tamil. Summing up the vowels, consonants and compound letters, there are in total 247 characters in standard Tamil alphabet. In addition to the standard characters, six characters are taken from the *Grantha* script which are used in modern Tamil to represent sounds not native to Tamil i.e. words borrowed from Sanskrit and other languages.

2.2 Pronunciation in Tamil

Retroflex consonants are the main characteristics of Tamil phonology. Only a few consonant clusters are permitted in Tamil words and these can never be word initial. Unlike most other Indian languages, Tamil does not distinguish between aspirated and un-aspirated consonants. In addition, the voicing of plosives is governed by strict rules. Plosives

Table 1 Syllabic rules of Tamil language

Description	Pattern	Example (with transliterated Tamil and meaning)
Short vowel, long vowel followed by consonant(s)[a] (*Nirai*)	SV + LV + C(s)	புலால் (pulal) (meat)
Short vowel followed by a long vowel, (*Nirai*)	SV + LV	விழா (vizha) (function)
Two short vowels followed by consonant(s)[a], (*Nirai*)	SV + SV + C(s)	களம் (kaLam) (field)
Two short vowels, (*Nirai*)	SV + SV	கல (kala) (echo sound)
Short vowel followed by consonant(s)[a], (*Ner*)	SV + C(s)	கல் (kal) (stone)
Long vowel followed by consonant(s)[a], (*Ner*)	LV + C(s)	வாள் (vaL) (sword)
Long vowel, (*Ner*)	LV	வா (va) (come)
Short vowel, (*Ner*)	SV	க (ka)

[a]Maximum two Consonants will appear

are unvoiced if they occur word-initially or doubled, elsewhere they are voiced. The Tamil script does not have distinct letters for voiced and unvoiced plosives, although both are present in the spoken language as allophones.

Generally, languages structure the utterance of words by giving greater prominence to some constituents than others. This is true in the case of English: one or more phones standout as more prominent than the rest. This is typically described as lexical stress. The same is true for higher level prosody in a sentence where one or more syllables may bear sentence stress or accent. As far as Tamil language is concerned, it is assumed that there is no lexical stress or accent in Tamil (Arden 1934; Arokianathan 1981; Soundaraj 2000) at word level and all syllables are pronounced with the same emphasis. However there are other opinions that (Marthandan 1983) the position of stress in the word is by no means fixed to any syllable of individual word. In connected speech, the stress is found more often in the initial syllable (Balasubramanian 1980). In some studies (Asher and Keane 2005) it is shown that there is marked reduction in a vowel's duration of non-initial syllables compared to initial syllables. Regardless of the experimental results hinting at the presence of lexical stress, we assume no stress on any syllable of a word because we are dealing with read speech.

In Tamil, there is a unique feature called a prosodic syllable (*asai*). These prosodic syllables were used in ancient literatures to write poetry. The grammar of poetry is strictly based on prosodic syllables. In general, the structure of prosodic syllables is almost same as that of syllables, but sometimes a prosodic syllable may have two vowel nuclei with an ambisyllabic consonant in the middle. The prosodic syllabic representation in Tamil language comprises the combinations of short vowel (SV), long vowel (LV) and consonant (C). Depending on the vowel and consonant combinations, a prosodic syllable is classified into two types: *Ner Asai* and *Nirai Asai*. The prosodic syllabic representations take any of the only eight patterns which are tabulated in Table 1. While segmenting a word into prosodic syllables, larger syllable patterns should be compared first and smaller syllable patterns should be compared next. Hence the application of syllable patterns should proceed in the same order as they appear in Table 1.

Root words in Tamil may consist of up to five prosodic syllables. Due to *sandhi*, word boundaries are often merged in spoken language. This is again a unique feature of Tamil language where word boundaries which appear in the written text are ignored in speech. The following examples illustrate this feature.

அரசு (*government*) + பணி (*service*)

– அரசுப்பணி

ஆபரணம் (*ornament*) + தங்கம் (*gold*)

– ஆபரணத்தங்கம்

In the both examples, the two distinct words on the left hand side are pronounced as a single word when spoken. *Sandhi* does not occur between any pair of adjacent words, but its occurrence is more frequent and is strictly governed by linguistic rules. Hence based on our study, it is our assumption that the contextual effect between adjacent prosodic syllables is minimal and contextual effects exist in word boundaries due to *sandhi*.

