International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 3 Issue 1 January 2012

CONTINUOUS SPEECH RECOGNITION SYSTEM FOR MALAYALAM LANGUAGE USING PLP CEPSTRAL COEFFICIENT

Cini Kurian and Kannan Balakriahnan

Department of Computer Applications. Cochin University Science and Technology, Cochin , Kerala , India

ABSTRACT

Development of Malayalam speech recognition system is in its infancy stage; although many works have been done in other Indian languages. In this paper we present the first work on speaker independent Malayalam continuous speech recognizer based on PLP (Perceptual Linear Predictive) Cepstral Coefficient. The performance of the developed system has been evaluated with different number of states of HMM (Hidden Markov Model), Different number of Gaussian mixtures, and tied states. We have also evaluated the performance of the system with bigram and trigram language models. Moreover this paper compares the recognition accuracy of context independent and context dependent tied state models. The system employs Hidden Markov Model (HMM) for pattern recognition. The system is trained with 21 male and female speakers in the age group ranging from 19 to 41 years. The system obtained a word recognition accuracy of 89% and a sentence recognition accuracy of 83%, when tested with continuous speech data from unseen speakers

KEYWORDS

Speech Recognition, Malayalam, PLP , HMM

1. INTRODUCTION

Malayalam is one among the 22 languages spoken in India with about 38 million speakers. Malayalam belongs to the Dravidian family of languages and is one of the four major languages of this family with a rich literary tradition. The majority of Malayalam speakers live in the Kerala,

one of the southern states of India and in the union territory of Lakshadweep. There are 37 consonants and 16 vowels in the language. It is a syllable based language and written with syllabic alphabet in which all consonants have an inherent vowel /a/. There are different spoken forms in Malayalam although the literary dialect throughout Kerala is almost uniform.[5]

Humans interact with environment in several ways: sight, audio, smell and touch. Humans send out signals or information visually, auditory or through gestures [26]. Because of the increased volume data, human has to depend on machines to get the data processed. Human – computer interaction generally use keyboard and pointing devices. In fact, speech has the potential to be a better interface other than keyboard and pointing devices [16].

Keyboard, a popular medium requires a certain amount of skill for effective usage. Use of mouse also requires good hand-eye coordination. Physically challenged people find it difficult to use computer. It is difficult for partially blind people to read from monitor. Moreover current computer interface assumes a certain level of literacy from the user. It expects the user to have certain level of proficiency in English apart from typing skill. Speech interface helps to resolve these issues. Speech synthesis and speech recognition together form a speech interface. Speech synthesizer converts text into speech. Speech recognizer accepts spoken words in an audio format and converts into text format [19].

Speech interface supports many valuable applications - for example, telephone directory assistance, spoken database querying for novice users, “hands busy“ applications in medical line, office dictation devices, automatic voice translation into foreign languages etc. Speech enabled applications in public areas such as railways; airport and tourist information centers might serve customers with answers to their spoken query. Physically handicapped or elderly people might able to access services easily, since keyboard is not required. In Indian scenario, where there are about 1670 dialects of spoken form, it has greater potential. It could be a vital step in bridging the digital divide between non English speaking Indian masses and others. Since there is no standard input in Indian language, it eliminates the key board mapping of different fonts of Indian languages [7].

ASR is a branch of Artificial Intelligence (AI) and is related with number of fields of knowledge such as acoustics, linguistics, pattern recognition etc [2]. Speech is the most complex signal to deal with since several transformations occurring at semantic, linguistic, acoustic and articulator levels. In addition to the inherent physiological complexity of the human vocal tract, physical production system also varies from one person to another [5, 6]. The utterance of a word found to be different, even when it is produced by the same speaker at different occasions. Apart from the vast inherent difference across different speakers and different dialects, the speech signal is influenced by the transducers used to capture the signal, channels used to transmit the signal and even the environment too can change the signals. The speech also changes with age, sex, and socio economic conditions, the context and the speaking rate. Hence the task of speech recognition is not easy due to many of the above constraints during recognition [17].

Speech recognition systems perform two fundamental operations: Signal modeling and pattern matching. Signal modelling represents process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter sets from memory which closely matches the parameter set obtained from the input speech signal. A number of methods exist for encoding speech signals, such as Linear Prediction coding (LPC) [10], and Mel-Frequency Cepstrum Coefficients (MFCC) [15] and Perceptual Linear Predictive (PLP) coefficient [14]. PLP technique is more adapted to human hearing, since it uses psycho-acoustic concepts to estimate the auditory spectrum.

In most of the current speech recognition systems, the acoustic component of the recognizer is exclusively based on HMM [10, 12, 14]. The temporal evolution of speech is modelled by the Markov process in which each state is connected by transitions, arranged into a strict hierarchy of phones, words and sentences.

Artificial neural networks (ANN) [3, 18] and support Vector machines (SVM) [2, 10] are other techniques which are being applied to speech recognition problems. In ANN, temporal variation of speech can not be properly represented. SVM, being a binary static classifier, adaptation of

the variability of the duration of speech utterance is very complex and confusing. Being a binary classifier, SVM cannot be applied directly to ASR, as ASR faces multi-class issues.

Many research studies have been taken place in various Indian languages during the recent years [20, 21, 23]. However, Malayalam speech recognition is still in its beginning stage and very less work has been reported. A wavelet based word recognizer [18], a number recognition systems [6], and a digit recognizer [5] based on SVM are the reported works in Malayalam.

