Chapter 2

Speech Recognition for Agglutinative Languages

R. Thangarajan

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/50140

1. Introduction

Speech technology is a broader area comprising many applications like speech recognition, Text to Speech (TTS) Synthesis, speaker identification and verification and language identification. Different applications of speech technology impose different constraints on the problem and these are tackled by different algorithms. In this chapter, the focus is on automatically transcribing speech utterances to text. This process is called Automatic Speech Recognition (ASR). ASR deals with transcribing speech utterances into text of a given language. Even after years of extensive research and development, ASR still remains a challenging field of research. But in the recent years, ASR technology has matured to a level where success rate is higher in certain domains. A well-known example is human-computer interaction where speech is used as an interface along with or without other pointing devices. ASR is fundamentally a statistical problem. Its objective is to find the most likely sequence of words, called hypothesis, for a given sequence of observations. The sequence of observations involves acoustic feature vectors representing the speech utterance. The performance of an ASR system can be measured by aligning the hypothesis with the reference text and by counting errors like deletion, insertion and substitution of words in the hypothesis.

ASR is a subject involving signal processing and feature extraction, acoustics, information theory, linguistics and computer science. Speech signal processing helps in extracting relevant and discriminative information, which is called features, from speech signal in a robust manner. Robustness involves spectral analysis used to characterize time varying properties of speech signal and speech enhancement techniques for making features resilient to noise. Acoustics provides the necessary understanding of the relationship between speech utterances and the physiological processes in speech production and speech perception. Information theory provides the necessary procedures for estimating parameters of statistical models during training phase. Computer science plays a major role in ASR with its implementation of efficient algorithms in software or hardware for decoding speech in real-time.

Currently, an important area in speech recognition is Large Vocabulary Continuous Speech Recognition (LVCSR). A large vocabulary means that system has a vocabulary ranging from 5,000 to 60,000 words. Continuous speech means that the utterances have words which are run together naturally. This is different from isolated word speech recognition where each word is demarcated with initial and final silence. Speech recognition algorithms can also be speaker independent, i.e. they are even able to recognize the speech of new users with whom the system is not exposed. Xuedong et al (2001) present a very good reference on algorithms and techniques in speech processing.

1.1. Issues in speech variability

Even though state-of-the-art speech recognition systems cannot match human performance, still they can recognize spoken input accurately but with some constraints. Some of the constraints may be speaker dependency, language dependency, speaking style, and applicability to a particular task and environment. Therefore, building a speech recognizer that could recognize the speech of any speaker, speaking in any language, style, domain and environment is far from realization.

a. Context variability

In any language, words with different meanings have the same phonetic realization. Their usage depends on the context. There is even more context dependency at the phone level. The acoustic realization of a phone is dependent on its neighboring phones. This is because of the physiology of articulators involved in speech production.

b. Style variability

In isolated speech recognition with a small vocabulary, a user pauses between every word while speaking. Thus it is easy to detect the boundary between words and decode them using the silence context. This is not possible in continuous speech recognition. The speaking rate affects the word recognition accuracy. That is, the higher the speaking rate, the higher the WER.

c. Speaker variability

Every speaker's utterance is unique per se. The speech produced by a speaker is dependent on a number of factors, namely vocal tract physiologies, age, sex, dialect, health, education, etc. For speaker independent speech recognition, more than 500 speakers from different age groups, their sex, educational background, and dialect are necessary to build a combined model. The speaker independent system includes a user enrollment process where a new user can train his voice with the system for 30 minutes before using it.

d. Environment variability

Many practical speech recognizers lack robustness against changes in the acoustic environment. This has always been a major limitation of speech based interfaces used in mobile communication devices. The acoustic environment variability is highly unpredictable and it cannot be accounted for during training of models. A mismatch will always occur between the trained speech models and test speech.

1.2. Measure of performance

The ultimate measures of success for any speech recognition algorithms are accuracy and robustness. Therefore, it is important to evaluate the performance of such a system. The WER is one of the most widely used measures for accuracy. The system may be tested on sample utterances from the training data for understanding the system and identification of bugs during the development process. This would result in a better performance than what one can get with test data. In addition, a development set can also be used to test the system and also fine-tune its parameters. Finally, the system can be tested on a test set comprising around 500 speech utterances of 5-10 different users in order to reliably estimate accuracy. This test set should be completely new with respect to training and development. There are three types of word recognition errors in speech recognition:

- Substitution (*Subs*): An incorrect word substituted for a correct word.
- Insertion (*Ins*): An extra word added in the recognized sentence.
- Deletion (*Dels*): A correct word omitted in the recognized sentence.

Generally, a hypothesis sentence is aligned with the correct reference sentence. The number of insertions, substitutions and deletions are computed using maximum substring matching. This is implemented using dynamic programming. The WER is computed as shown in equation (1):

$$WER = \frac{Subs+Dels+Ins}{N} \times 100 \quad (1)$$

Other performance measures are speed and memory footprints. The speed is an important factor which quantifies the turn around time of the system once the speech is uttered. It is calculated as shown in equation (2):

$$speed = \frac{time\ taken\ for\ processing}{utterance\ duration} \times real\ time \quad (2)$$

Obviously, the time taken for processing should be shorter than the utterance duration for a quicker response from the system. Memory footprints show the amount of memory required to load the model parameters.

