MultiCraft

International Journal of Engineering, Science and Technology
Vol. 2, No. 5, 2010, pp. 244-257

INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND TECHNOLOGY

www.ijest-ng.com

Design of language models at various phases of Tamil speech recognition system

S. Saraswathi[1]*, T.V. Geetha[2]

[1]Department of Information Technology, Pondicherry Engineering College, Pondicherry, INDIA*
[2] Department of Computer Science and Engineering, Anna University, Chennai, INDIA
[*]Corresponding Author: e-mail: swathimuk@yahoo.com,

Abstract

This paper describes the use of language models in various phases of Tamil speech recognition system for improving its performance. In this work, the language models are applied at various levels of speech recognition such as segmentation phase, recognition phase and the syllable and word level error correction phase. The speech signals were segmented at phonetic level based on their acoustic characteristics. The wrongly identified segmentation points were detected and corrected using articulatory feature based phoneme language model. The segmented signals were mapped to their phonemes. The ambiguities in the recognized phonemes were reduced by using inter and intra word based language models. The recognized phonemes were grouped together to form syllables and then words. The errors in the syllables and words were detected and corrected by using the syllable and morpheme based language models developed for Tamil language. The performance of the Tamil speech recognition system was improved by using the language models at different phases of speech recognition. Recognition rate of 74.11% was obtained by applying language models at segmentation phase, which was further improved to 84.11% at phoneme recognition phase and finally to 87.1% at syllable level and word level recognition phase. Thus the use of language models has drastically reduced the error rates at various levels and improved the recognition rate of Tamil speech recognition system.

Keywords: Language model, articulatory features, morphemes, syllables.

1. Introduction

There are various techniques for the development of a speech recognition system such as the template matching (Marcus *et al.*, 1994), knowledge based (Lahiri 1999; Reetz 1999) and statistical approaches (Gales, 1998). The selection of a technique for recognition depends on the language to be recognized. The characteristics of the language such as its phoneme types, allophonic characteristics; syllable patterns and inflectional characteristics decide the type of the technique to be used for recognition. The distinct characteristics of a language can be analyzed from the order of occurrence of phonemes, syllable patterns and words in that language. The use of statistical approach - language models will help us to detect the co-occurrence of phoneme, syllable patterns and words in a language. One of the best-known statistical approaches for speech recognition is the Hidden Markov Model (HMM), which uses word or phoneme as the modeling unit (Rabiner 1989; Young 1996; Axelrod *et al.*, 2004). HMM is automatically trained from data, resulting in improved accuracies, if training and decoding are treated in the same framework. HMM is a double stochastic model, in which the generation of the underlying phoneme string and the frame-by-frame, surface acoustic realizations are both represented probabilistically as Markov processes. HMM uses language models to reduce the search space and resolve acoustic ambiguity. However, in HMM, the probability distribution related to duration of each phonetic state does not match correctly with speech phoneme duration. Obviously, speech is not really a string of stationary states and the acoustic realization of a phoneme can be dynamically modified by adjacent phonemes. Moreover, HMM is an inaccurate model as it relies on the use of statistics to model the variability of speech such as co-articulation effects and inter speaker differences and the technique does not deal with the actual mechanism of speech production or perception (Bilmes, 2002). Speech recognition approaches that incorporate linguistic and articulatory knowledge as an integral part of the speech recognition system have gained more attention and interest in recent times (Metze *et al.*, 2002; Stuker *et al.*, 2003). These types of approaches are referred to as

artificial intelligence approaches, in which knowledge based information regarding the articulatory features of the phonemes are combined with statistical information to correct the errors that occur during the recognition phase. A limitation in the number of phonemes and the existence of unique articulatory features to identify those phonemes motivated researchers to classify the signals at the phonetic level based on their articulatory phonetic features (Ali *et al.*, 1999). This approach is currently used to improve the performance of large vocabulary speaker independent speech recognition. Statistical information regarding the units to be recognized can be obtained by using the concept of Statistical Language Models (SLM). A statistical language model is one of the major modules in the speech recognition process that is used to predict and correct errors that occur at various levels of speech recognition (Gotoh, 2000). They can be easily integrated with other components in the recognition system to improve the performance of error correction. Hence language models were applied at various phases of speech recognition in Tamil language.

The paper is organized as follows: In section 2, the proposed system details are discussed, in section 3, the details regarding the use of language models in speech segmentation is discussed. In section 4, the use of language models for phoneme recognition is discussed. In section 5, the use of syllable and morpheme based language models to correct the errors at syllable and word level is discussed. In the final section the results are analyzed.

2. Proposed Work

An Artificial Intelligence based speech recognition approach which is a hybrid of knowledge based approach and statistical approach is designed to recognize speech in Tamil language. Language models can be applied from all levels of granularity, from the phoneme level, to syllable level, to morpheme level and till the word level in the proposed Tamil speech recognition system. The type of language model to be applied depends on the language being used. Current research in speech recognition focuses on methods to improve the performance of the speech recognition system using improved language models. The characteristic of a language decides the type of language model designed for it. Most languages are generally assumed to have a basic word order. For example, in English SVO (subject-verb-object) is the commonly used fixed word pattern for normal type of sentences. Hence the use of n-gram language models and class based language models based on the syntactic category of the words and the position of the word within the sentence will improve the performance of the English speech recognition system (Charniak *et al.*, 2003). Some languages like Tamil which are partially free word order in nature, require different language modeling approaches to extract linguistic characteristics that would improve the performance of speech recognition. This characteristic of Tamil languages would require the use of a language model at finer levels of linguistic granularity. This led to the search for new approaches to capture phoneme, morphological and syntactic constraints and intra and inter word associations between these linguistic units. This in turn resulted in the development of new models to tackle these distinguishing characteristics of languages like Tamil. Phoneme was chosen as the fundamental unit, for building the language models for Tamil due to the restriction in the number of phonemes in most relatively free word order languages, compared to the large number of morphological variants of words. There are 50 phonemes in Tamil (12 vowels and 38 consonants). Therefore phonemes as basic units would be a natural choice for speech recognition of partially free word order languages. The language models developed at phoneme/syllable/word level will detect and correct the errors in Tamil language. The use of language models at the phoneme level segmentation and recognition phase itself would try to reduce the errors that are propagated to word level recognition process. This motivated us to use language modeling approaches at various phases of speech recognition.

The basic block diagram of the system is shown in Figure1. The preprocessing module will eliminate the noise and silence from the signals and segment the signals from sentence level to word level. The pre-processed signals are then segmented at the phonetic level using a phoneme segmentation algorithm. The segmentation algorithm designed for this work, identifies the under segmented and over segmented signals based on their spectral, temporal and acoustic characteristics and by using category-based phoneme level language models. The segmented phonemes are recognized based on their acoustic features and further processed using inter and intra word based language models. Syllable and morpheme based language models are used to avoid the ambiguities in the recognition process and in general to improve the performance of the system.