This paper describes the process of development and the evaluation of a medium vocabulary, speaker independent Automatic Mayalam Speech Recognition (AMSR) system. In Section 2, we describe the methodologies used for this work. In Section 3 a detailed overview of the system development is given, while section 4, introduces training and testing methods used. Finally, Sect. 5 provides a detailed evaluation of the developed AMSR system.

2. METHODOLOGIES USED

Speech recognition systems perform two fundamental operations: Signal modelling and pattern matching. Signal modelling represents process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter sets from memory which closely matches the parameter set obtained from the input speech signal. Hence the two important methodologies used in this work is PLP Cepstral Coefficient for signal modelling and Hidden Markov model for pattern matching. Section 2.1 highlights the fundamental concepts of PLP Cepstral Coefficnet and in section 2.2 we introduce the theoretical frame work as to how HMM can be applied in speech recognition problems.

2.1. PLP Cepstral Coefficient

The prime concern while designing speech recognition system is how to parameterise the speech signal before its recognition is attempted. An ideal parametric representation should be perceptually meaningful, robustness and capable of capturing change of the spectrum with time.

The Perceptual Linear Prediction (PLP) method proposed by Wheatley and Picone [27],

converts speech signal in a meaningful perceptual way. It takes advantages of the principal characteristics derived from the psychoacoustic properties of the human hearing. viz; Critical band analysis, Equal loudness pre-emphasis and Intensity loudness conversion. In contrast to pure linear predictive analysis of speech, perceptual linear prediction (PLP) modifies the short-term spectrum of the speech by several psychophysically based transformations. The different stages of PLP extraction is shown in Fig 1.

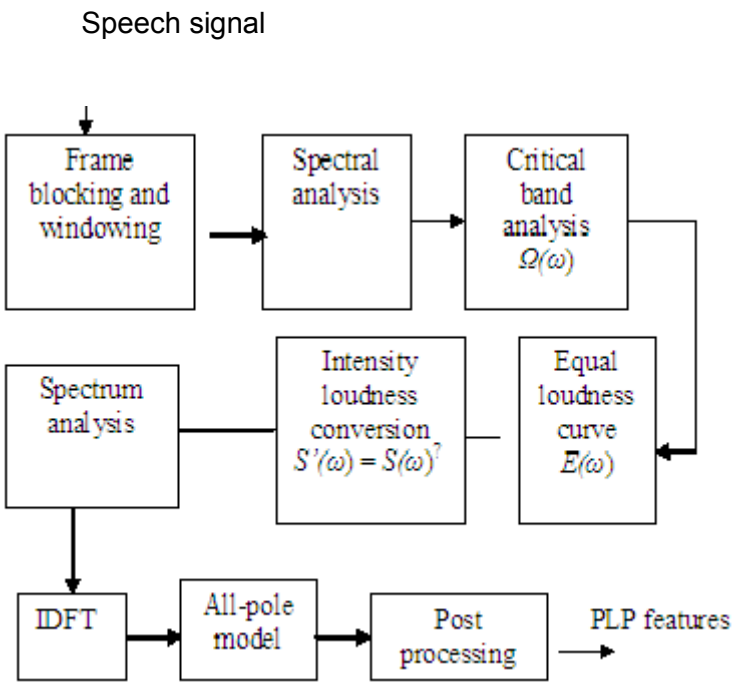


Fig.1. Block diagram of PLP extraction

The primary step in any feature extraction process is blocking the frame. Here audio signals which are basically non stationary are cut into fragments are called frames. Then frames are passed through Hamming Window. During spectral analysis, signal is passed though Fourier Transform process and then power spectrum of the signal is computed. The various steps of PLP feature extraction used for this work are depicted below.

2.1.1 Critical band integration (Bark frequency weighing)

Experiments in human perception have shown that frequencies of a complex sound within a certain bandwidth (critical bandwidth) of 10% to 20% frequency cannot be individually identified. If any one of the components of this sound falls outside this band width, it cannot be individually distinguished. Hence a mapping is done from acoustic frequency to a 'perceptual frequency' called bark frequency scale, represented as equation (1)

$$Bark=13atan(0.76f/1000) + 3.5atan(f^2/7500^2) \quad (1)$$

Thus the speech signal is passed through some trapezoidal filters equally spaced in bark scale to produce a critical band spectrum approximation.

2.1.2 Equal loudness pre-emphasis

At conventional speech levels, human hearing is more sensitive to the middle frequency range of the audible spectrum. PLP incorporates the effect of this phenomenon by multiplying the critical band spectrum by an equal loudness curve that suppresses both the low and high frequency regions relative to the midrange from 400 to 1200 Hz. In short different frequency components of speech spectrum are pre-emphasized by an equal -loudness curve, which is an approximation to the unequal sensitivity of human hearing at different frequencies, closer to 40dB level.