There are a number of well-known factors which affect the accuracy of an ASR system. The prominent factors are those which include variations in context, speakers and noise in the environment. Research in ASR is classified into different types depending on the nature of the problems, like a small or a large vocabulary task, isolated or continuous speech, speaker dependent or independent and robustness to environmental variations. The state-of-the-art speech recognition systems can recognize spoken input accurately with some constraints. The constraints can be speaker dependency, language dependency, speaking style, task or environment. Therefore, building an automatic speech recognizer which can recognize the speech of different speakers, speaking in different languages, with a variety of accent, in any domain and in any ambience environmental background is far from reality.

ASR for languages like English, French and Czech is well matured. A lot of research and development have also been reported for oriental languages like Chinese and Japanese. But in the Indian scenario, ASR is still in its nascent stage due to the inherent agglutinative nature of most of its official languages. Agglutination refers to the extensive morphological inflection in which one can find a one-to-one correspondence between affixes and syntactic categories. This nature results in a large number of words in the dictionary which hinders modeling and training of utterances, and also creates Out-Of-Vocabulary (OOV) words when deployed.

2. Speech units

The objective of this chapter is to discuss a few methods to improve the accuracy of ASR systems for agglutinative languages. The language presented here as a case study is Tamil (ISO 639-3 *tam*). Tamil is a Dravidian language spoken predominantly in the state of Tamilnadu in India and in Sri Lanka. It is the official language of the Indian state of Tamilnadu and also has official status in Sri Lanka, Malaysia and Singapore. With more than 77 million speakers, Tamil is one of the widely spoken languages in the world. Tamil language has also been conferred the status of classical language by the government of India.

Currently, there is a growing interest among Indian researchers for building reliable ASR systems for Indian languages like Hindi, Telegu, Bengali and Tamil. Kumar et al (2004) reported the implementation of a Large Vocabulary Continuous Speech Recognition (LVCSR) system for Hindi. Many efforts have been put to build continuous speech recognition systems for Tamil language with a limited and restricted vocabulary (Nayeemulla Khan and Yegnanarayana 2001, Kumar and Foo Say Wei 2003, Saraswathi and Geetha 2004, Plauche et al 2006). Despite repeated efforts, a LVCSR system for the foresaid languages is yet to be explored to a significant level. Keeping agglutination apart, there are other issues to be addressed like aspirated and un-aspirated consonants, and retroflex consonants.

2.1. Phones, phonemes and syllables

Before embarking on the concepts, it is always better to review a few terminologies pertaining to linguistics. In any language, there are acoustic properties of speech units and symbolic representation of lexical units. For more information, please refer (Xuedong et al, 2001).

a. Phonemes and phones

From the acoustics point of view, a phoneme is defined as the smallest segmental unit of sound employed to tell apart meaningfully between utterances. A phoneme can be considered as a group of slightly different sounds which are all perceived to have the same function by the speakers of a language or a dialect. An example of a phoneme is the /k/ sound in the words *kit* and *skill*. It is customary to place phonemes between slashes in

transcriptions. However, the phoneme /***k***/ in each of these words is actually pronounced differently i.e. it has different realizations. It is because the articulators which generate the phoneme cannot move from one position to another instantaneously. Each of these different realizations of the phoneme is called a phone or technically an allophone (in transcriptions, a phone is placed inside a square bracket like [*k*]). A phone can also be defined as an instance of a phoneme. In ***k****it* [*k*] is aspirated while in *s****k****ill* [*k*] is un-aspirated. Aspiration is a period of voiceless-ness after a stop closure and before the onset of voicing of the following vowel. Aspiration sounds like a puff of air after the [*k*] and before the vowel. An aspirated phone is represented as [k^h]. In some languages, aspirated and un-aspirated consonants are treated as different phonemes. Hindi, for instance, has four realizations for [*k*] and they are considered as different phonemes. Tamil does not discriminate them and treats them as allophones.

b. Words

Next comes the representation in symbolic form. According to linguistics, word is defined as a sequence of morphemes. The sequence is determined by the morpho-tactics. A morpheme is an independent unit which makes sense in any language. It could refer to the root word or any of the valid prefixes or suffixes. Therefore what is called a word is quite arbitrary and depends on the language in context. In agglutinative languages a word could consists of a root along with its suffixes – a process known as inflectional morphology. Syntax deals with sentence formation using lexical units. Agglutinative languages, on one hand, exhibit inflectional morphology to a higher extent. On the other hand, the syntactic structure is quite simple which enables free-word ordering in a sentence. The English language, which is not agglutinative, has simpler lexical morphology but the complexity of the syntactic structure is significantly higher.

c. Syllables

A syllable is a unit of organization for a sequence of speech sounds. It is composed of three parts: the onset, the nucleus and the coda. A syllable has a hierarchical structure as shown in Figure 1. It can also be expressed in Backus-Naur Form (BNF) as follows:

<syllable> ::= <onset> <rhyme>

<rhyme> ::= <nucleus> <coda>

A vowel forms the nucleus of the syllable while an optional consonant or consonant cluster forms the onset and coda. In some syllables, the onset or the coda will be absent and the syllables may start and/or end with a vowel.