3 Problem formulation

After carefully analyzing the methods for syllable modeling in the existing studies, we found the following deficien-

Table 2 Results of Baseline systems

Details	CD Phone Models	CI Word Models
Utterances in the Test set	400	400
Words in Test Set	3,085	3,085
Words Correctly Recognized	2,794	2,162
No. of word errors and type	291	923
	(Sub: 229, Ins: 10, Del: 52)	(Sub: 295 Ins: 10 Del: 618)
WER	09.44%	29.92%
Speed	0.65 × Real Time	2.46 × Real time

cies. Since there are a large number of syllables, there are shortcomings in training them. Conventional methods used for CD phones cannot be used for training syllables models. Existing methods circumvent the problem by mixing units of different length and context. Therefore a sound method to integrate syllable with phone is very much necessary. For resource deficient languages like Tamil, LVCSR is still a challenging task. The agglutinative nature of the language further deepens the problem. Existing methods address the problem by segmenting speech utterances into syllable-like units in the front-end or signal processing domain. Since Tamil language is computationally appealing with its linguistic rules which help segment words into prosodic syllables, it is motivating to apply prosodic syllables in speech recognition.

In this paper, we reveal the importance of prosodic syllable modeling in continuous speech recognition of Tamil language. In Tamil, a word however long is ultimately composed of one or more prosodic syllables and there are strong linguistic rules to form such syllables, unlike English where syllabification is based on vowel-consonant clusters and is rather vague. Through the prosodic syllable modeling we attempt to address two different problems in LVCSR systems for Tamil language. Hereafter prosodic syllable will be referred to as syllable.

(i) Managing the issue of large vocabulary size due to inflectional morphology by transcribing speech utterances into syllable units using CI syllable based acoustic model.
(ii) Seamless integration of syllable and triphone modeling.

4 Development of speech corpus

Since speech corpora are not available for Tamil, an in-house speech corpus has been created. A corpus containing 22.5 hours of continuous read speech of 50 people—25 males and 25 females—for training has been created. All the speakers spoke from the same set of unique 550 sentences. In total, there are 27,500 utterances. The sentences are drawn from two domains viz. Agriculture and Magic show from the text corpus (CIIL; http://www.ciilcorpora.net/tamsam.htm). The test set comprises of 0.33 hours (20 minutes) of continuous speech spoken by 25 speakers—13 males and 12 females. Every speaker in the test set spoke 16 different sentences whose lexicon is covered in the training set. The recording was carried out in a noise free lab environment. Finally, sentence level transcriptions have been done manually.

4.1 Acoustic features

A HMM based acoustic model trainer from Carnegie Mellon University (CMU), *SphinxTrain*, has been employed. The input file format is MSWAV sampled at 16,000 Hz, with a depth of 16 bits and mono channel. The acoustic feature consists of 13 dimensional Mel Frequency Cepstral Coefficients (MFCC), their first and second derivatives. The MFCCs were estimated with a window size of 25 ms and a frame shift of 10 ms.

4.2 Baseline systems

A medium vocabulary CD phone based acoustic model and a small vocabulary CI word models for Tamil have been successfully built. The HMM of CD phone models have three states with continuous density of eight Gaussians per state. For CI word models, 20 states HMM with single Gaussian have been used. The training of CD phone models involved CI monophones, CD untied models, state-tying with decision tress, density mixture generation and CD tied models. On the other hand, word models were trained as CI monophones. A word based trigram language model with a perplexity of 72 is used in the both systems. These systems have been used as the baseline system for our experiments. The hypothesis word sequences from the decoder are aligned with reference sentences and the results are shown in Table 2.

5 Proposed methodology

Two methodologies are proposed which demonstrate the syllable's significance in speech recognition. In the first

methodology, modeling syllable as an acoustic unit is suggested and CI syllable models are trained and tested. The second methodology proposes integration of syllable information in the conventional triphone or CD phone modeling.