2.1.3 Intensity loudness conversion (cube-root amplitude compression)

Cube-root compression of the modified speech spectrum is carried out according to the power law of hearing [25], which simulates the non-linear relation between the intensity of sound and its perceived loudness. Together with the psychophysical equal-loudness pre-emphasis, cube-root amplitude compression operation reduces spectral amplitude variation of critical-band spectrum

2.2. Hidden Markov Model and Statistical Speech Recognition

Hidden Markov Models are widely used for automatic speech recognition and inherently incorporate the sequential and statistical character of the speech signal. Speech recognition system treats the recognition process as one of the maximum a-posteriori estimation, where the most likely sequence of words is estimated, given the sequence of feature vectors for the speech signal. The speech signal to be recognized is converted by a front-end signal processor into a sequence of acoustic vectors, *Y = y1,y2,y3--.* Assuming that the utterance consists of sequence of words *W = w1, w2, w3 ----wn,*the problem here is to determine the most probable word sequence, W, given the sequence of feature vectors for the speech signal. Applying Bays' rule [15] to decompose the required probability

$S = \arg w\max P(Y/O) = \arg w\max (P(Y/W)P(W) / P(Y)$

$S = \arg w\max P(Y/W)P(W)$

(2)

Posterior prior

The right hand side of equation (2) has two components: i) the probability of the utterance of the word sequence given the acoustic model of the word sequence and ii), and the probability of sequence of words. The first component P*(O/W*), known as the observation likelihood, which is computed by the acoustic model. The Second component P *(W)* is estimated using the language model. The acoustic modeling of this Speech Recognition system is done using HMM. The figure 2 illustrates these concepts. The topology of a basic HMM with five states is shown in figure 3.

Each transition in the state diagram of a HMM has an associated transition probability [17, 24]. These transition probabilities are denoted by matrix A. Here A is defined as $A = a_{ij}$ where $aij = P(t_{t+1} = j \mid j = i)$, the probability of being in state j at time $t+1$, given that we were in state *i* at time *t*. It is assumed that a_{ij}'s are independent of time. Each state is associated with a set of discrete symbols with an observation probability assigned to each symbol, or is associated with the set of continuous observation with a continuous observation probability density. These observation symbol probabilities are denoted by the parameter B. Here B is defined as $B = b_j(k)$, where $b_j(k) = P(v_k \text{ at } t \mid i_t = j)$, the probability of observing the symbol v_k , given that it is in the

state *j*. The initial state probability is denoted by the matrix π, where π is , defined as $\pi = \pi_i$ where $\pi_i = P(i_t = 1)$, the probability of being in state *t* at *t = 1.* Using the three parameters *A, B* , and π a HMM can be compactly denoted as $\lambda = \{ A, B, \pi \}$

Hence the three fundamental ASR problems can be well addressed with HMM. They are (i) Scoring and evaluation i.e computing the likelihood of an observation sequence, given a particular HMM. This problem occurs during the recognition phase. Here for the given parameter vector sequence (observation sequence), derived from the test speech utterance, the likelihood value of each HMM is computed using forward algorithm. The symbol associated with the HMM, for which the likelihood is maximum, is identified as the recognized symbol corresponding to the input speech utterance. Problem (ii) is associated with training of the HMM for the given speech unit. Several examples of the same speech segments with different phonetic contexts are taken, and the parameters of HMM, λ, have been interactively refined for maximum likelihood estimation, using the Baum-Wetch algorithm [13]. Problem (iii) is associated with decoding or hidden state determination, where the best HMM state is decided.