Generally speaking, a syllable is a vowel-like sound together with some of the surrounding consonants that are most closely associated with it. For example, in the word *parsley*, there are two syllables [**pars.ley**] – CVCC and CVC. In the word *tarragon*, there are three syllables [**tar.ra.gon**] – CVC, CV and CVC. The process of segmenting a word into syllables is called *syllabification*. In English, the syllabification is a hard task because there

is no agreed upon definition of syllable boundaries. Furthermore, there are some words like *meal*, *hour* and *tire* which can be viewed as containing one syllable or two (Ladefoged 1993).

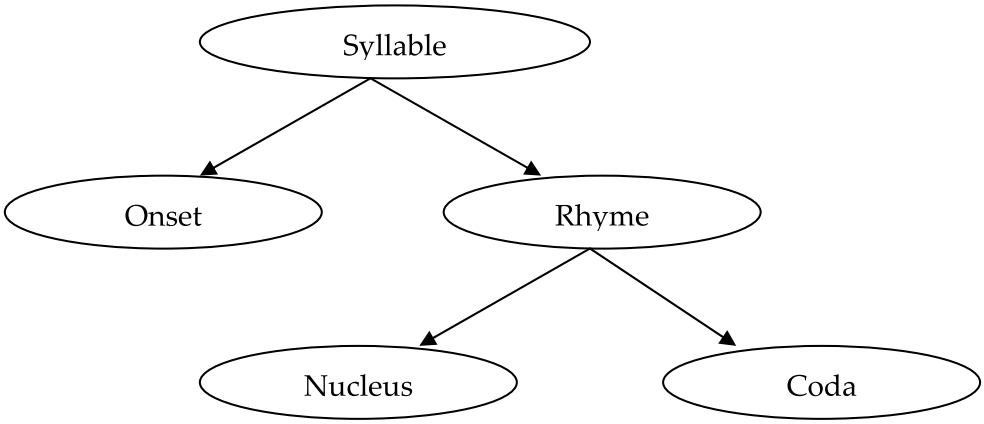


Figure 1. A Syllable's Hierarchical Structure

A syllable is usually a larger unit than a phone, since it may encompass two or more phonemes. There are a few cases where a syllable may only consist of single phoneme. Syllables are often considered the phonological building blocks of words. Syllables have a vital role in a language's rhythm, prosody, poetic meter and stress.

The syllable, as a unit, inherently accounts for the severe contextual effects among its phones as in the case of words. Already it has been observed that a syllable accounts for pronunciation variation more systematically than a phone (Greenberg 1998). Moreover, syllables are intuitive and more stable units than phones and their integrity is firmly based on both the production and perception of speech. This is what sets a syllable apart from a triphone. Several research works using syllable as a speech unit have been successfully carried out for English and other oriental languages like Chinese and Japanese by researchers across the world.

In Japanese language, for instance, the number of distinct syllables is 100, which is very small (Nakagawa et al 1999). However in a language like English, syllables are large in number. In some studies, it is shown that they are of the order of 30,000 syllables in English. The number of lexically attested syllables is of the order of 10,000. When there are a large number of syllables, it becomes difficult to train syllable models for ASR (Ganapathiraju et al 2001).

3. Agglutinative languages – Tamil

Tamil language, for instance, employs agglutinative grammar, where suffixes are used to mark noun class, number, and case, verb tense and other grammatical categories. As a

result, a large number of inflectional variants for each word exist. The use of suffixes is governed by morpho-tactic rules. Typically, a *STEM* in Tamil may have the following structure.

STEM +negative+participle +nominalization +plural +locative +ablative +inclusive

For each stem, there are at least $2^7 = 128$ inflected word forms, assuming only two affixes of each type. Actually, there may be more than two options, but there may be gaps. In contrast, English has maximally 4 word forms for a verb as in *swim*, *swims*, *swam* and *swum*, and for nouns as in *man*, *man's*, *men* and *men's*. Hence, for a lexical vocabulary of 1,000, the actual Tamil words list of inflected forms will be of the order of 1,28,000.

3.1. Inflectional morphology

The Parts of Speech (POS) categories in Tamil take different forms due to inflections. According to Rajendran (2004) morphological inflections on nouns include gender and number. Prepositions take either independent or noun combined forms with various cases like accusative, dative, instrumental, sociative, locative, ablative, benefactive, genitive, vocative, clitics and selective. Table 1 arrays a list of examples of cases and their possible suffixes.