3. Language models in Speech segmentation

There are various approaches to perform speech segmentation: based on syllables using group delay and MFCC features on HMM (Hedge *et al.,* 2004) and based on phonemes - using linguistically constrained segmentation methods (Pellom *et al.*, 1998), text independent method of segmentation (Aversano *et al.*, 2001) and cluster based phonetic segmentation (Eberman 1996). This paper is based on the text independent phoneme segmentation algorithm proposed by Guido Aversano, which was modified to suit segmentation of Tamil phonemes. By using the text independent segmentation algorithm on Tamil speech, it was found that some of the segmented phonemes were over-segmented and some were under-segmented. The under-segmented signals were detected based on their duration and acoustic characteristics and were again re-segmented that led to the design of two level segmentation algorithm (Saraswathi *et al.*, 2006a). The over-segmented signals were checked for their likelihood of joining with the neighboring segments based on their spectral characteristics, which led to the design of phoneme combination algorithm (Saraswathi *et al.*, 2006b).

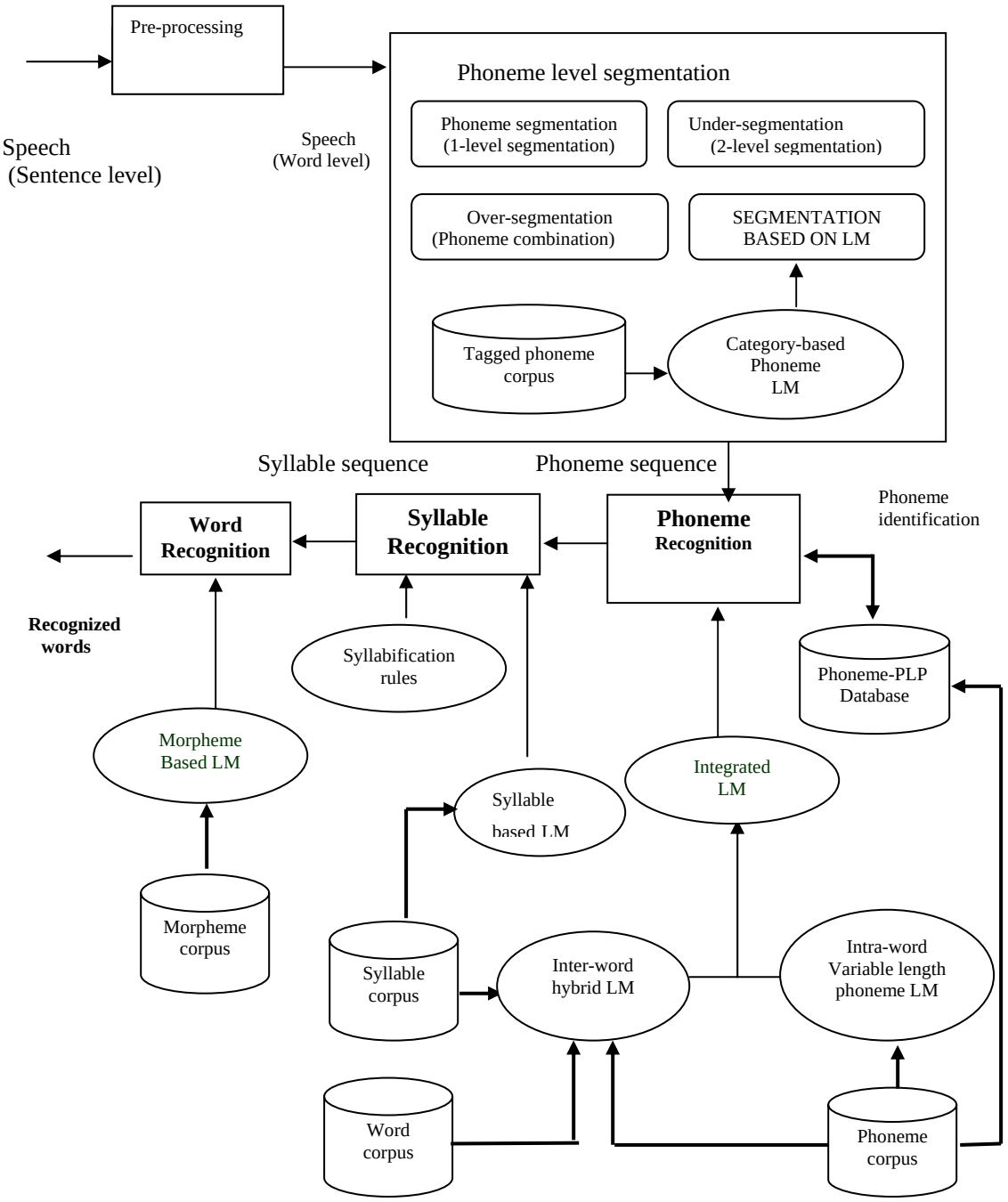

Figure 1. Block diagram of the Tamil Speech Recognition system

Incorrectly detected segmentation points were corrected by using category-based phoneme language models at the segmentation points as discussed in the next section

Language models for segmentation: Due to segmentation errors, the segments obtained may not correspond to actual phonemes. In this work, we have used information associated with phonological categories such as voicing, place and manner of articulation to alter the segmentation points. We consider the co-occurrence pattern of the phonemes based on their type (vowel/consonant), voicing characteristic (voiced/unvoiced) as well as the place and manner of articulation to detect the incorrect segmentation points. The signal segments obtained from segmentation algorithm are mapped to their phoneme category based on both the acoustic features and hand-coded articulatory features (Ali, 1998; Ali, 1997) of the phoneme segments, as show in Table 1.