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 3 Issue 1 January 2012

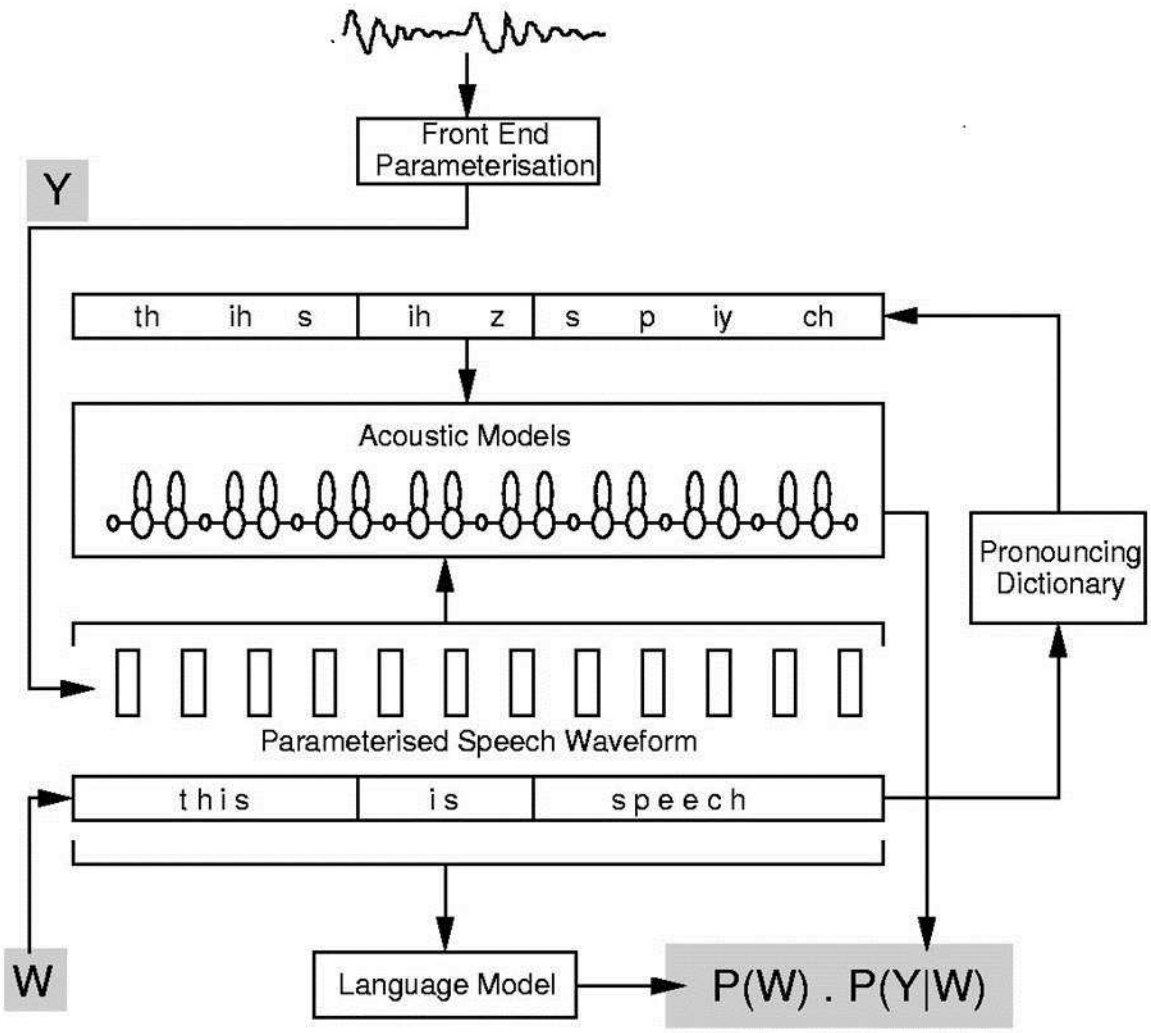
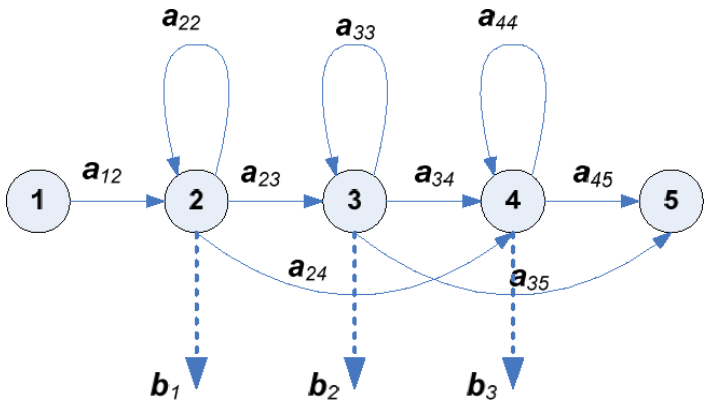


Fig. 2: Acoustic and Language Models for a speech recognition system

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 3 Issue 1 January 2012



Fig, 3. Topology of a 5 state HMM

3. SYSTEM DEVELOPMENT

The system development architecture is detailed in figure 4. As illustrated in this figure, the developed AMSR (Malayalam Speech recognition) system consists of two important phases. They are Training phase and Recognition phase. The training phase can be further subdivided into four sub modules. i.e data collection and data preparation, feature extraction, acoustic modelling and language modeling