Cases	**Suffixes**
Accusative	ஏ, ஐ
Dative	க்கு, ற்கு
Instrumental	ஆல்
Sociative	ஓடு, உடன்
Locative	இல், உள், இடம்
Ablative	இருந்து
Benefactive	க்காக, ற்காக
Genetive	இன், அது, உடைய
Vocative	ஏ
Clitics	உம், ஓ, தான்
Selective	ஆவது
Interrogative	ஆ, ஓ

Table 1. Case Suffixes used with Noun in Tamil

The verbs in Tamil take various forms like simple, transitive, intransitive, causative, infinitive, imperative and reportive. Verbs are also formed with a stem and various suffix patterns. Some of the verbal suffix patterns are shown in Table 2. Rajendran et al (2003) had done a detailed study on computational morphology of verbal patterns in Tamil.

Suffix	**Categories**	**Sub categories**	**Suffixes**
Tense	Present		கிறு, கின்று, ஆனின்று
	Past		த் , ந் , ற் , இன்
	Future		ப் , வ்
Person	First	Singular	ஏன்
		Plural	ஓம்
	Second	Singular	ஆய்
		Plural	ஈர்கள்
		Honorific	ஈர்
	Third	Male Singular	ஆன் ,அன்
		Female Singular	ஆள், அள்
		Common Plural	ஆர்கள்
		Honorific	ஆர் , அர்
		Neutral Singular	அது
		Neutral Plural	அன
Others	Causative		இ
	Verbal Noun Untensed		அல்
	Infinitive		உ
	Imperative	Plural	உங்கள்
		Negative	ஆதே, ஆது
	Passive		படு
	Future	Negative	மாட், இல்லை
	Optative		முடியும், வேண்டும், கூடும், ஆம்
		negative	முடியாது, கூடாது, வேண்டாம்
	Morpho-phonology (*Sandhi*)		ந், க், ம், ச், த்
	Plural		கள்

Table 2. Verbal Suffixes in Tamil

Adjectives and adverbs are generally obtained by attaching the suffix – **ஆன** and **ஆக** to noun forms respectively. Tamil often uses a verb, an adjective or an adverb as the head of a noun phrase. This process is called nominalization where a noun is produced from another POS with morphological inflections.

3.2. Morpho-phonology

Morpho-phonology, also known as *sandhi*, wherein two consecutive words combine by deletion, insertion or substitution of phonemes at word boundaries to form a new word is very common in Tamil. In English, one can find morpho-phonology to a limited extent. For example, the negative prefix (*in*) when attached to different words, changes according to the first letter of the word.

in + proper → improper

in + logical → illogical

in + rational → irrational

in + mature → immature

However in Tamil, use of morpho-phonology is more common among two adjacent words. The last phoneme of the first word and the first phoneme of the following word combine and undergo a transformation. Based on certain phono-tactic rules and context, there may be no transformation, or an insertion of a consonant, or a deletion of a vowel/consonant, or a substitution of a vowel/consonant. The following examples illustrate this phenomenon.

அரசு (*government*) + **பணி** (*service*) → **அரசுப்பணி**

ஆபரணம் (*ornament*) + **தங்கம்** (*gold*) → **ஆபரணத்தங்கம்**

ஒன்று (*first*) + **ஆவது** (*selective case suffix*) → **ஒன்றாவது**

In the first example, there is an insertion of a consonant (**ப்**) between the two words. The second example shows a substitution of the last consonant (**ம்**) of the first word by another consonant (**த்**). In the third example, there is a deletion of the last vowel (**உ** i.e. **று** → **ற்** + **உ**) of the first word and the consonant (**ற்**) merges with the incoming vowel (**ஆ** i.e. **ற்** + **ஆ** → **றா**) of the second word. As a result of morpho-phonology, two distinct words combine and sound as a single word. In fact, Morpho-phonology has evolved as a result of context dependencies among the phonetic units at the boundary of adjacent words or morphemes.

To a certain extent morpho-phonology is based on phono-tactics. The following rules encompass most of them.

- 'உ' removal rule: This rule states that when one morpheme ends in the vowel 'உ' and the following morpheme starts with a vowel, then the 'உ' would be removed from the combination.
- 'வ்' and 'ய்' addition rule: When one morpheme ends with a vowel from a particular a set of vowels and the morpheme it joins starts with a vowel then the morpheme 'வ்' or 'ய்' would be added to the end of the first morpheme.
- Doubling rule: According to this rule, when one morpheme ends with either 'ண்', 'ன்', 'ம்' or 'ய்' and the next morpheme starts with a vowel then the 'ண்', 'ன்', 'ம்' or 'ய்' is doubled.
- Insertion of க், ச், ட் or ப்: This rule states that when one morpheme ends with a vowel and the next morpheme begins with either க், ச், ட் or ப் followed by a vowel there is a doubling of the corresponding க், ச், ட் or ப்.

Apart from these rules, there are instances of morpho-phonology based on the context. For example, **பழங் கூடை** (old basket) and **பழக் கூடை** (fruit basket). Therefore, modelling morpho-phonology in Tamil is still a challenging issue for research.