Table 1. Features for identifying the phoneme categories

S.No	Feature	Identification
1	Formant Frequencies	Vowels
2	Spectral center of gravity (SCG)	Place of articulation (dental, alveolar, palatal)
3	Maximum normalized spectral slope	Labial and dental
4	Most dominant peak frequency (MPPF)	Palatals and nonpalatal
5	Duration of unvoiced portion (DUP)	Voicing/ unvoiced – fricatives
6	Formant Transitions and Burst frequency	Stop consonants

A database that contains the acoustic features of the phonemes along with their articulatory features was maintained. The phonemes were classified into groups based on the characteristics of their acoustic and articulatory features. The segmented speech signals were mapped to a phoneme type based on their acoustic and articulatory features. The signal sequence may contain certain improbable co-occurrences, which is determined by using category- based phoneme language model, where the category corresponds to the articulatory features of the phonemes. This indicates that there is an error in the segmentation and calls for re-assignment of the segmentation points. To build the category-based language model, a Tamil newspaper text corpus with 5 lakh words was selected. The Tamil text characters were converted to their corresponding graphemes and the graphemes were mapped to their corresponding phonemes in International Phonetic Alphabet (IPA) text format based on rules proposed by Kothandaraman (1997). The phonemes were tagged based on their type (vowel/consonant), voicing characteristics and place and manner of articulation. The category-based language model was used to determine the probability of occurrence of a phoneme based on the previous sequence of phonemes which was estimated using the equation 1.

$$P(Ph_i \mid Ph_{i-1}.....Ph_1) = P(Ph_i \mid Ar_{i-1}.....Ar_1) = P(Ph_i \mid Ar_i).P(Ar_i \mid Ar_{i-1}.....Ar_1) \quad (1)$$

where Ph_i corresponds to the phoneme recognized at the i^{th} segment point and v_i corresponds to the articulatory feature of the phoneme Ph_i. Thus a phoneme is recognized based on the category to which it is mapped on and on the history of the category of the previous phoneme sequence. Bigram category-based language model was applied on the given corpus to detect the possible co-occurrence of phoneme sequences in Tamil language. The category-based language model determines the improbable sequence of phoneme co-occurrence based on voicing/unvoicing and place and manner of articulation. Thus the model has been used to find certain phoneme sequences that are not possible. In general as proposed by Kothandaraman (1997) and learned from the category-based bigram phoneme language models built for Tamil, the following phoneme sequences are not possible in Tamil

1. The consonant alveolar- approximant, nasal, trill, tap and fricative do not occur in word starting position.
2. The consonant bilabial- fricative and plosive do not occur in word starting position.
3. The consonant velar-plosive, velar- fricative, palatal - nasal, velar-nasal, dental-fricative, alveolar-plosive, alveolar-trill and bilabial-nasal do not occur in word endings.
4. Consonants other that alveolar-nasal, bilabial-fricative are not possible after the consonant alveolar-approximant.
5. No consonants occur after alveolar-plosive, bilabial-fricative, dental-fricative, glottal-fricative and alveolar-trill. Only vowels occur after them.
6. Consonant other than alveolar-plosive and post alveolar-fricative do not occur after palatal-nasal.
7. Consonants other than velar-plosive do not occur after velar-nasal.

When identifying the phonemes at the segmentation points, an analysis was done based on the language models to check if it is a valid phoneme to occur at that position. If it is not a valid phoneme then it implies an error in segmentation at that point. Some of the possible errors are

1. Due to the joining of some identical phonemes, which may occur, in gemination in Tamil. Such segments are again re-segmented based on the amplitude peaks in them.
2. Due to wrong identification of segmentation points, by considering the maximum amplitude peaks as the threshold value for segmentation. Such segments are combined with the neighboring segments and they are again resegmented based on the next lower level amplitude peaks present in them.

Thus most of the errors due to over-segmentation and under-segmentation were reduced using language models. The speech

corpus for testing purposes was recorded under clean acoustic conditions in 16 kHz, 16-bit audio quality. Read speech data was recorded from 100 speakers - 60 male and 40 female speakers, of different age groups in the range 15-50. The articles in a Tamil newspaper that contained 200 sentences were read aloud by the speakers and they were recorded. The speech signals were transcribed at the word level. 2500 words were obtained from the speech corpus of each speaker. The test material contained a total of 20,000 phonemes for each speaker and the segmentation algorithm had technical prerequisites to recognize of about 15,920 of them, which gives Phoneme segmentation rate of 79.6%.

4. Phoneme Recognition based on Language models

There are various approaches to phoneme recognition: based on time delay neural networks (Waibel *et al*., 1989), based on LPC spectral matching measures (Shikano, 1985) and based on acoustic phonetic features (Ali 2001). This work performs phoneme recognition based on the acoustic features of the phonemes and uses language models to decrease the errors in phoneme recognition phase. The first stage in phoneme recognition process is to aggregate all the phoneme signals, by segmenting the speech signals and then compare the Perceptual Linear Prediction (PLP) features of the segmented phoneme signals with the PLP feature of the existing phonemes in the database. The database was built by manually segmenting the words collected to cover all phonemes in Tamil language, extracting their PLP co-efficient and storing them in a database. A Euclidean distance measure was used to compare the incoming phoneme signals with phoneme signals available in the database. However, the phonemes recognized by using the PLP features led to errors such as phoneme insertion, deletion and replacement, due to coarticulation effect associated with the phonemes. These types of errors typically degrade the performance of the speech recognition. Hence, we used the concept of language models to correct phoneme recognition errors. In this work, intra and inter-word based language models have been used to predict the presence of a phoneme based on its preceding context. This prediction probability is used to reduce errors in the recognition process. The language model used in the recognition process considers the probability of the presence of a phoneme P based on the preceding phoneme sequence P_s of the word in which P occurs. Depending on the length of the word W in which P occurs and the position of P in W, the sequence length of P_s considered for probability prediction of P varies. In turn the probability of the sequence P_s is determined by the last syllable S of the previous word W_p. In turn the independent probability of occurrence of syllable S as a last syllable is determined by considering the probability of occurrence of S as last syllable in any word. These features of phoneme context led to the design of a new language model for phoneme recognition based on the context of phonemes, syllables and words as shown in Figure 2.