Both the methodologies use an algorithm for segmenting Tamil text into syllables. This proposed algorithm is unique in the sense that it uses linguistic rules of the language (discussed in Sect. 2.2) to segment a word into prosodic syllables which are natural units of pronunciation in Tamil as opposed to consonant-vowel clusters used in by Lakshmi and Hema (2006) on transliterated or Romanized Tamil text. The algorithm initially scans all the letters of a given word and categorizes them into long vowels, short vowels and consonant categories. The next step of the algorithm combines the letters into syllables with the help of syllabic rules which are tabulated in Table 1. Syllable patterns in a given word are checked from the biggest syllable to the smallest one. The algorithm stores the syllables in an array and also returns their count. Another advantage of the proposed algorithm is that it works with Unicode encoding which is more natural compared to transliterated text.

The consonant-vowel clusters (e.g. VCV, CVC(C), etc.) would yield more numbers of syllable-like units which would increase the number of units to be modeled. The following example illustrates the difference. When the word செய்தியாளர்கள் (*reporters*) is segmented, the proposed algorithm segments into 4 units whereas the method based on consonant-vowel cluster segments into 5 units.

Proposed Method:
செய்தியாளர்கள் செய் தியா ளர் கள்

Consonant-vowel cluster
செய்தியாளர்கள் செய் தி யா ளர் கள்

Sometimes, consonant-vowel clustering also leads to ambi-syllabic consonants. Ambi-syllabic consonants present a 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 தி and யா units. On expanding தியா as த் (C) + இ (V) + ய் (C) + ஆ (V), it is found that consonant ய் is common to both units namely த் (C) + இ (V) + ய் (C) and ய் (C) + ஆ (V). On the other hand, the proposed method naturally segments தியா (short + long vowel) as a single prosody syllabic unit (CVCV) thus avoiding the problem of ambi-syllabic consonants.

5.1 Creating syllable based models

The objective of this experiment is to address the problem of large vocabulary in Tamil due to inflectional morphology and *Sandhi*. Since Tamil words are composed of syllables, a speech recognition system can be made domain independent if it can transcribe speech utterances to a stream of syllables. We have analyzed a text corpus (CIIL; http://www.ciilcorpora.net/tamsam.htm) with 2.6 million words. This corpus is a collection of Tamil text documents collected from various domains viz. agriculture, biographies, cooking tips, new articles, etc. Table 3 shows the details of analysis made on the corpus. There are 26,153 unique syllables in the corpus with a minimum occurrence of one time. It is further found that there are only 4,023 numbers of syllables have frequency greater than 25.

Table 3 Corpus statistics

Details	Counts
Documents	686
Sentences	455,504
Words	2,652,370
Unique Syllables	26,153
Unique Syllables (Frequency greater than 25)	4,023

Therefore a syllable-based lexicon has been created with our algorithm where every word in the dictionary is segmented into its constituent syllables. Along with the dictionary, phone-set (in this case a syllable-set) comprising 1398 syllables and continuous speech with sentence aligned transcription are used to the training the models. The transcription are force aligned with *Baum-Welch* training followed by Viterbi alignment.

In order to keep the complexity low, modeling of CI syllables with single Gaussian continuous density has been preferred. Here the novelty of our approach lies in the syllable segmentation algorithm of the text. 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 and the time stamps of segmented syllable boundaries are obtained as a by-product of Viterbi decoding. Durations of the syllables are found to vary from 290 ms to 315 ms. This is an important finding in our study. Even though the length of a syllable varies from one phoneme to four phonemes, the duration is equal to 300 ms on average. This is due to vowel duration reduction which occurs in non-initial syllables. Typically, duration of long vowels which is supposed to be twice than that of short vowels is actually shortened. These considerations have made us to decide the number of states per HMM to be eight. Figure 1 shows the schematic block diagram of a syllable based recognizer.

For simplicity, an acoustic model was trained with 1398 unique syllables drawn from agriculture domain. These syllables almost cover the training data and the test set fully. The results of baseline word and triphone models, and proposed syllable model on test sentences is shown in Table 4.