3.3. Pronunciation in Tamil

Generally languages structure the utterance of words by giving greater prominence to some constituents than others. This is true in English where one or more syllables standout as more prominent than the rest. This is typically known as word stress. The same is true for higher level prosody in a sentence where one or more syllables may bear sentence stress or accent. Spoken Tamil language has a number of dialects. People from different parts of Tamilnadu state in India speak different accents. Harold Schiffman (2006) is a good reference for studying the grammar of spoken Tamil. As far as formal Tamil language is concerned, it is assumed that there is no stress or accent (Arden 1934, Arokianathan 1981, Soundaraj 2000) at word level and all syllables are pronounced with the same emphasis. However, there are other opinions that the position of stress in the word is by no means fixed to any syllable of individual word (Marthandan 1983). In connected speech, the stress is found more often in the initial syllable (Balasubramaniam 1980). In some studies (Asher and Keane 2005) it is shown that there is a marked reduction in vowel's duration of non-initial syllables compared to initial syllables.

4. Novel methods for improving accuracy of ASR systems

This section concentrates on two novel approaches adopted in ASR systems for agglutinative languages. In the first approach, an enhanced bi-gram or tri-gram morpheme based language model is designed to reduce the vocabulary size and reliably predict the strings in an agglutinative language. The second approach leverages the syllabic structure of the word and builds a syllable based acoustic model.

4.1. Language models

A language model is defined as the probability distribution over strings in a language. Frequencies of patterns of words as they occur in any training corpus are recorded as probability distributions. A language model is an indispensable part of ASR and machine translation systems. Language model helps in reducing the decoder's search space and in generating optimal sequence of words. For strict word-order languages like English, statistical language models have been effectively used because *trigram* or *bigram* models accurately model short distance relationship in a sentence. Other sophisticated language models also exist like class based model, distance based model and dependency based model. The performance of a statistical language model is measured in terms of *perplexity*. The perplexity refers to the branching factor in the search graph. If the perplexity is low, better the performance of a language model.

Owing to resource deficiency in text and annotated corpora in Tamil, building reliable statistical language models is a difficult task. Even with the available corpus, the agglutinative nature of the language further deepens the problem. Statistical studies using large corpora are still in the nascent stage. However, Rajendran (2006) has given a review of various recent works carried out in morphological analysis, morphological disambiguation, shallow parsing, POS tagging and syntactic parsing in Tamil.

For instance, let W_{i-1} and W_i be two consecutive words in a text. The probability $P(W_i|W_{i-1})$ is the bi-gram that measures the correlation between the words W_{i-1} and W_i. This bi-gram measure is sufficient for modeling strings of words in a language where inflectional morphology is low. However in agglutinative languages, like Tamil, a more minute measure is warranted. This issue has been successfully resolved by Saraswati and Geetha (2007) with their enhanced morpheme based language model. The size of the vocabulary was reduced by decomposing the words into stems and endings. These sub-word units (morphemes) are stored in the vocabulary separately. The enhanced morpheme-based language model is designed and trained on the decomposed corpus. A Tamil text corpus is decomposed into stem and its associated suffixes using an existing Tamil morphological analyzer (Anandan P et al, 2002). The decomposition helps reduce the number of distinct words by around 40% on two different corpora namely News and Politics. The stems and its endings are marked with a special character '#' for stems and '$' for suffixes in order to co-join them back after recognition is done.

A general morpheme based language model is one where the stem and suffixes are treated as independent words. No distinction is made between a stem and a morpheme. Figure 2 depicts the various probability measures involved in a morpheme based language model. The word W_{i-1} is split into the stem $S_{i-1}$and suffix E_{i-1}, and the word $W_i$is split into the stem S_i and suffix E_i.

In this case, the prediction of suffix E_{i-1} will be based on S_{i-1} which is strongly correlated since a stem can have a few suffixes among the possible 7 suffixes. This information is

modeled in $P(E_{i-1}|S_{i-1})$ and $P(E_i|S_i)$ which can be reliably gathered from a corpus. However, the correlation between stem S_i and the suffix of the previous word E_{i-1} is weak, because the suffix bears very little information of the next word in a sentence. But when it comes to *stem to stem* correlation, the probability $P(S_i|S_{i-1})$ can be reliably used, since there is contextual information that exists between adjacent words in a sentence and stem is the primary part of a word. In Tamil language there is strong subject-predicate agreement which leads to contextual information between suffixes of words in a sentence. This information is available in $P(E_i|E_{i-1})$

The perplexity and WER are obtained using a Tamil speech recognition system. While figure 3.a portrays the perplexity of the language models, figure 3.b compares the WER of the ASR system employing both language models.

The results has confirmed that the modified morpheme-based trigram language model with Katz back-off smoothing technique is better perplexity and lower WER on two Tamil corpora. The results confirm that the proposed enhanced morpheme-based language model is much better than the word-based language models for agglutinative languages.