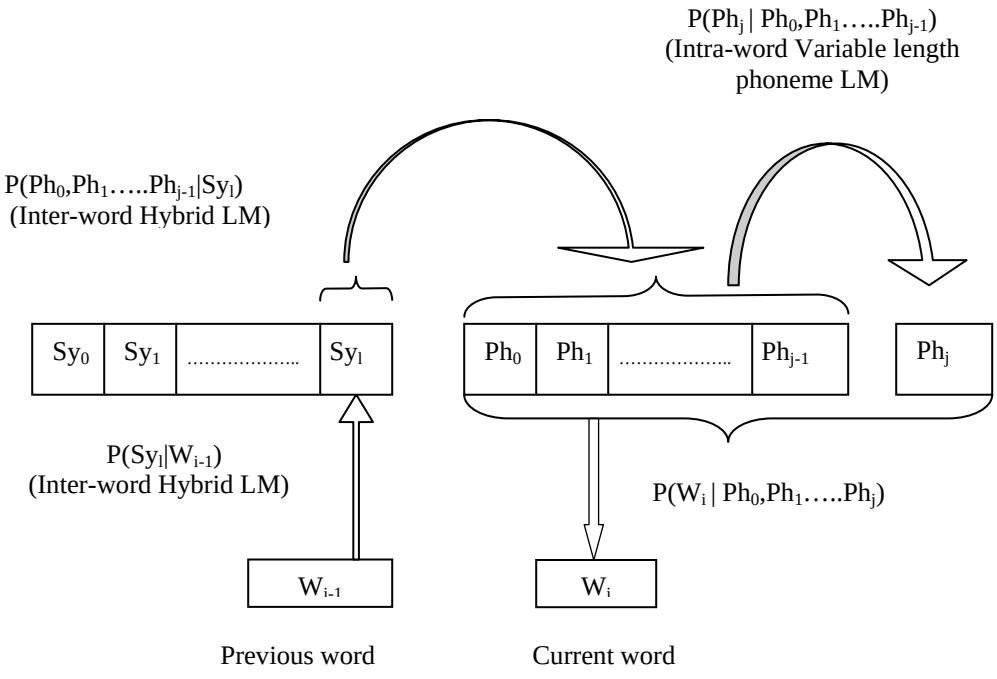

Figure 2. Design of language model for phoneme recognition

The probability of occurrence of a word (W_i) and its associated phoneme sequence $Ph_0,Ph_1,.....Ph_j$ is evaluated using the equation 2.

$$P(W_i \mid Ph0, Ph1, \ldots Ph_j) = \prod_{k=1}^{j} P(Ph_k \mid Ph0, \ldots Ph_{k-1}).P(Ph0, \ldots Ph_{k-1} \mid Sy_l).P(Sy_l \mid W_{i-1}) \quad (2)$$

where $Ph_0, Ph_1, \ldots Ph_j$ corresponds to the phoneme sequence recognized for the word W_i and Sy_l corresponds to ending syllable pattern of the previous word W_{i-1}.Thus the language model designed for the phoneme recognition system in our work is a combination of

- Intra-word variable length phoneme based language model: To detect the phonemes based on previous sequence of phonemes of the current word.
- Inter-word Hybrid language model: To detect the phonemes based on the syllable pattern and the word present in the preceding position.

Intra word phoneme based language model: In this section, we first describe the use of preceding sequence of phonemes of the word as the context to determine the probability of occurrence of the phoneme under consideration. Since the length of the word in which the phoneme under consideration occurs and the position of this phoneme in the word varies, the sequence length of the preceding phonemes taken as context for the probability prediction of phoneme under consideration varies. Thus the varying length of the preceding phoneme sequence results in an intra-word variable length language model. The bigram language model will predict the occurrence of a large number of phonemes as next phoneme thus leading to ambiguity in prediction. In other words, the use of bigram language models results in a higher perplexity value for detection of next possible phoneme. The use of bigram language model obtained a perplexity value of 22.3 for predicting the phonemes as shown in Table 2.

The training was done on a newspaper text corpus that contained 5 lakh words. The words were mapped to their phoneme sequences using the word to phoneme mapping algorithm designed for Tamil. The bigram probability value for the phoneme sequences were analyzed for the training set. The testing was done on the phoneme segments obtained from the speech signals segmented using the phoneme level segmentation algorithm discussed in the previous section.

The use of variable length language model obtained perplexity value of 4.1 for predicting the succeeding phoneme as shown in Table 3. This reduction in perplexity essentially reduces the ambiguity in phoneme detection. In addition, in order to improve the prediction probability of the sequence of phonemes used in the variable length language model the inter-word hybrid language model was designed.

Inter word hybrid language model: Large vocabulary speech recognition and search was implemented based on a Phonetic approach by Seide *et al.*, (2004). Closer inspection of the results showed that word-lattice based search and phonetic search produced essentially complementary search results: Phonetic search tends towards low miss rates with many false alarms, and word-level search tends towards few false alarms but high miss rates. This motivated the use of a hybrid language model based on a word and its phoneme sequence, by Bazzi (2000), to improve speech recognition. The probability of occurrence of a word was evaluated based on the probability of the phoneme sequence that forms a part of the word, which is essentially an intra word language model. The main drawback in this hybrid approach was that, it led to lower recognition rate, as it could not predict the errors due to deletion and replacement of phonemes in the recognized sequence. In order to overcome this drawback we designed an inter-word hybrid language model. This hybrid language model improved the performance of the phoneme recognition system. Here, we assume that some level of language dependent error corrections of phoneme sequences have already taken place at the segmentation phase. This hybrid model essentially predicts the probability of occurrence of the phoneme sequence used as context in variable length language model based on the syllables at the end of the previous word and independent probability of syllables occurring as an ending syllable in a word. In general the prediction of phonemes at an instant also depends on the information regarding the history of words recognised so far. The information regarding the syllable pattern that occurs at the ending of the previous word has some impact in recognising the phoneme for the current word as the occurrence of word depends on the ending of the previous word in a highly inflectional language, Tamil. This led to the design of inter-word hybrid language model for phoneme recognition.

Table 2. Perplexity values in predicting phoneme using bigram language model

Phoneme (IPA format)	Grapheme	Perplexity value
/ɳ/	<ண>	24
/ k /	<க்>	30
/ n/	<ன்>	29
/d̪ /	<த்>	20
/ɑ:/	<ஆ>	27
/ʌy/	<ஐ>	19
/ ʈ, /	<ட்>	29
/ o:/	<ஓ>	25
/ m /	<ம்>	26
/ ɭ /	<ள்>	20
/ p /	<ப்>	27
/ʌ/	<அ>	27
/ i /	<இ>	26
/u/	<உ>	24
/s/	<ச்>	17
/e:/	<ஏ>	23
/o/	<ஒ>	21
/ j/	<ய்>	29
/g/	<க்>	14
/ŋ/	<ங்>	12

Table 3. Perplexity values in predicting phoneme using intra-word variable length phoneme language model

Phoneme **Sequence (IPA format)**	Grapheme **Sequence**	**Perplexity value**
/k i ʈ /	<க்இட்>	7
/^ ɽ i t̪ /	<அட்இத்>	3
/ ʊ i ɽ /	<வ்இட்>	7
/ ɳ i n d̪/	<ந்இன்ற் >	3
/^ ð i x ɑ: r/	<அத்இக்ஆர்>	2
/p ^ ɽ /	< ப்அட் >	5
/ ʊ i ɻ /	<வ்இ ழ் >	4
/ ɳ i r /	<ந்இர்>	3
/ t̪ i ɽ /	<த்இ ட்>	5
/ʊ i ɽ /	< வ்இட் >	6

Table 4. Perplexity values for phoneme prediction using integration of intra-word and inter-word language models