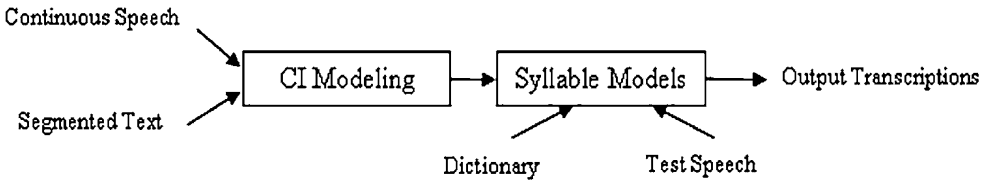

Fig. 1 Syllable modeling and recognition system

Table 4 Performance of syllable models and triphone models

Details	Word Models	Triphone Models	Syllable Models
Sentences in the Test set	400	400	400
Words in Test Set	3,085	3,085	3,085
Words correctly recognized	2,162	2794	2,488
No. of Errors and type	923 (Sub: 295 Ins: 10 Del: 618)	291 (Sub: 229, Ins: 10, Del: 52)	597 (Sub: 410 Ins: 30 Del: 157)
WER	29.92%	9.44%	19.35%
Speed	2.46 × Real time	0.65 × Real time	3.52 × Real Time

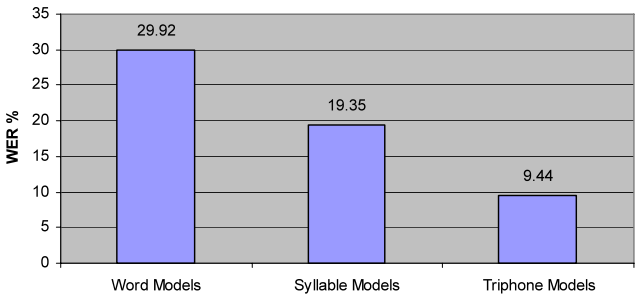

Fig. 2 Comparison of WER in syllable models vs. triphone models

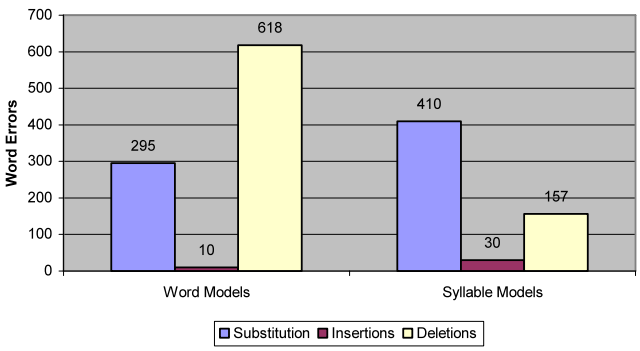

Fig. 3 Types of word errors in word models and syllable models

Comparison of WER of the three units is shown in Fig. 2. On comparing the WER to that of baseline word and triphone models, it is found that WER of syllable models have been considerably reduced compared to word models by 10%.

It can also be observed that in syllable models there is large 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 is shown in Fig. 3. Majority of deletion errors in word models signify OOV rate due to morphological inflections which are accounted for in syllable models. This proves the fact that syllables are effective as sub-word units.

However, compared to triphone models, there is an increase in WER by 10% approximately. 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.

5.2 Integrating syllables with triphone modelling

The objective of the approach described in this section is to demonstrate a novel method of integrating syllable information in the conventional triphone modeling procedure. A typical CD phone modeling involves the following steps:

1. Flat-start monophone training is done first. In this step monophone seed models with nominal values are generated, and re-estimation of these models using reference transcriptions is done.
2. Baum-Welch training of monophones is carried out in second step. Adjustment of the silence model, re-estimation of single-Gaussian monophones is done using the standard Viterbi alignment process.
3. Triphone generation is the third step. Creation of triphone transcriptions from monophone transcriptions, initial triphone training, triphone clustering, state tying, training of state-tied triphones are carried out.
4. Mixture generation is done finally by splitting single Gaussian distributions into mixture distributions using an iterative divide-by-two clustering algorithm followed by re-estimation of triphone models with mixture distributions.