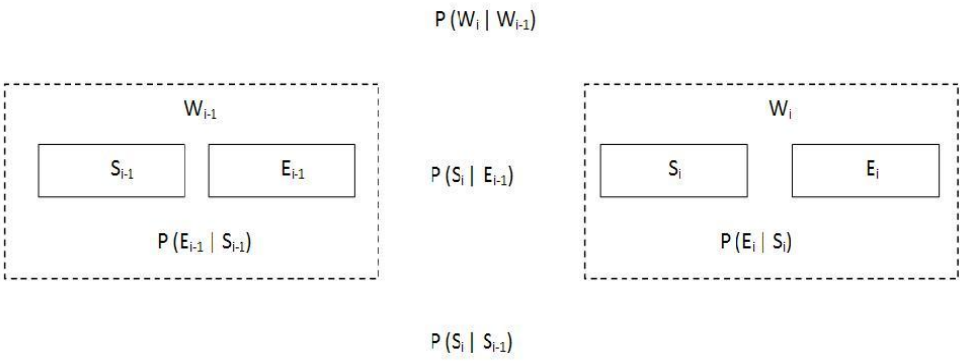
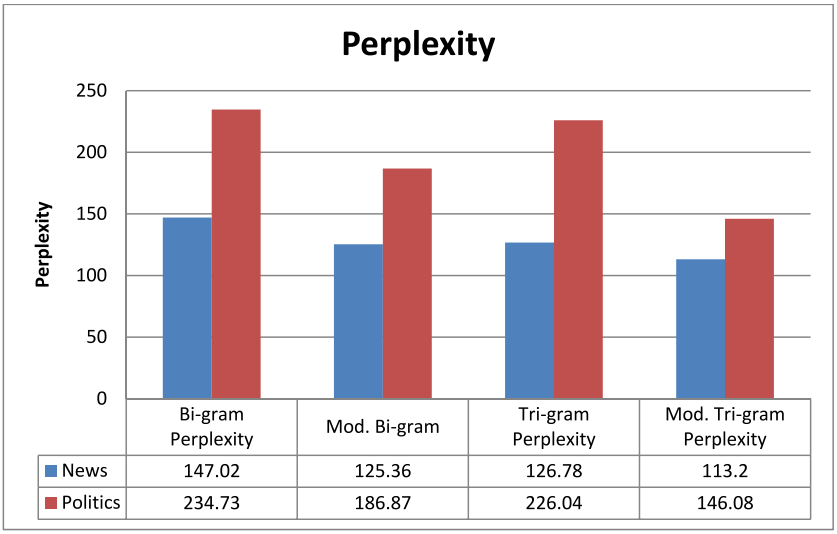
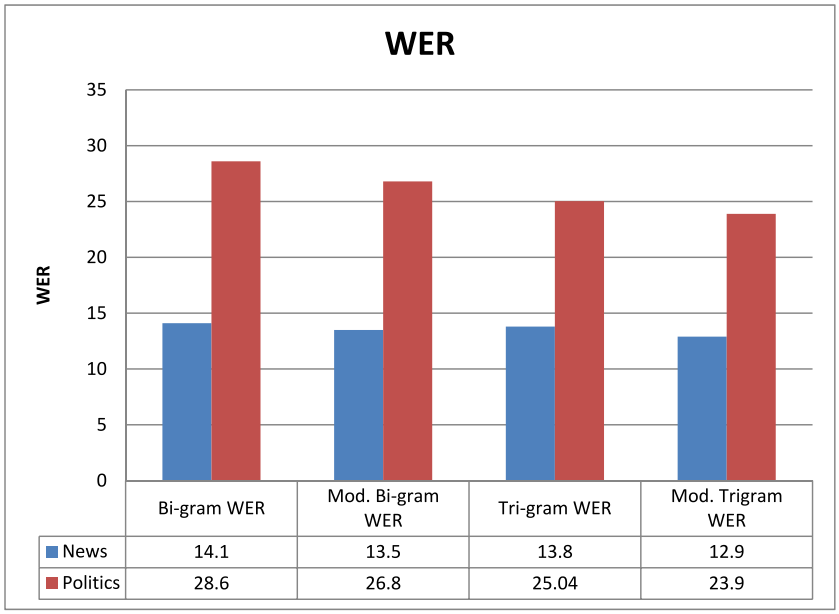


Figure 2. Morpheme based language Modeling



(a) Comparison of Perplexity



(b) Comparison of WER

Figure 3.

4.2. Syllable modeling

The importance of syllable as a unit in ASR was felt in early researches starting with (Fujimura 1975) where irregularities in phonemes have been discussed and it has been claimed that a syllable will serve as the viable minimal unit of speech in time domain.

In the paper (Greenberg 1998), it is stated that pronunciation variation in Switchboard corpus is more systematic at 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. Greenberg (1998) proposed that syllables are frequently realized in their standard or canonical form but in the case of phones, canonical realization is mostly unusual.

In the paper by Ganapathiraju et al (2001), the first successful robust LVCSR system that used syllable level acoustic unit in telephone bandwidth spontaneous speech is reported. The paper begins with a conjecture that syllable based system would perform better than existing triphone systems and concludes with experimental verification after comparing a syllable based system performance with that of a word-internal and a cross-word triphone system on publicly available databases, viz. Switchboard and Alphadigits. A number of syllable based experiments involving syllables and 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.