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 3 Issue 1 January 2012

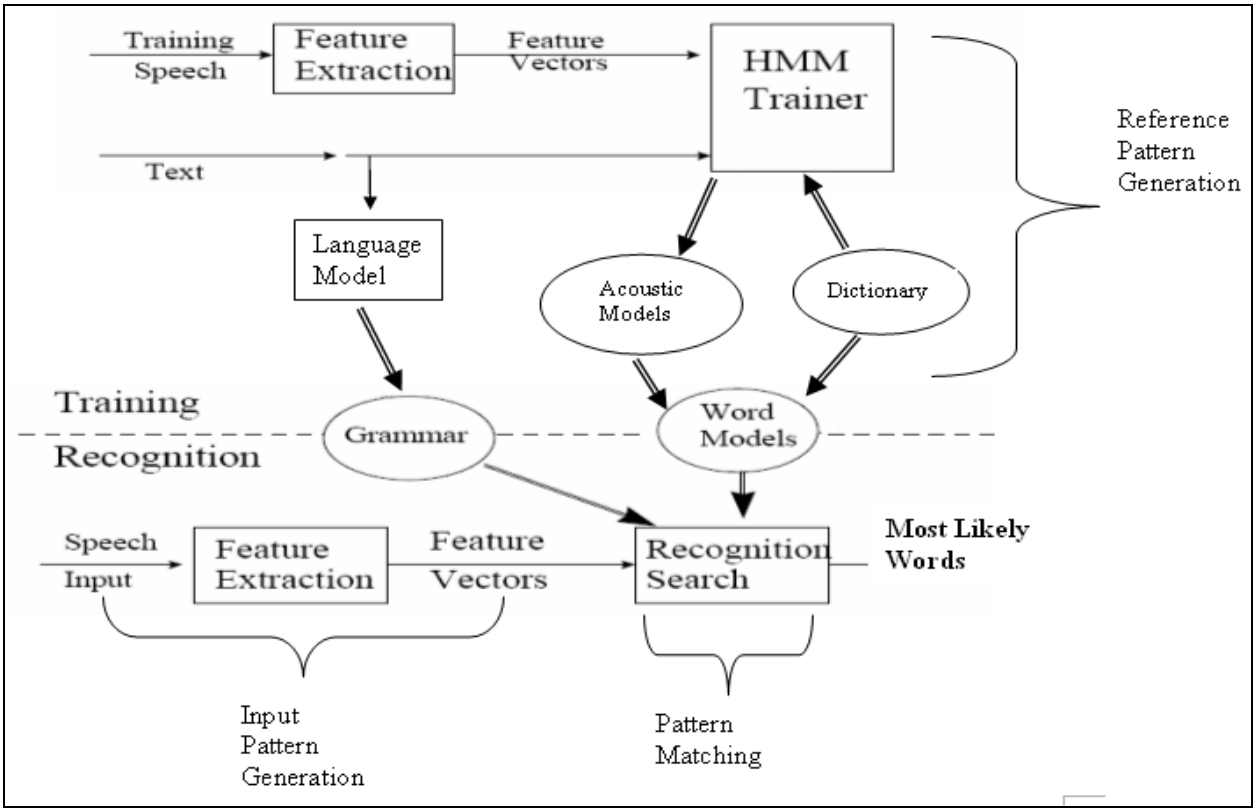


Fig. 4: Architecture view of the AMSR system

3.1 Data collection

We have collected two types of data for the MSR system i.e. text corpus and speech corpus. Corpus creation especially collection speech corpus is laborious and time consuming process. If such resources are readily available, the system development would have been easier. Unfortunately, no such error free resources are available for Malayalam language. We collected text materials from on-line Malayalam newspapers. The corpus balancing tool, CorpusCrt [22] is used to extract a set of phonetically rich sentences, from the text materials. Accordingly, 20 phonetically rich sentences are selected for training.

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 3 Issue 1 January 2012

Speech data for MSR system is collected from 10 male and 11 female speakers between the age group of 19-41. Recording is done in normal office environment. (The signal to noise (SNR) ratio range from 10 to 40 dB) using head set, having microphone with 70Hz to 1600Hz of frequency range and with 16 kHz sampling frequency quantized by 16 bit. The speech signal is saved in Microsoft wave format

3.2 Data Preparation

Four types of files had to be prepared for the training phase. They are; phoneme list, phonetic dictionary and transcription file. A well defined and accurate phonetic dictionary contributes a lot to the accuracy of the recognizing system. Defining the phonemes of the language and creating phonetic dictionary is another trivial task. Accordingly all the phonemes of the Malayalam language is identified and defined. The Table 1 shows the phoneme list used for this work. Unique phonemes of Malayalam language like ഴ(zha)(approximant retroflex), റ്റ **(tt'),** ൾ**(l_),** ൺ**(n_)** etc are identified and defined. Table 2 shows the IPA chart of the Malayalam alphabets.

Malayalam	Phonetic Notation	**Malayalam**	Phonetic Notation	**Malayalam**	Phonetic Notation
അ	a	ച്	clch ch	മ്	m
ആ	aa	ഛ്	clch chch	യ്	y
ഇ	i	ജ്	vbj j	ര്	r
ഈ	ii	ഝ്	clch chch	ല്	l
ഉ	u	ഞ്	nj'	വ്	v
ഊ	uu	ട്	clt' t'	ശ്	sh
ഏ	e	ഠ്	clt' t'h	ഷ്	s'h

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 3 Issue 1 January 2012

ഏ	e'	ഡ്	vbd' d'	സ്	s
ഐ	ai	ഢ്	clt' t'h	സ്	s1
ഒ	o	ണ്	n'	ഹ്	h
ഓ	o'	ത്	clt t	ള്	l'
ഔ	au	ഥ്	clt th	റ്	r'
ം	m	ദ്	vbd d	ഴ്	z
ഃ	-	ധ്	clt th	റ്റ്	clr1 r1
ു	u'	ന്	n (Dental nasal)	ഫ്	ph1 (fan)
ക	clk k	ന്	n1(alveolar nasal	ൾ	l'
ഖ്	clk kh	പ്	clp p	ൺ	n'
ഗ്	vbg g	ഫ്	clp ph	ൻ	n1
ഘ്	clk kh	ബ്	vbb b	ർ	r'
ങ്	ng'	ഭ്	clp ph	ൽ	l

Table 1: Phoneme list

Theoretically, creation of phonetic dictionary is just a mapping of grapheme to phoneme. But this alone would not work especially for a language like Malayalam because many phonemes pronounced differently at different contexts. For example, ഫ (ph'a) pronounced differently in ഫലം and ഫാൻ . ന (na) (Nasal dental and Nasal alveolar) is pronounced differently even though the grapheme notation is same (eg. നനക്കുക(nan'aykkuka). Hence for creating pronunciation dictionary, initially we have done mapping of all grapheme into the corresponding

phoneme units. Then some phonological rules have been applied manually and edited the dictionary. Thus, we have created a phonetic dictionary for the 102 words of the vocabulary. Multiple pronunciations are also incorporated in the dictionary.