Syllable–Phoneme sequence (IPA format)	Syllable –Grapheme sequence	Perplexity value
< ʈi > /k i ʈ /	< ட்இ> / க்இட்/	2
< ku > /k ^ ɽ i t̪ /	<க் உ> / க்அட்இத் /	1
< ki > / ʊ i ɽ /	< க் இ> / வ்இட்/	3
< ɭil> / ɳ i n ɖ/	< ள்இல்> / ந்இன்ற் /	1
< ʈ^> /^ ð i x ɑ: ɾ/	< ட்அ> / அத்இக்ஆர்/	2
<k ^> /p ^ ɽ /	<க்அ> / ப்அட் /	4
< t̪^ > / ʊ i ɻ /	<த்அ> / வ்இ ழ் /	2
< ɻum > / ɳ i ɾ /	< ழ்உம் > / ந்இர்/	2
< ^l > / t̪ i ɽ /	<அல்> / த்இ ட்/	1
<jɯ> /ʊ i ɽ /	< ய்உ> / வ்இட்/	5

To construct the intra and inter-word based language models, the newspaper text corpus with 5,00,000 words was taken and the phonemes that form the words were obtained. The syllables that constitute the words were built from the phonemes. The words, phonemes and syllable sequences were stored as

கிட்டத்தட்ட	<k i ʈt̪^ t̪ t̪^ ʈt̪^ >	<kiʈ t̪^ʈt̪^t̪ t̪^>
(word)	<phoneme sequence>	<syllable sequence>

The probability of the occurrence of a phoneme in a word was evaluated using a combination of the variable length phoneme language model and the hybrid phoneme/syllable/word based language model as shown in Equation 2. An overall perplexity value of 2.09 was obtained by integrating the variable length phoneme language model with the hybrid model as shown in Table 4 when tested on the phoneme segments obtained from the phoneme level segmentation algorithm. A lower perplexity value in predicting the phoneme reduced the ambiguity in identifying the succeeding phoneme.

5. Error correction using syllable and morpheme based language models

In this section, we move from phoneme based language models to language models based on larger units such as syllables and morphemes. In the previous section, we had obtained a probable sequence of phonemes corresponding to a word. In this section, we use specially designed language models for error correction based on probability of syllables and morphemes in the formation of words.

5.1 Syllable based language models

A basic Tamil pronunciation unit is a syllable that can be represented in the form of (C) V (C) (C) where C corresponds to Consonant and V corresponds to Vowel. Following are the possible syllable patterns in Tamil

V, CV, CVC, CVCC, VC, VCC

For performing syllable based language models, Tamil news paper text corpus with 5 lakhs word was selected. A grapheme to syllable converter was used to convert the text data to sequence of syllables. In order to implement a Tamil Grapheme to syllable conversion system many language specific problems need to be solved. A rule based grapheme to syllable converter was used to perform the conversion. Following were some of the rules used to perform grapheme to syllable conversion

1. Nucleus can be Vowel(V) or Consonant (C)
2. If onset is C then nucleus is V to yield a syllable of type CV
3. Coda can be empty or C
4. If characters after CV pattern are of type CV then the syllables are split as CV and CV.
5. If the CV pattern if followed by CCV then syllables are split as CVC and CV.
6. If the CV pattern is followed by CCCV then the syllables are split as CVCC and CV
7. If the VC pattern is followed by V then the syllables are split as V and CV.
8. If the VC pattern is followed by CVC then the syllables are split as VC and CVC

The phonemes obtained after the application of phonological rules are converted to the syllable sequences based on the syllabification rules listed above. The probability of generation of the syllables (Sy) from the phoneme sequence is given in

equation 3.

$$P(Sy) = \arg \max_{Sy} \{P(Sy \mid Ph_1^{Ar_1}, Ph_2^{Ar_2} ..... Ph_n^{Ar_n})\} \quad (3)$$

where Ph_1^{Ar1} corresponds to the phoneme Ph_1 with articulatory feature Ar_1 and '*n*' take any value between 1 to 4, as the maximum possible phonemes to form syllable is 4. Based on the rule based algorithm for mapping the group of graphemes to form syllables and the probability of generating the syllables from the phoneme sequences, the syllables were recognized. Recognition of some syllable sequences is not possible due to error in recognized phonemes. They are detected using the articulatory-acoustic features of the phonemes.

Various studies (Chang 2002; Greeberg 2003) show that there is a systematic relation between articulatory-acoustic features and syllables. The way articulatory-acoustic features give insight into nature of pronunciation variation at the level of syllables is discussed by Greenberg (Greeberg 2003) was used in formation of syllables from the phoneme sequences. A few points given by him are

- In a syllable, onsets are often produced canonically, whereas the nucleus and coda are often reduced, the coda often being deleted.
- Voicing is the articulatory foundation of the syllabic nucleus
- It is rare for two segments of the same manner class to occur in adjacent positions within a syllable.
- Based on place of Articulation we can distinguish among words, particularly at onset. In coda position there is a general preference for central place of articulation.

To extract the acoustic and articulatory information for the syllables, templates are derived from manual transcriptions by rewriting the strings of phonetic segments in terms of articulatory-acoustic features and bundling them together as syllables. The features for manner of articulation, place of articulation and voicing are considered. Syllable sequences and their templates that are not possible in the Tamil language are listed in Table 5. The syllable patterns that are not possible in Tamil are identified based on the rules given in Table II and they are corrected using syllable based language model.

$$P(Sy_i \mid Sy_{i-1}) = \frac{C(Sy_{i-1}, Sy_i)}{C(Sy_{i-1})} \quad (4)$$

where $C(Sy_{i-1},Sy_i)$ is the count of no. of syllable sequences Sy_{i-1},Sy_i in the given corpus and $C(Sy_i)$ is the count of syllable Sy_i in the given corpus.

The probability of occurrence of the syllable is thus calculated by the equation (5)

$$P(Sy_i) = \arg\max_s \{P(Sy_i \mid Ph_1^{Ar_1}, Ph_2^{Ar_2}, ...Ph_n^{Ar_n}) P(Sy_i \mid Sy_{i-1})\} \quad (5)$$

where Ph_i^{Ari} corresponds to the phoneme Ph_i with articulatory feature Ar_i. Based on the previous syllable pattern and the generation of syllable from the phoneme sequence the current one was identified and the ambiguities were resolved. The syllables were combined to form words and the errors in word level were detected and corrected using morpheme based language model as discussed in the next section.