4.2.1. Justification for using prosodic syllable as a speech unit

Thangarajan et al (2008a) have proposed a syllable based language model for combating the agglutinative nature of Tamil language. The basic syllable consonant-vowel phono-tactics in Tamil is characterized by a regular expression shown in equation (3). 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 183 × 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.

$$RE(S) = [C]V[C[C]] \tag{3}$$

Properties that constitute prosody are fundamental frequency or formant f0 (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.

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 3. An uninflected Tamil word may comprise one to four prosodic syllables.

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

* At the maximum, two consonants can occur

Table 3. The Linguistic Rules of Tamil Prosodic Syllables

4.2.2. Formal representations of prosodic syllables

Prosodic syllables are composed of phonemes. Based on the linguistic rules tabulated in Table 3, a regular expression can be formulated as shown in equation (4).

$$RE(S) = [SV](SV|LV)[C[C]] \tag{4}$$

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.

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 phono-tactics and morpho-tactics. Hence, it is essential to estimate the number of prosodic syllables with the help of a corpus.

4.2.3. Analysis of Tamil text corpus

In this section, a Tamil text corpus provided by Central Institute for Indian Languages (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.

```
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 3. 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 were gathered. The algorithm segmented 26,153 numbers of unique prosodic syllables in the corpus. Since the text corpus used here was not clean, it contained a lot of abbreviations, digits and other foreign characters. Therefore, the prosodic syllable patterns with frequency less than 10 were eliminated. Then, it was found that there were only 10,015 numbers of unique prosodic syllables as shown in Table 4.

Details	**Frequency**
Documents	686
Sentences	455,504
Words	2,652,370
No. of unique prosodic syllables segmented by the algorithm	26,153
No. of unique prosodic syllables validated by the DFA	10,015

Table 4. Prosodic Syllables in CIIL Corpus

4.2.4. Creating context independent syllable models

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

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

Based on these considerations, eight states per HMM were decided to be adequate for the experiment. Figure 4 shows the schematic block diagram of a syllable based recognizer.

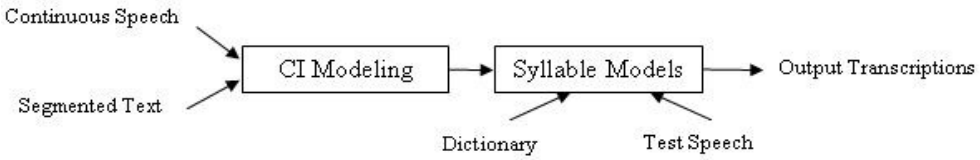


Figure 4. The Syllable Modeling and Recognition System

4.2.5. Results and discussions

For simplicity, an acoustic model was trained with 1,398 unique prosodic syllables drawn from agriculture domain. These prosodic syllables almost covered the agriculture data and the test set completely. In the experiment, the number of models to be trained was significantly reduced compared to the triphone models. The baseline triphone model had 3,171 numbers of unique triphones extracted from the transcript. The experiment was carried out using syllable based continuous speech recognition for Tamil. The dictionary and the transcripts were segmented into prosodic syllables with the proposed algorithm and models were trained. The syllable based acoustic model was deployed on a conventional continuous speech recognizer and tested with the same test set comprising 400 sentences. When comparing the WER, it is found that the WER of syllable models were considerably reduced (by 10%) compared to word models. However the triphone models performed well with a WER of 9.44%

It was also observed that in the prosodic syllable models, there were larger number of substitution errors than that of insertions and deletions whereas in the case of word models, there was a majority of deletion errors. This comparison is shown in Figure 5. 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 according to Thangarajan et al (2008b)

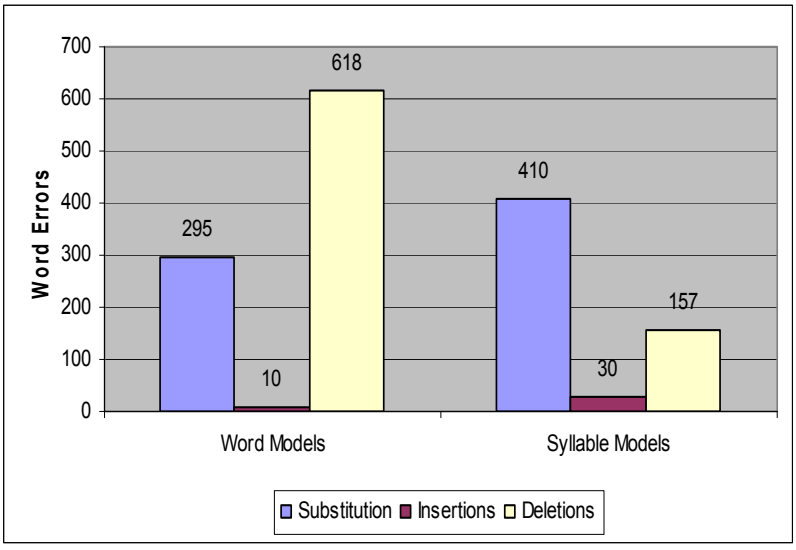


Figure 5. The Types of Word Errors in Word Models and Syllable Models

The increase in WER by 10% approximately in syllable models compared to triphone models 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. This is an avenue for future research.

5. Summary

