Malayalam Isolated Digit Recognition using HMM and PLP cepstral coefficient

Cini Kurian, Kannan Balakrishnan

Department of Computer Applications, Cochin University of Science & Technology,
Cochin – 682 022
cinikurian@gmail.com,bkannan@cusat.ac.in

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 isolated speech recognizer based on PLP (Perceptual Linear Predictive) Cepstral Coefficient and Hidden Markov Model (HMM). The performance of the developed system has been evaluated with different number of states of HMM (Hidden Markov Model). The system is trained with 21 male and female speakers in the age group ranging from 19 to 41 years. The system obtained an accuracy of 99.5% with the unseen data.

Key words:

Speech Recognition, Malayalam, PLP

1. Introduction

Humans interact with environment in several ways: sight, audio, smell and touch. Humans send out signals or information visually, auditory or through gestures [17]. 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 [12].

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 [14].

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

DOI : 10.5121/ijait.2011.1504

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

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 [13].

In most of the current speech recognition systems, the acoustic component of the recognizer is exclusively based on HMM [7,9,10]. The temporal evolution of speech is modeled 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) [3,15] and support Vector machines (SVM) [2,7] 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. SVM is a binary classifier while ASR, faces multiclass issues.

For processing speech, the signal has to be represented in some parametric form. Wide range of methods exists for parametric representation of speech signals, such as Linear Prediction coding (LPC) [10], and Mel-Frequency Cepstrum Coefficients (MFCC) [11] and Perceptual Linear Predictive (PLP) coefficient [10]. Since PLP is more adapted to human hearing, PLP parameterization technique is used in this work.

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.

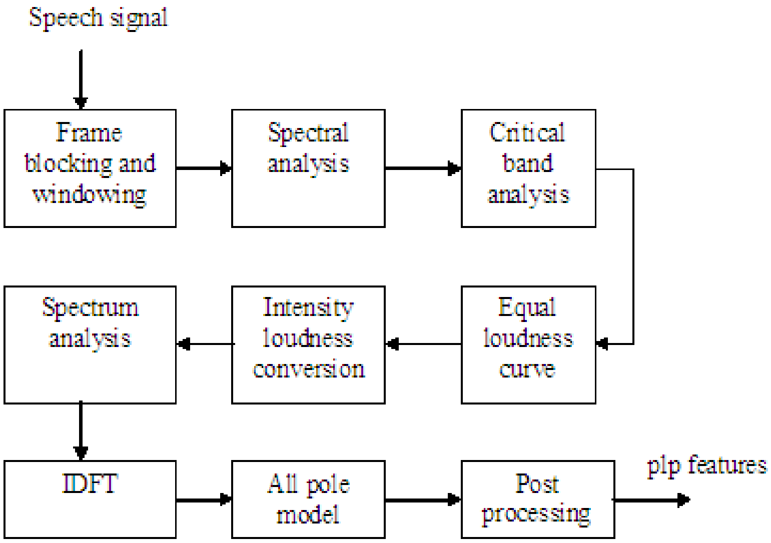


Fig.1. Block diagram of PLP extraction

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)

$$\text{Bark}=13\text{atan}(0.76f/1000) + 3.5\text{atan}(f^2/7500^2) \qquad (1)$$

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
One of first decisions in any pattern recognition system is the choice of what features to use. The PLP feature converts speech signal in meaningful perceptual way through some psychoacoustic process. The various stages of this method are based on our perceptual auditory characteristics. The different stages of PLP extraction is shown in fig 1.

2.2. Hidden Markov Model and Statistical Speech Recognition

An unknown speech wave form is converted by a front-end signal processor into a sequence of acoustic vectors, *O = o1,o2,o3,….* The utterance consists of sequence of words *W = w1, w2, w3 ----wn*. In ASR it is required to determine the most probable word sequence, W, given the observed acoustic signal *O*. Applying Bays' rule to decompose the required probability, [11]

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

$$S = \arg_{w} \max \underbrace{P(O/W)}_{Posterior} \underbrace{P(W)}_{prior}$$

Hence a speech recognizer should have two components: *P (W),* the prior probability, is computed by language model, while P *(O/W),* the observation likelihood, is computed by the acoustic model. Here the acoustic modeling of the recognition unit is done using HMM.