The transcription file is a record of what is actually recorded. Errors during the transcription will mislead the training process. Hence the transcription process is done manually considering even silence, noise or a breath. Thus 420 files were manually transcribed. The total words in the corpus of work is 2300 words, while the vocabulary is for 102 words

		Labial		Dental		Alveolar	Retroflex		Palatel	
Nasal		/m/ മ m		/n̪/ ന n		/n/ ന * n	/ɳ/ ണ ṇ		/ɲ/ ഞ ñ	
Stop	**plain**	/p/ പ p	/b/ ബ b	/t̪/ ത t	/d̪/ ദ d	/t/ * t	/ʈ/ ട ṭ	/ɖ/ ഡ ḍ	/t͡ʃ/ ച c	/d͡ʒ/ ജ j
	aspirated	/pʰ/ ഫ ph	/bʰ/ ഭ bh	/t̪ʰ/ ഥ th	/d̪ʰ/ ധ dh		/ʈʰ/ Oṭh	/ɖʰ/ ഢḍh	/t͡ʃʰ/ ഛ ch	/d͡ʒʰ/ ഝ jh
Fricative		/f/ ഫ* f		/s̪/ സ s			/ʂ/ ഷ ṣ		/ɕ/ ശ ś	
Approximant	**central**	/ʋ/ വ v					/ɻ/ ഴ l̤		/j/ യ y	
	lateral					/l/ ല l	/ɭ/ ള l̤			
Rhotic						/ɾ/ ര r	/r/ റ r̲			

Table 2 : IPA chart of Malayalam consonants

3.3 Feature Extraction:

Feature extraction module of this work is done as per the methodology detailed section 2.1 above. For feature extraction, speech is digitized at a sampling rate of 16 kHz. Then PLP Cepstral Features are extracted from the speech signal using a Hamming window size of

25msec and a window shift of 10msec. A pre-emphasis filter $H(z) = 1 - 0.97 z^{-1}$ is applied. From each frame of speech, 12 cepstral coefficients and 1 energy coefficient is obtained. The delta and acceleration coefficients are appended to the derived cepstral coefficients to obtain a 39 dimensional vector coefficient. These acoustic vectors are used for representing the voice characteristic of the speaker [15]. Therefore, each input utterance is transformed into a sequence of acoustic vectors.

3.4 Language modeling

Importance of a language model in a speech recognition system is vital as acoustic model alone cannot handle the problem of word ambiguity. Word ambiguity may occur in several forms such as similar sounding sounds and word boundaries. With similar sounding sounds, words are indistinguishable to the ear, but are different in spelling and meaning. The words "paat'am' (പാടം) and " paat'ham' (പാഠം) " are such examples. In continuous speech, word boundaries are also challenging. For instance, the word " talasthaanam' (തലസ്ഥാനം) can be misconstrued as "tala sthaanam' (തലസ്ഥാനം). The use of language model resolves these issues by considering phrases and words that are more likely to be uttered.

Language Model is used to support the Recognition process. As explained earlier, in a recognition process, initially a sequence of features from the utterance is taken and it is compared with the existing acoustic model. Then it generates the possible phones in the Utterance. Lexicon/Dictionary is used for the identification of the utterance (spoken word). After the word identification it checks with the Language Model whether the corresponding word/sentence format is a valid one or not. If the word/sentence is found in the Language Model, the recognition system identifies that it is a valid word/sentence format and accordingly the recognized result will be produced. For a given a text $W^T = W_1, W_2, \ldots\ldots W_t, \ldots\ldots W_T$, the probability is computed by $\Pr(W^T) = \Pr(W_1) \prod \Pr(W_t \mid h_t)$, where $h_t = w_1 \ldots\ldots w_{t-1}$ indicates history of the word W_t. $\Pr(W_t \mid h_t)$ is difficult to calculate as the history of h_t grows. Hence usually the history is restricted to bigram or trigram. The trigram model based on the previous two words is powerful, as most words have a strong dependence on the previous

two words and it can be estimated reasonably well with an attainable corpus. The trigram based language model with back-off is used for recognition in this work. The language model is created using the CMU statistical LM toolkit [9].

3.4 Acoustic modeling

The acoustic modeling component of the system has four important stages. The first stage is to train the context independent model and then training context dependent models. Decision trees are built on the third stage and finally context independent tied models are created. Here, continuous Hidden Markov models is chosen to represent context dependent phones (triphones). The phone likelihood is computed. using HMM. The likelihood of the word is computed from the combined likelihood of all the phonemes. The acoustic model thus built is a 3 state continuous HMM, with states clustered using decision tree.

4. TRAINING AND TESTING

For training and testing the system, the data base is divided into three equal parts- 1, 2, & 3 and the training is conducted in a round robin fashion. For each trial, 2/3rd of the data is taken for training and 1/3rd of the remaining the data is used for testing. For eg. In trial I, part 1 and part 2 of the data is given for training. Then Part 3 of the database is used for testing the trained system. In trial II, part 1 and part 3 of the data base is used for training and part II of the database is used for testing. In experiment III, part 2 and part 3 of the database is taken for training and the system is tested with part 1 of the database. The result in terms of Word Recognition Accuracy, Sentence recognition Accuracy, Number of words deleted, inserted, substituted are obtained from each experiment.

For all the performance evaluation reports detailed in the following sections, we have adopted this procedure and the result reported is the average of testing experiments of I, II and III.

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 3 Issue 1 January 2012

5. PERFORMANCNE EVALUATION OF THE SYSTEM

The performance of the speech recognition system is affected by a number of parameters. This section describes the evaluation of performance of speech recognition system with different types of parameters. Section 5.1 describes the performance evaluation metrics used for evaluation. Further in section 5.2 we introduce the baseline system and its parameter. In Section 5.3, performance is evaluated with different number of Gaussian mixture models. In section 5.4 we introduced the performance of the system with different number of tied sates and finally we compared the performance of context independent and context dependent tied models.