Table 5 Triphones in baseline and proposed systems

S. No.	Triphones	Baseline System	Proposed System	% of Reduction in Triphones
Triphones extracted from the dictionary				
a.	Unique single word triphones	700	700	–
b.	Unique word—initial triphones	4,312	4,312	–
c.	Unique word—internal triphones	1,651	1,651	–
d.	Unique word—final triphones	3,825	3,825	–
e.	Unique Triphones extracted from dictionary—Virtual triphones (Sum of a. to d.)	10,488	10,488	–
Triphones extracted from the transcript				
f.	Single word triphones	14	14	–
g.	Word—initial triphones	41,076	41,076	–
h.	Word—internal triphones	237,412	132,790	44.06
i.	Word—final triphones	41,076	41,076	–
j.	Triphones extracted from transcripts (Sum of f. to i.)	319,578	214,956	44.06
k.	Unique triphones extracted from transcripts—Real triphones	3,171	2,449	22.76

Based on the role of prosodic syllables in the pronunciation of a Tamil word, it is conjectured that inter-syllable triphones will not affect the accuracy of the recognition because of minimal contextual effects. Therefore, an augmentation is made to the conventional training procedure of triphones. The input word from the transcription is segmented into syllables. Using the syllable information, inter-syllable triphones (triphones whose left and right context phones span across adjacent syllables) are marked and both the left and right contexts are reduced to monophones. This will reduce the number of triphone models considerably.

5.2.1 Reducing cross syllable triphones to monophones

Prior to the creation of triphone models, HMM trainer generates a list of all possible triphones from the dictionary. These triphones are called virtual triphones because they may or may not occur in the training data. The next step consists of counting the triphones as they occur in the training data. This procedure scans the training data, namely the transcription file and increments a count against each triphone in the virtual triphone list when the same occurs in the training data.

Once the counting is finished, a threshold value is set and triphone models are created for those triphones whose count is greater than the threshold value. In order to create as many triphones as possible from the training set, in this experiment, the threshold value is set to 1 for both the baseline and proposed system. These triphones are called real triphones which are used to compute the CD untied models. Table 5 shows the various statistics of triphone modeling in the baseline and proposed system.

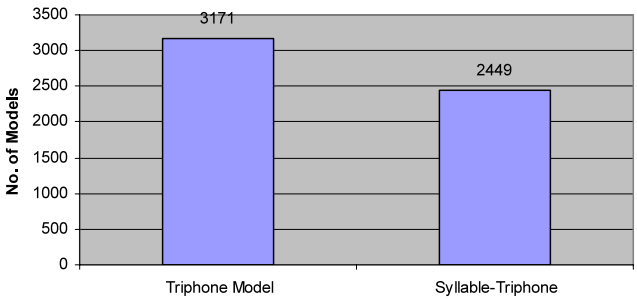

Fig. 4 No. of CD untied models in baseline vs. proposed system

There is a drastic reduction by 22.76% in the unique number of real triphones compared to the baseline system. Figure 4 shows the comparison of number of models in triphone and proposed approach.

5.2.2 Illustration with an example

Taking the Tamil word செய்தியாளர்கள் the working of the proposed algorithm is explained. In Table 6, the first row shows the letters and the second row shows their corresponding phones. The word செய்தியாளர்கள் consists of eight Tamil letters and 14 phones. The conventional algorithm will enumerate 12 triphones viz. {c eh y}, {eh y T}, {y T h} . . . {r, k, ah} and {k ah L} as shown in the third row of the Table 6.

In the proposed approach, the word செய்தியாளர்கள் is segmented into syllables viz. செய், தியா, ளர் and கள் by the proposed algorithm. From the enumerated triphones shown in third row, cross-syllable triphones (superscripted with b) are identified and modeled as monophones. The

Table 6 Process of triphone reduction

செய்தியாளர்கள்	செ	ய்	தி	யா	ள	ர்	க	ள்	8 letters
Phones	c eh	y	T h ih	y aa	L ah	r	k ah	L	14 phones
Triphones	{c eh y}[a], {eh y T}[b], {y T h}[b], {T h ih}, {h ih y}, {ih y aa}, {y aa L}[b], {aa L ah}[b], {L ah r}, {ah r k}[b], {r k ah}[b] and {k ah L}[a]								12 triphones
Triphones which constitute syllabic units	{c eh y}[a], {T h ih}, {h ih y}, {ih y aa}, {L ah r}, and {k ah L}[a]								6 triphones

[a]Word-initial and word-final triphones

[b]Cross-syllable triphones

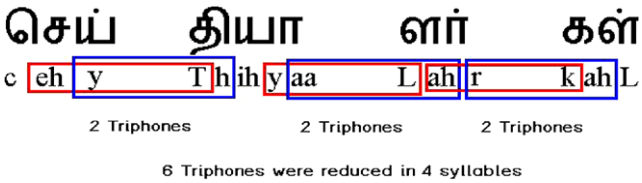

Fig. 5 Illustration of inter-syllable triphone reduction

cross-syllable triphones are also depicted in Fig. 5. This approach reduces the number of triphones to six. Due to possible contextual effects between words as a result of *sandhi*, the word-initial phone has a left context with the last phone of the preceding word or silence (SIL) and the word-final phone has a right context with the first phone of the following word or SIL. Thus, the cross-word modeling of the baseline system is preserved in the proposed system.