5.2 Morpheme based language models

Tamil words typically have more morphological patterns than English words. For example, a Tamil word will often contain the following, easily identifiable, constituent parts: a stem, which can be thought of as responsible for the nuclear meaning of the verb, attached to which may be zero or more derivational prefix (es) and zero or one suffix, which together form a word. The stem often acquires an entirely new lexical meaning with the presence of these affixes. Of most relevance to language modeling, however, is the inflection (inflectional suffix), which is appended to the stem and which determines the grammatical case, gender (masculine, feminine, or neuter), number, etc. of the word. The presence of the inflection results in many different word forms for a word in Tamil compared to English (Arden *et al.,* 1969). The direct consequence of this is the coverage of more words in Tamil vocabulary than that of the same sized English vocabulary. The size of the Tamil vocabulary was reduced by using morpheme based language models (Saraswathi *et al.,* 2004). The Tamil newspaper text corpus that had 5 lakh words with 1,49156 distinct words contained only about 93795 distinct stems and 1515 distinct endings. The use of morpheme based language model designed by us (Saraswathi *et al.,* 2007) will predict the errors in recognised word sequence due to replacement of phonemes and correct them to improve the recognition rate of the Tamil speech recognition system.

Table 5. Syllable patterns not possible in Tamil

SYLLABLE	TYPE	TEMPLATE		
		VOICE	MANNER	Place
After any short vowel 'ழ்'	VC	+voi	vow_appr	Vow_alv
After any long vowel 'வ்'	VC	+voi	vow_appr	Vow_labiode
'ங்அழ்' (any V between C)	CVC	+voi	Nas_vow_appr	vel_vow_alv
'ண்அழ்' (any V between C)	CVC	+voi	Nas_vow_appr	retr_vow_alv
'ன்அழ்' (any V between C)	CVC	-voi +voi	Nas_vow_appr	alv_vow_alv
'ஞ்அழ்' (any V between C)	CVC	+voi	Nas_vow_appr	pal_vow_alv
'ங்அவ்' (any V between C)	CVC	+voi	Nas_vow_appr	vel_vow_labioden
'ண்அவ்' (any V between C)	CVC	+voi	Nas_vow_appr	retr_vow_ labioden
'ன்அவ்' (any V between C)	CVC	-voi +voi	Nas_vow_appr	alv_vow_ labioden
'ஞ்அவ்' (any V between C)	CVC	+voi	Nas_vow_appr	pal_vow_ labioden
'ற்ஊவ் '	CVC	+voi	Tri_vow_appr	Alv_long-close-rounded_labioden
'ற்ஒவ்'	CVC	+voi	Tri_vow_appr	Alv_short-closemid-rounded_labioden
'ர்அழ்' (any V between C)	CVC	+voi	tap_vow_appr	Alv_vow_alv
'ட்அழ்' (any V between C)	CVC	+voi	tap_vow_appr	Retr_vow_alv
'ல்அழ்' (any V between C)	CVC	+voi	Lat-appr_vow_appr	Alv_vow_alv
'ள்அழ்' (any V between C)	CVC	+voi	Lat-appr_vow_appr	Retr_vow_alv
'ங்' followed by any vowel (' ங்அ')	CV	+voi	Nas_vow	Vel_vow
'ண்' followed by any vowel ('ண்அ')	CV	+voi	Nas_vow	Retr_vow
'ழ் ' FOLLOWED BY ANY VOWEL (' ')	CV	+voi	Nas_vow	Bilab_vow
' ' followed by any vowel('ள்அ')	CV	+voi	Lat-appr_vow	Retr_vow
'ற்' FOLLOWED BY ANY VOWEL (' ')	CV	+voi	trill_vow	Alv_vow
'ன்' FOLLOWED BY ANY VOWEL (' ')	CV	-voi +voi	nas_vow	Alv_vow
'அண்ற்' any vowel followed by nasal and plosive	VCC	+voi	Vow_nas_plos	Vow - Bilab Alv Pal Vel Retro _ Bilab Alv Pal Vel Retro

In general, in a word, the prediction of the stem S_i should not only be based on the knowledge of the preceding ending E_{i-1}, but also dependent on the previous stem S_{i-1}, i. e., the language model should also consider the probability $P(S_i \mid S_{i-1})$ to stem S_i. Since the stem gives the major part of the information about the word, quality of such dependency should be comparable to word bigram. Prediction of the ending is more complicated. Ending E_i should depend on the corresponding stem S_i. In addition, the Tamil language makes extensive use of agreement, for example a noun and its adjectival or pronominal attribute must agree in gender, number and case. The morphological categories often affect word-ending E_i. So the ending E_i of the word W_i should also be based on ending E_{i-1} of the preceding word W_{i-1}.

Consider the following decomposition of word W_i and W_{i-1}:
Word W_{i-1} decomposed as stem S_{i-1} and ending E_{i-1}. Word W_i decomposed as stem S_i and ending E_i
According to the morpheme based language model designed by us the prediction of stem S_i depends on S_{i-1} and also E_{i-1}:
$P(S_i)$ depends on $P(S_i \mid S_{i-1})$ and $P(S_i \mid E_{i-1})$
and prediction of the ending E_i depends on the previous ending E_{i-1} and the stem S_i:
$P(Ei)$ depends on $P(Ei \mid E_{i-1})$ and $P(Ei \mid Si)$
Since there is a strong dependency between the stem and its endings, all possible endings of the stems present in the training corpus is found using the existing Tamil morphological generator (Anandan et al., 2001). The generator generates all the possible endings for the given stem word. All stem - ending combinations are used in evaluating the probability of occurrence of the morphs in the modified morpheme based language model.
The bigram probability estimation for the modified morpheme based language is calculated for stems as follows:

$$P(S) = \alpha.P(S_i \mid S_{i-1}) + (1-\alpha).P(S_i \mid E_{i-1}) \tag{6}$$

The bigram probability estimation for the endings is calculated as:

$$P(E) = \xi.P(E_i \mid S_i) + (1-\xi).P(E_i \mid E_{i-1}) \tag{7}$$

where α and ξ are parameters in the range *0 to 1*. The knowledge of the preceding ending gives less information on the occurrence of a stem and also the occurrence of the preceding ending gives less information on the occurrence of the next word ending. So, the value of α and ξ were set to 0.9, for which improved perplexity values were obtained. The bigram probability of occurrence of a word based on the stem-end combination was calculated by:

$$P(W_i \mid W_{i-1}) = \begin{cases} P(S)P(E) & \text{if } cnt(S_i, E_i) \text{ and } cnt(S_{i-1}, E_{i-1}) > 0 \\ P(E_i \mid S_i) & \text{if } cnt(S_i, E_i) > 0 \text{ and } cnt(S_{i-1}, E_{i-1}) = 0 \\ P(S_i) & \text{otherwise} \end{cases} \tag{8}$$

The probability of occurrence of a word is based on the combined probability of occurrence of its stem and endings. If it is not possible to identify the word, based on the stem-ending combinations of the words W_i and W_{i-1}, the number of occurrence of the word (W_i) with the stem (S_i) and ending (E_i) pair is estimated. If no words exist with that stem-ending combination then the number of occurrence of the stem (S_i) of the word (W_i) is estimated. The bigram probability of the occurrence of the word W_i, after smoothing is represented as:

$$P(W_i \mid W_{i-1}) = \varphi 1 P(S)P(E) + \varphi 2 P(E_i \mid S_i) + \varphi 3 P(S_i) \tag{9}$$

where $\phi 1 + \phi 2 + \phi 3 = 1$.
The results were analyzed for different values of $\phi 1$, $\phi 2$ and $\phi 3$ and it was found that for values of $\phi 1 = 0.5$, $\phi 2 = 0.3$ and $\phi 3 = 0.2$ improved perplexity and WER results were produced. The modified morpheme based language model was trained on newspaper text corpus with 5 lakhs words and tested on the speech corpus collected from 200 speakers. The results obtained for the modified morpheme based language model for newspaper corpus is shown in Table 6. The morpheme based Trigram language model showed an improvement in the perplexity and WER values.