5.1 PERFORMANCE MATRICS

Word Error Rate (WER) is the standard evaluation metric used here for speech recognition. It is computed by SCLITE [13], a scoring and evaluating tool from National Institute of Standards and Technology (NIST). Sclite is designed to compare text output from a speech recognizer such as hypothesis text to the original text (reference text) and generate a report summarizing the performance. The comparing of the reference to the hypothesis text is called the alignment process. Then result of the alignment process is obtained in terms of WER, SER, and number of word deletions, insertions and substitutions. If N is the number of words in the correct transcript; S, the number of substitutions; and D, the number of Deletions, then,

$$WER = ((S + D + I)N) /100 \quad (3)$$

Sentence Error Rate (S.E.R) = (Number of sentences with at least one word error/ total Number of sentences) * 100

5.2 EVALUATION OF BASE LINE SYSTEM

Performance of the base line system with various parameters are detailed in Table 3. The base line system uses continuous context dependent tied state HMM (3 state per model) model. The state probability distribution uses continuous density of 16 Gaussian mixture (GM) distributions. The state distributions are tied to about 1500 senons. We have used a trigram language model and a language weight of 10 .

PARAMETERS USED	SENTENCE ACCURACY %
STATES PER HMM = 3 , GM= 16 , LANGAGE MODEL = TRIGRAM , NUMBER OF SENONES = 1500	70.5

Table 3. Performance of the baseline systems and the various parameters used.

5.3 PERFORMANCE WITH DIFFERENT GAUSSIAN MIXTURES

The baseline system shows only 70.5% accuracy. Hence the system is tested with changing the number of Gaussian mixtures to 2, 4,8 and 16. A 17.6 % improvement (in comparison of base line model)obtained with a Gaussian mixture of 8 as detailed in table 4.

PARAMETERS USED	GM =2	GM = 4	GM = 8	GM =16
STATES PER HMM =3 , LANGUAGE MODEL = TRIGRAM, NUMBER OF SENONES = 1500	77.1	79.8	82.9	70.5

Table 4. Sentence Recognition Accuracy (%) with different number of Gaussian Mixture models

5.4 EFFECT OF TIED STATES (SENONS)

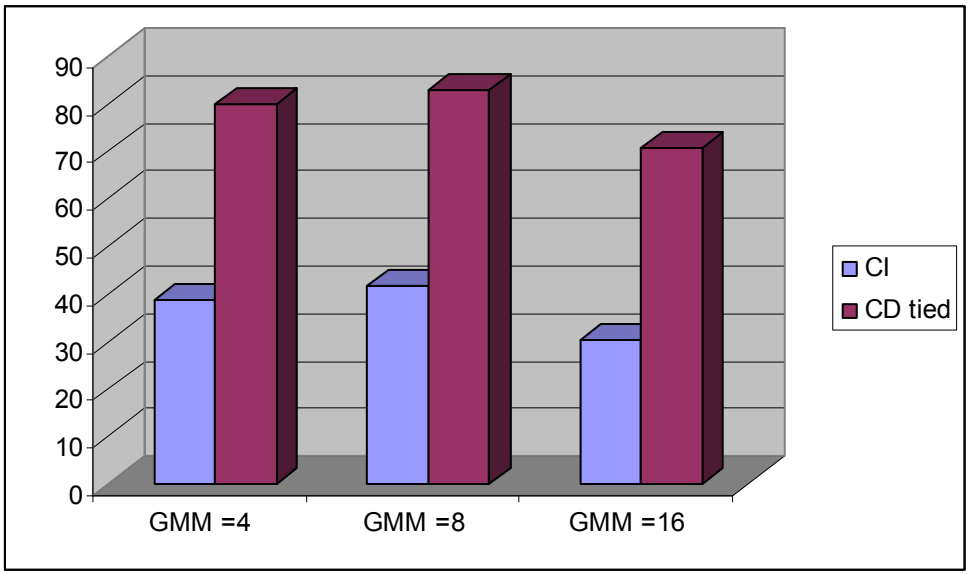
Accuracy of the system is also tested by varying the number of senons and number of Gaussian mixtures. The obtained results are detailed in table 5. The best performance is obtained for the combination of 8 Gaussian and 1500 senon . Another observation is that increasing the number of senons beyond 3000 does not have any effect on the performance.

No of senons	GMM = 2	GMM = 4	GMM = 8	GMM = 16
1200	76.9	79.5	82.6	79.5
1500	77.1	79.8	**82.9**	70.5
2000	77.9	81	82.1	57.9
3000	78.3	80.7	80.5	33.1
4000	78.3	80.7	80.5	33.1

Table 5. Recognition Accuracy (number of senon vs. number of Gaussian Mixtures)

5.5 COMPARISON OF DIFFERENT TYPES OF MODELS

We have evaluated the performance of the two important types of HMM models namely, Context independent models (CI models) and context Dependent tied models (CD tied models). For that we have created 71 independent phonemes and 2376 context dependent phonemes. Then similar phonemes are tied, and the number of phoneme is reduced to 966. Other parameters used are : 3 states per HMM ; a Gaussian mixture of 8 ; 1500 senons . Figure 5 compares the performance in terms sentence recognition accuracy. CD tied models outperform CD models by about 50%.