In this chapter, the nature of agglutinative languages is discussed with Tamil language taken as a case study. The inflectional morphology of Tamil language is described in great detail. The challenges that are faced in ASR systems for such languages are highlighted. Two different approaches – enhanced morpheme based languages model and syllable based models - used in ASR for agglutinative languages are elaborated along with their results. The merits and scope for further research is also discussed.

Author details

R. Thangarajan
Department of Computer Science and Engineering,
Kongu Engineering College, Perundurai, Erode, Tamilnadu, India

6. References

[1] Anandan P., Saravanan K., Parthasarathy R., and Geetha T.V., (2002), 'Morphological Analyzer for Tamil', in the proceedings of ICON 2002, Chennai.

[2] Arden A. H. (1934), 'A progressive grammar of common Tamil' 4th edition, Christian Literature Society, Madras, India, pp. 59.

[3] Arokianathan S. (1981), 'Tamil clitics', Dravidian Linguistics Association, Trivandrum, India, pp. 5

[4] Asher R.E. and Keane E.L. (2005), 'Diphthongs in colloquial Tamil', (Hardcastle W.J. and Mackenzie Beck J. eds.), pp. 141-171.

[5] Balasubramanian T. (1980), 'Timing in Tamil', Journal of Phonetics, Vol. 8, pp.449-467.

[6] Fujimura O. (1975), 'Syllable as a unit of Speech Recognition', IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 1, pp. 82-87.

[7] Ganapathiraju A., Jonathan Hamaker, Joseph Picone, Mark Ordowski and George R. Doddington (2001), 'Syllable Based Large Vocabulary Continuous Speech Recognition', IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 4, pp. 358-366.

[8] Greenberg S. (1998), 'Speaking in Short Hand - A Syllable Centric Perspective for Understanding Pronunciation Variation', Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Kekrade, pp. 47-56.

[9] Harold F. Schiffman (2006), 'A Reference Grammar of Spoken Tamil', Cambridge University Press (ISBN-10: 0521027527).

[10] Kumar C.S. and Foo Say Wei (2003), 'A Bilingual Speech Recognition System for English and Tamil', Proceedings of Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia, Vol. 3, pp. 1641-1644.

[11] Kumar M., Rajput N. and Verma A. (2004), 'A large-vocabulary continuous speech recognition system for Hindi', IBM Journal of Research and Development, Vol. 48, No.5/6, pp. 703-715.
[12] Ladefoged, Peter. (1993), 'A course in phonetics.' 3rd edition, Fort Worth, TX: Harcourt, Brace, and Jovanovich
[13] Marthandan, C.R. (1983), 'Phonetics of casual Tamil', Ph.D. Thesis, University of London.
[14] Nakagawa S. and Hashimoto Y. (1988), 'A method for continuous speech segmentation using HMM', presented at IEEE International Conference on Pattern Recognition.
[15] Nakagawa S., Hanai K., Yamamoto K. and Minematsu N. (1999), 'Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition', Proceedings of International Workshop Automatic Speech Recognition and Understanding, pp. 393-396.
[16] Nayeemulla Khan A. and Yegnanarayana B. (2001), 'Development of Speech Recognition System for Tamil for Small Restricted Task', Proceedings of National Conference on Communication, India.
[17] Plauche M., Udhyakummar N., Wooters C., Pal J. and Ramachadran D. (2006), 'Speech Recognition for Illiterate Access to Information and Technology', Proceedings of First International Conference on ICT and Development.
[18] Rajendran S, Viswanathan S and Ramesh Kumar (2003) 'Computational Morphology of Tamil Verbal Complex', Language in India, Vol. 3:4
[19] Rajendran S (2004) 'Strategies in the Formation of Compound Nouns in Tamil', Languages in India, Volume 4:6
[20] Rajendran S (2006) 'Parsing In Tamil: Present State of Art', Language in India, Vol. 6:8.
[21] Saraswathi S. and Geetha T.V. (2004), 'Implementation of Tamil Speech Recognition System Using Neural Networks', Lecture Notes in Computer Science, Vol. 3285.
[22] Saraswathi S. and 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 Vol. 6 No. 3, Article 9.
[23] Soundaraj F. (2000) 'Accent in Tamil: Speech Research for Speech Technology', In: (Nagamma Reddy K. ed.), Speech Technology: Issues and implications in Indian languages International School of Dravidian Linguistics, Thiruvananthapuram, pp. 246-256.
[24] Thangarajan R., Natarajan A. M. and Selvam M. (2008a), 'Word and Triphone based Approaches in Continuous Speech Recognition for Tamil Language', WSEAS Transactions on Signal Processing, Issue 3, Vol.4, 2008, pp. 76-85.
[25] Thangarajan R. and Natarajan A. M. (2008b), 'Syllable Based Continuous Speech Recognition for Tamil Language', South Asian Language Review (SALR), Vol. XVIII, No. 1, pp. 71-85.
[26] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon (2001), 'Spoken Language Processing - A Guide to Theory, Algorithm and System Development', Prentice Hall PTR (ISBN 0-13-022616-5).