The overall process of the proposed approach of integrating syllable with triphone modeling comprises syllabification of an input word and inter-syllable triphone identification. This process is illustrated in the flowchart in Fig. 6.

5.2.3 Decision tree based clustering and state tying

Next step in the training process is the decision tree based clustering of triphones for modeling unseen triphones, followed by Gaussian mixture generation and modeling. The unseen triphones are the ones which are in the virtual triphone list but not present in the training data. With the CD untied models and set of phonetic classes (which share some common properties), decision tree for state based clustering is built whose leaf nodes are CD tied states or senones. The CD tied states are finally modeled. Since there is a reasonable size training data, continuous density HMM modeling with Gaussian mixture densities is opted for. Performance is also a function of number of mixtures. Therefore it is decided the number of Gaussians to be 8 which has given good performance with less computational complexity. Mixture generation has been done by splitting single Gaussian distribution into mixture distribution using an iterative divide-by-two clustering algorithm followed by re-estimation of triphone models with mixture distribution.

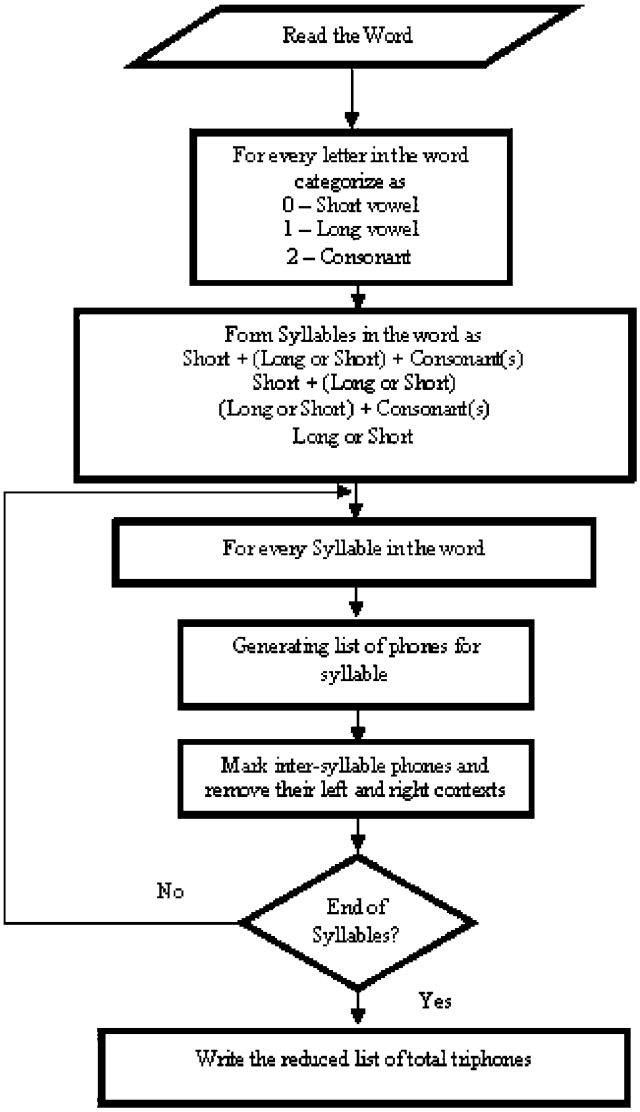

Fig. 6 Flowchart of syllabification and triphone reduction

The resulting acoustic model is deployed on CMU Sphinx decoder (Lamere et al. 2003) and tested with the same test set comprising 400 sentences. The result shows that the proposed syllable-triphone integrated approach