6. Results

Tamil text of about 200 sentences was read aloud by 100 speakers (60 male and 40 female speakers) and the signals were segmented using our segmentation algorithm. The segmented phoneme signals were mapped to their phonemes represented in text format in IPA form. Error distribution at different positions in the word due to addition, deletion and replacement of phonemes were analyzed. The distribution of the phoneme recognition errors is 11.76% due to deletion, 48.24% due to addition and 40% due to replacement. When compared to the consonants, the distinct characteristics of the vowels were accurately evaluated using the

acoustic features.

Table 6. Results of modified morpheme based Bigram and Trigram models using Katz backoff smoothing technique.

Corpus	Morpheme Based Bigram		Morpheme Based Trigram	
	Perplexity	WER	Perplexity	WER
News	125.36	13.50	113.20	12.90

This resulted in reduction of errors due to vowel insertion and replacement, and no errors to occur due to vowel deletion. The use of category-based phoneme language models at the segmentation points detected and corrected the errors that occurred at the beginning and ending of the words. However, the errors that occur in the middle of the words were not detected during the segmentation phase. Phoneme recognition rate of 74.11% was obtained by the use of language models at the segmentation phase. In order to improve the performance of the phoneme recognition rate, variable length language model integrated with inter-word hybrid language models were applied after the recognition phase. This language model for phoneme error correction resulted in phoneme recognition rate of 84.11%. The use of syllable/morpheme based language models for error correction resulted in recognition rates of 87.1%. Comparison of the percentage of errors of various types corrected at different phases of speech recognition using language models is shown in Figure 3. Percentage of phoneme classification errors in three word positions and percentage of errors corrected by using language models in the segmentation and recognition phase is listed in Table 7. The use of variable length language model integrated with hybrid language model essentially removed the errors introduced due to phoneme deletion and insertion and the use of syllable and morpheme based language models for error correction removed the errors due to phoneme replacement.

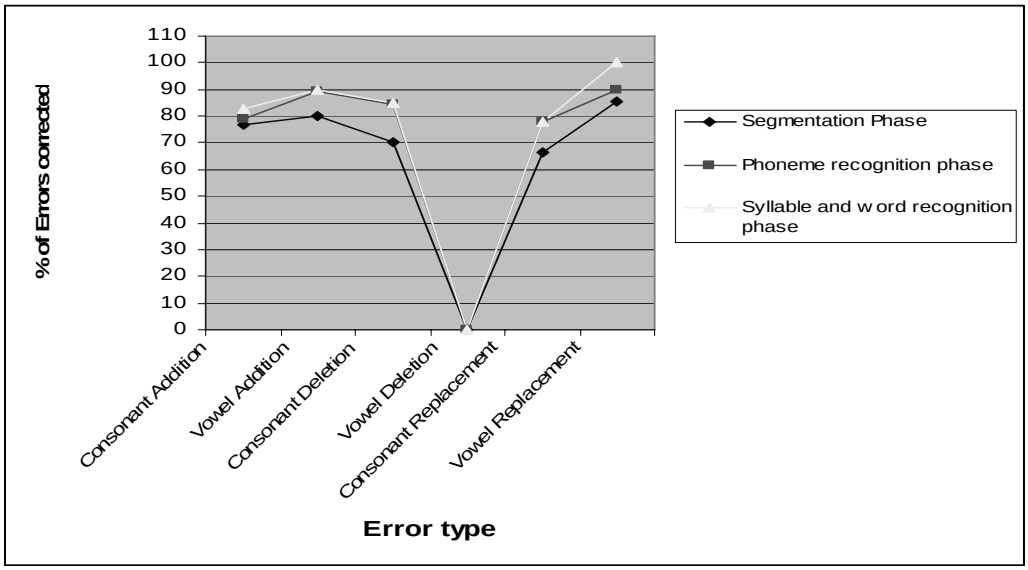

Figure 3. Comparison of different types of errors corrected during the various phases of speech recognition

Table 7. Percentage of phoneme classification errors in three word positions and Percentage of errors corrected by using language models in the segmentation and recognition phase

Error type	Word position			Error corrected by using language model at		
	B	M	E	Segmentation phase	Recognition phase	
					Phoneme LM	Syllable and morpheme LMS
Consonant Addition	23.10	63.50	13.40	76.90	79.00	82.69
Vowel Addition	40.00	33.40	26.60	80.00	89.40	90.00
Consonant Deletion	10.00	45.00	45.00	70.00	84.50	85.00
Vowel Deletion	-	-	-	-	-	-
Consonant Replacement	24.10	70.40	5.50	66.66	77.65	77.77
Vowel Replacement	50.00	35.70	14.30	85.71	90.00	100.00

7. Conclusion

In this paper, language models are used in different phases of speech recognition such as phoneme segmentation, recognition and error correction. Different types of language models such as category-based phoneme language models, inter and intra word language models, syllable and morpheme based language models were used to improve the performance of Tamil speech recognition system. Recognition rate of 87.1 % was obtained by applying language models at various phases of recognition. The speech recognition system is designed for Tamil speech corpus transcribed at the word level. The system can be further enhanced, by modifying it for continuous speech recognition. Design of language models at the sub-word level to identify wrongly segmented words can further improve the performance of continuous speech recognition for the Tamil language. A combination of the morpheme based language model and the syntactic and semantic class based models can further improve the error correction rate in the Tamil speech recognition system, due to the inflectional characteristics of Tamil.

References

Ali A.M.A. 1997, Acoustic features for automatic recognition of fricatives, *Technical Report, TR-CST27AUG97, Center for Sensor Technologies, University of Pennsylvania*.