Fig, 5. Comparison of CI, and CD tied models

6. CONCLUSION

This paper presents the methodology of continuous speech recognition system for Malayalam language implemented using PLP Cepstral coefficients and HMM. The system is evaluated with different parameters. From the accuracy of the results it is clear that the system performed with maximum accuracy for CD-TIED models with a 3 state HMM and for a GMM of 8. It is evident from the results of the experiments that that PLP and HMM are ideal candidates for Malayalam continuous speech recognition. The system achieved 89% word recognition accuracy and 83% sentence recognition accuracy. The system can be further enhanced with vocabulary size and could be developed as a complete recognition system. To achieve the same, our future plan is to develop a tool for automated pronunciation dictionary. Other than this we propose to improve the model accuracy by utilizing more information of the linguistic knowledge such as tone, prosody, and to implement more efficient approach into the acoustic modeling process.

References

[1] A. Ganapathiraju, J. Hamaker and J.Picone, ‘ Support Vector machines for speech Recogntion,” Proceedings of the International Conferences on Spoken Language processing, pp pp.292-296,Sdney,Australia, November , 1999

[2] B. Gold, N. Morgan, Speech and audio signal processing, John Wiley and Sons, N.Y., 2002.

[3] Behrman, L. Nash, J. Steck, V. Chandrashekar, and S. Skinner, “Simulations of Quantum Neural Networks”, Information Sciences, 128(3-4): pp. 257-269, October 2000

[4] B. Gold, N. Morgan, Speech and audio signal processing, John Wiley and Sons, N.Y., 2002

[5] Cini Kurian, Kannan Balakrishnan , (2009), “Speech Recognition of Malayalam Numbers”, IEEE Transaction on Nature and Biologically Inspired computing NaBIC-2009), pp 1475-1479

[6] Cini Kurian, F. Shah, A.;Balakrishnan, K. (2010), “ Isolated Malayalam digit recognition using Support Vector Machines, IEEE Transaction on Communication Control and Computing Technologies (ICCCCT-2010), pp 692 -695

[7] Cini Kurian;Kannan Balakrishnan, K ; "Natural Language Processing in India Prospects and Challanges" Proceedings of the International Conference on “Recent Trends in Computational Science 2008“(ICRTCS-2008), Kochin, India.June 11-June 13

[8] Cini Kurian , Kannan Balakrishnan K, “Automated Transcription System for MalayalamLanguage ” International Journal of Computer Applications(*IJCA*), ISSN-0975-8887, volume 19- No.5, April 2011

[9] Clarkson, P., & Rosenfeld, R. (1997). Statistical language modeling using the CMU-Cambridge toolkit. In Proceedings of the 5th European conference on speech communication and technology, Rhodes, Greece, Sept. 1997.

[10] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” Knowledge Discovery Data Mining, vol. 2, no. 2, pp. 121–167, 1998

[11] Davis S and Mermelstein P, "Comparison of parametric representations for Monosyllabic word Recognition in continuously spoken sentences", IEEE Trans On ASSP,vol. 28, pp.357 – 366

[12] Dimov, D., and Azamanov, I.(2005). "Experimental specifics of using HMM in isolated word Speech recognition" .International Conference on Computer system and Technologies –CompSysTech „2005"

[13] Fiscus, J. (1998) Sclite Scoring Package Version 1.5, US National Institute of Standards and Technology (NIST), URL - http://www.itl.nist.gov/iaui/894.01/tools/.

[14] F.Felinek, "Statistical Methods for Speech recognition" MIT Press, cambridge Massachusetts, USA, 1997

[15] Huang, X., Alex, A., and Hon, H. W. (2001). "Spoken Language Processing; A Guide to Theory, Algorithm and System Development", Prentice Hall, Upper Saddle River, New Jersey

[16] Jurasky, D, and Martin, J.H (2007). Speech and Language Processing : An introduction to natural language Processing, Computational linguistics, and speech recognition, 2nd edition

[17] Jyoti, Singhai Rakesh,"Automatic Speaker Recognition: An Approach using DWT Based Feature Extraction and Vector Quantization", IETE Technical Review, 24, No. 5, Sept-Oct 2007, pp 395-402.

[18] Krishnan, V.R.V. Jayakumar A, Anto P B (2008) , "Speech Recognition of isolated Malayalam Words Using Wavlet features and Artificial Neural Network". DELTA2008. 4th IEEE International Symposium on Electronic Design, Test and Applications, 2008.Volume, Issue, 23-25 Jan. 2008 Page(s):240 – 243

[19] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Pearson Education 2008,

[20] M Kumar., et al "A Large Vocabulary Continuous Speech recognition system for Hindi", IBM Research and Development Journal, September 2004

[21] Saumudravijaya K, "Hindi Speech Recognition" (2001), J. Acoustic Society India, 29(1), pp 385-95

[22] Sesma Bailador , Alberto "CorpusCrt" Politechnic University of Catalonia. 1998.

[23] Singh, S. P., et al "Building Large Vocabulary Speech Recognition Systems for Indian " International Conference on Natural Language Processing, 1:245-254, 2004.

[24] Sperduti and A. Starita, "Supervised Neural Networks for classification of Structures", IEEETransactions on Neural Networks, 8(3): pp.714-735, May 1997.

[25] S.S. Stevens, "On the psychophysical law," Psychological Review, vol. 64,no. 3,pp. 153-181,1957

[26] Sukhminder Singh Grewal, Dinesh Kumar. "Isolated word Recognition System for English language" International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 447-450

[27] Wheatley, B., & Picon, J.(1991). Voice across America: "Toward robust speaker independent speech recognition for telecommunications applications". Digital signal processing: A review Journal, I(2), 45-64