Table 7 Performance of baseline triphone models and syllable-triphone approach

Details	Triphone Models	Syllable-Triphone
Sentences in the Test set	400	400
Words in Test Set	3,085	3,085
Words correctly recognized	2,794	2,757
No. of Errors and type	291 (Sub: 229, Ins: 10, Del: 52)	328 (Sub: 271 Ins: 12 Del: 45)
WER	9.44%	10.63%
Speed	0.65 × Real time	0.67 × Real Time

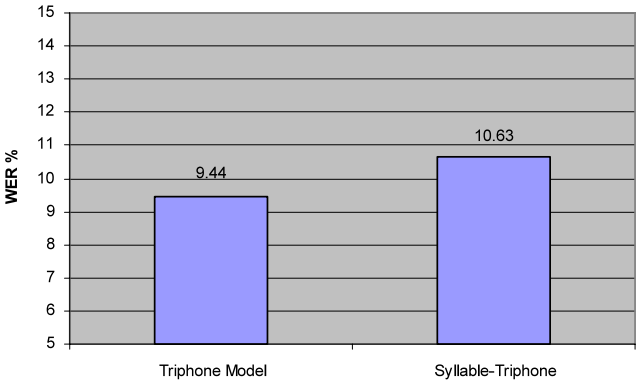

Fig. 7 WER in baseline vs. proposed system

gives comparable performance to that of baseline system with only a marginal increase (approximately 1%) in the WER. The comparison of performance is shown in Table 7.

Despite a considerable reduction in the number of models, the WER of the proposed system is 10.63%. This is shown in Fig. 7.

6 Conclusion

As discussed in Sects. 5.1 and 5.2, two experiments in continuous speech recognition for Tamil language have been carried out using prosodic syllables as sub-word units. The first experiment demonstrates a conventional continuous speech recognizer working with CI syllable models. The dictionary and the transcripts are segmented into syllables with the proposed algorithm and syllable models are trained. This method addresses the problem of root word inflections in the vocabulary. The analysis of a 2.6 million word corpus shows that only 4023 number of syllables occur more than 25 times. The number of models to be trained has been substantially reduced. Compared to word models there is significant improvement (10% reduction in WER) in the proposed CI syllable based system. The OOV word rate has also been substantially reduced.

However, when compared to triphone models, the proposed CI syllable based system has a few shortcomings. There is an increase in WER by 10%. The increase in WER and slower recognition speed are quite expected because of the larger size of syllables compared to phones and minimal context variability among syllables. The increase in WER can also be attributed to large number of syllables to be modeled with a limited training set. This system also has the few other short comings. There is no sharing and tying of the states of the syllable models.

In the second experiment, syllable information has been integrated into the conventional CD phone modeling by means of segmenting the word into syllables and eliminating inter-syllable triphones' left and right contexts. The phones on syllable boundaries are thus modeled as monophones. This experiment is conducted to demonstrate that in the pronunciation of Tamil words, there is minimal contextual effect between two adjacent syllables. Comparing the results shown in Table 7, it is found that the performance of the recognizer is comparable (marginally lower by 1.19%) to the baseline model despite a substantial reduction in the number of CD untied models (22.76%). Hence the conjecture that syllable is an intuitive unit for speech recognition is experimentally verified.

An improvement to the CI syllable model could be done by modeling context between syllables units, but this would introduce thousands of new parameters to be modeled. Further research could be driven in this direction.

Acknowledgement The authors would like to thank Central Institute of Indian Languages (CIIL), Mysore, India for providing the Tamil text corpus and Tamil Virtual University, Chennai for funding the project on speech recognition.

References

Arden, A. H. (1934). *A progressive grammar of common Tamil* (4th ed.). Madras: Christian Literature Society, pp. 59.

Arokianathan, S. (1981). *Tamil clitics*. Trivandrum: Dravidian Linguistics Association, pp. 5.

Asher, R. E., & Keane, E. L. (2005). Diphthongs in colloquial Tamil. In W. J. Hardcastle & J. Mackenzie Beck (Eds.) (pp. 141–171).

Bahl, L. R., Bakis, R., Cohen, P. S., Cole, A. G., Jelinek, F., Lewis, B. L., & Mercer, R. L. (1980). Further results on the recognition of a continuously read natural corpus, presented at the IEEE international. In *Conference on acoustics, speech, signal processing*.