Since HMM is a statistical model in which it is assumed to be in a Markov process with unknown parameters, the challenge is to find all the appropriate hidden parameters from the observable states. Hence it can be considered as the simplest dynamic Bayesian network [7,9]. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. However, in a hidden Markov model, the state is not directly visible (so-called *hidden*), but the variables influenced by the states are visible. Each transition in the state diagram of a HMM has transition probability associated with it [13, 15]. 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 $b_j(k) = P(v_k \ at\ t \mid i_t = j)$, 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 $\lambda = \{A, B, \pi\}$ An Example of an HMM with five states is shown in Figure 2

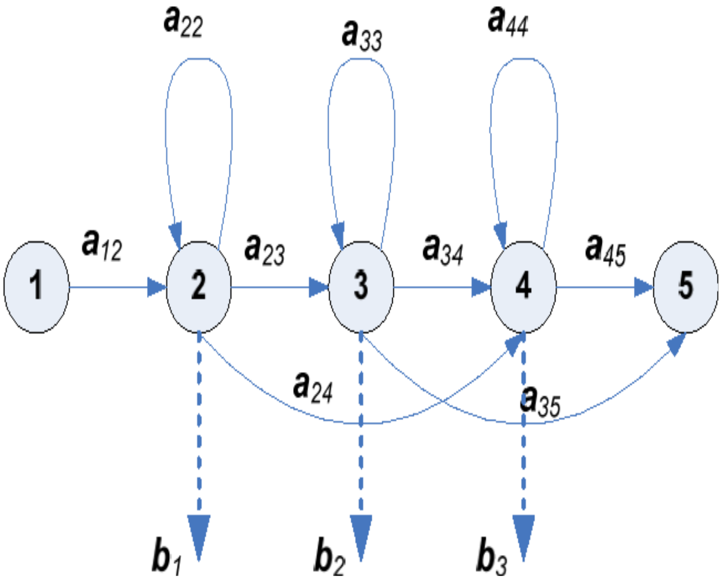
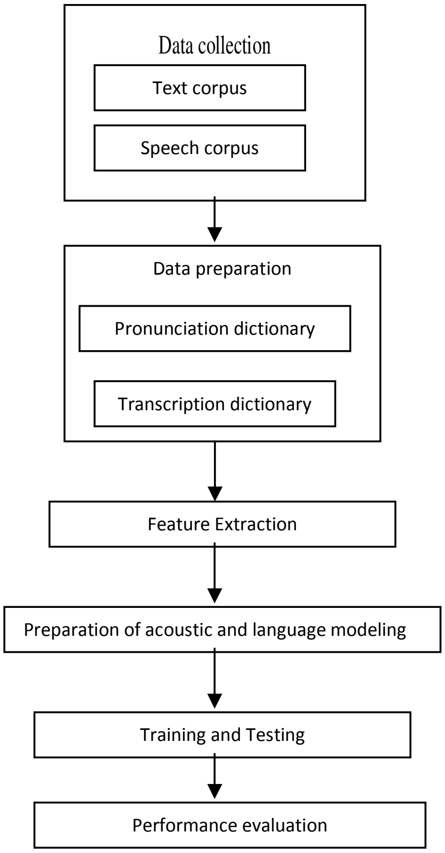


Fig. 2 Topology of a 5 state HMM

There are three fundamental ASR problems that can be addressed with HMM. Problem (i) is 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 the HMMs, 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.

3. SYSTEM DEVELOPMENT AND SPEECH DATA BASE

The different steps [5] involved in the development of the proposed system is shown in figure 3



.

Fig. 3. System Development

The database contains 210 isolated spoken words recorded from 21 speakers. Each speaker uttered numbers form zero to nine separately. The recording is done in normal environment with a high quality microphone at 16kHz sampling frequency and quantized at 16 bit . The data is quantized (at 16 kHz) and digitalized (with 16 bit) and the speech data is stored in WAV format

4. TESTING AND TRAIING

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

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 number of states of HMM.

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,

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

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 with 5 state per model. The state probability distribution uses continuous density of 16 Gaussian mixture (GM) distributions.

PARAMETERS USED	WORD RECOGNITION ACCURACY %

States per HMM = 5
GMM = 16
Senones = 1500
	82.5

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

5.3 PERFORMANCE WITH 3 STATES PER HMM

The baseline system shows only 82.5% accuracy. Hence the system is tested with changing the number of states of HMM . An average of 99.5% accuracy is obtained . Hence a 17% improvement in terms of recognition accuracy is obtained as detailed in table 4

	TRAINING				TESTING			
Experiment	Word Accuracy %	Number of Deletions	Number of Substitutions	Number of Insertions	Word Accuracy %	Number of Deletions	Number of Substitutions	Number of Insertions
1	99.29	0	1	0	100	0	0	0
2	99.29	0	1	0	98.57	0	1	0
3	100	0	0	0	100	0	0	0

Table 4: Performance Evaluation of the System with Training and Testing Data

6. CONCLUSION

This paper presents the methodology of digit 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 a HMM state of 3. It is evident from the results of the experiments that PLP and HMM are ideal candidates for Malayalam isolated speech recognition. We suggest that the model accuracy can be further improved by utilizing more information of the linguistic knowledge such as tone and prosody.

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.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, .;Firoz 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] 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

[8] 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

[9] 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‟.

[10] F.Felinek, “Statistical Methods for Speech recognition” MIT Press, cambridge Massachusetts, USA, 1997

[11] 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

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

[13] 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.

[14] Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of Speech Recognition”, Pearson Education, 2008.

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

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

[17] 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

[18] http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.html