Ali A.M.A. 1998, Segmentation and categorization of phonemes in continuous speech, *Technical Report TR-CST25JUL98, Center for Sensor Technology, University of Pennsylvania.*

Ali A.M.A., Vander S.J., Mueller P., Haentjens G. and Berman J. 1999, An acoustic-phonetic feature based system for automatic phoneme recognition in continuous speech, *IEEE Proceedings of International Symposium on Circuits and Systems (ISCAS),* Vol. 3, pp. 118-121.

Ali A.M.A. 2001, Acoustic–phonetic features for the automatic classification of stop consonants, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 8, pp. 833-841.

Anandan P., Geetha T.V. and Ranjani Parathasarathy 2001, Morphological generator for Tamil, *Proceedings of Tamil Inayam Conference, Malaysia,* pp. 46-54.

Arden A. H. Rev and Clayton A.C.1969, A progressive grammar of the Tamil language, *Christian Literature Society, Madras.*

Axelrod S. and Maison B. 2004, Combination of Hidden Markov Models with Dynamic Time Warping for Speech Recognition, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*, Vol. 1, pp. 173-178.

Bazzi I. 2000, Modeling out-of-vocabulary words for robust speech recognition, *Ph.D thesis, Massachusetts Institute of Technology*.

Bilmes J. 2002., What HMMs can do?, *UWEE Technical Report, Number UWEETR 2002-0003, University of Washington, Seattle, Washington.*

Chang S. 2002, A Syllable, Articulatory-feature, and stress accent model of speech recognition, *PhD thesis, University of California, Berkeley, California.*

Charniak E., Knight K. and Yamada K. 2003, Syntax-based language models for machine translation, *Proceedings of MT Summit IX*, pp. 40-46.

Eberman, B., and Goldenthal, W. 1996, Time-based clustering for phonetic segmentation, *Proceedings of ICSLP '96,* pp. 1225--1228,

Galescu L. and Ringger E. 1999, Augmenting words with linguistic information for N-gram language models, *Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech-99)***,** pp. 2171-2174.

Gotoh Y. and Renals S. 2000, Topic-based mixture language modeling, *Journal of Natural Language Engineering***,** Vol. 5, pp. 355-375.

Greenberg S. 2003, Pronunciation variation is key to understanding spoken language, *Proceedings of International Conference on Phonetic Science, Barcelona*, pp. 219-222.

Guido Aversano, Anna Esposito, Antonietta Esposito and Maria Marinaro 2001, A new Text independent method for phoneme segmentation, *Proceedings of the IEEE International Workshop on Circuits and Systems,* Vol. 2, pp. 516-519.

Hedge, Rajesh Mahanand, Hema A. Murthy and Gadde Venkata Ramana Rao. 2004, Continuous speech recognition using joint features derived from the modified group delay function and MFCC, *Proceedings of InterSpeech-2004*, pp. 905-908.

Kothandaraman Pon. 1997 , A grammar of contemporary literary Tamil, *International Institute of Tamil Studies, Chennai*.

Lahiri, A. 1999. Speech recognition with phonological features, *Proceedings of ICPhS 99*, pp. 715-718.

Marcus E. Hennecke, Venkatesh Prasad K. and David G. Stork 1994. Using deformable templates to infer visual speech dynamics, *28th Annual Asilomar Conference on Signals, Systems, and Computers*, vol 1, pp. 578-582, Pacific Grove,CA, IEEE Computer Society Press.

Metze F. and Waibel A. 2002, A flexible stream architecture for ASR using articulatory features, *Proceedings of 7th International Conference on spoken language processing (ICSLP 2002),* pp. 2133-2136

Pellom B.L. and Hansen J.H.L. 1998, Automatic segmentation of speech recorded in unknown noisy channel characteristics, *Speech Communication,* Vol. 25, pp. 97-116.

Rabiner L.R. 1989, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of IEEE,* Vol. 77, No. 2, pp. 257-286.

Reetz, H. 1999. Converting speech signals to phonological features, *Proceedings of ICPhS 99,* pp.1733–1736.

Saraswathi S. and Geetha, T.V. 2004, Building language models for Tamil speech recognition system, *Proceedings of Asian Applied Computing Conference (AACC-2004) and Published in Lecture Notes in Computer Science, Springer-Verlag,* Vol. 3285, pp. 161-168.

Saraswathi S., Geetha T.V. and Saravanan K. 2006a, Integrating language independent segmentation and language dependent phoneme based modeling for Tamil speech recognition, *Asian Journal on Information Technology,* Vol. 5, No. 1, pp. 38-43.

Saraswathi S., Rajeswari Sridhar and Geetha T.V. 2006b, Tamil phoneme segmentation by combining spectral and temporal features, *Proceedings of Frontiers of Research on Speech and Music (FRSM- 2006)*, pp. 54-57.

Saraswathi. S and Geetha T.V. 2007, Morpheme based language model for Tamil speech recognition system, *International Arab Journal of Information Technology,* Vol.4. , No.3, pp.214-219.

Seide F., Peng Yu, Chengyuan Ma and Chang E. 2004, Vocabulary-independent search in spontaneous speech, *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'04),* pp. 253-258.

Shikano.K, 1985. Evaluation of LPC spectral matching measures for phonetic unit recognition, *Tech. Rep., Carnegie-Mellon University*.

Stuker S., Metze F., Schultz T. and Waibel A. 2003, Integrating multilingual articulatory features into speech recognition, *Proceedings of the 8th European Conference on Speech Communication and Technology* (*EuroSpeech-2003)*, pp. 1033-1036.

Waibel A., Hanazawa T., Hinton G., Shikano K. and Lang K.J., 1989. Phoneme recognition using time-delay neural networks, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 31. No. 3. pp. 328-339.

Young S. 1996, A review of large vocabulary continuous speech recognition, *IEEE Signal Processing Magazine,* pp. 45-47.

Biographical notes

Dr.S.Saraswathi is an Assistant Professor, in the Department of Information Technology, Pondicherry Engineering College, Pondicherry, India. She completed her PhD, in the area of speech recognition for Tamil language at Anna University, Chennai, India.. Her areas of interest include speech processing, artificial intelligence and expert systems. Currently, she supervises seven PhD students in the areas of speech processing, information extraction, Natural language processing and Intelligent systems.

Dr. T.V. Geetha is a Professor in the Department of Computer Science and Engineering, Anna University, India. She has twenty years of teaching experience and has supervised six PhD students so far. She is interested in the area of Tamil computing and has done projects for Ministry of Information Technology, Government of India that includes development of Tamil corpora, Tamil office suite, speech engine, Tamil search engine and parser for Tamil. Currently, she supervises seven PhD students in the areas of speech processing, information extraction, visualization and game theory. Her research interests include artificial intelligence, speech processing, intelligent systems, and compiler design.

Received August 2010
Accepted August 2010
Final acceptance in revised form August 2010