Bahl, L. R., Brown, P. F., De Souza, P. V., & Mercer, R. L. (1988). *Acoustic Markov models used in the Tangora speech recognition system*. Presented at the IEEE international conference on acoustics, speech, signal processing, 1988.

Balasubramanian, T. (1980). Timing in Tamil. *Journal of Phonetics*, *8*, 449–467.

CIIL, Central Institute of Indian Languages, Mysore, India. http://www.ciilcorpora.net/tamsam.htm.

Fujimura, O. (1975). Syllable as a unit of speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *ASSP-23*(1), 82–87.

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., & Doddington, G. R. (2001). Syllable based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, *9*(4), 358–366.

Greenberg, S. (1998). Speaking in short hand—a syllable centric perspective for understanding pronunciation variation. In *Proceedings of the ESCA workshop on modeling pronunciation variation for automatic speech recognition*, Kekrade, 1998 (pp. 47–56).

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing—a guide to theory, algorithm and system development*. Englewood Cliffs: Prentice-Hall PTR. ISBN:0-13-022616-5.

Hwang, M. Y., & Huang, X. D. (1993). Shared distribution hidden Markov models for speech recognition. *IEEE Transactions on Speech and Audio Processing*, *1*(4), 414–420.

Khan, A. N., & Yegnanarayana, B. (2001). Development of speech recognition system for Tamil for small restricted task. In *Proceedings of national conference on communication*, India, 2001.

Lakshmi, A., & Hema, A. M. (2006). A syllable based continuous speech recognizer for Tamil. In *INTERSPEECH 2006*, Pittsburgh, Pennsylvania (pp. 1878–1881).

Lamere, P., Kwok, P., Walker, W., Gouvea, E., Singh, R., Raj, B., & Wolf, P. (2003). Design of the CMU Sphinx-4 decoder. In *EUROSPEECH 2003*.

Lee, K. F. (1990). Context dependent phonetic Markov models for speaker independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *38*(4), 599–609.

Lippmann, R. P., Martin, E. A., & Paul, D. P. (1987). Multi-style training for robust isolated-word speech recognition. In *Proc. IEEE international conference on acoustics, speech, signal processing* (pp. 705–708).

Marthandan, C. R. (1983). *Phonetics of casual Tamil*. Ph.D. thesis, University of London.

Nagarajan, T., Kamakshi Prasad, V., & Hema, A. M. (2001). The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation. In *Sixth biennial conference of signal processing and communications*.

Nagarajan, T., Hema, A. M., & Hegde, R. M. (2003). Segmentation speech into syllable-like units. In *EUROSPEECH-2003* (pp. 2893–2896).

Paul, D. B., & Martin, E. A. (1988). *Speaker stress-resistant continuous speech recognition*. Presented at the IEEE international conference on acoustics, speech, signal processing.

Plauche, M., Udhyakumar, N., Wooters, C., Pal, J., & Ramachadran, D. (2006). Speech recognition for illiterate access to information and technology. In *Proceedings of first international conference on ICT and development*.

Rabiner, L. R., Wilpon, J. G., & Soong, F. K. (1988). *High performance connected digit recognition using hidden Markov models*. Presented at the IEEE int. conf. acoustics, speech, signal processing.

Saraswathi, S., & Geetha, T. V. (2004). *Lecture notes in computer science: Vol.* 3285. *Implementation of Tamil speech recognition system using neural networks*.

Saraswathi, S., & Geetha, T. V. (2007). Comparison of performance of enhanced morpheme-based language model with different word-based language models for improving the performance of Tamil speech recognition system. *ACM Transaction on Asian Language Information Processing*, *6*(3), Article 9.

Schwartz, R. M., Chow, Y. L., Roucos, S., Krasner, M., & Makhoul, J. (1984). *Improved hidden Markov modeling phonemes for continuous speech recognition*. Presented at the IEEE international conference acoustics, speech, signal processing.

Soundaraj, F. (2000). Accent in Tamil: Speech research for speech technology. In K. Nagamma Reddy (Ed.), *Speech technology: Issues and implications in Indian languages* (pp. 246–256). Thiruvananthapuram: International School of Dravidian Linguistics.