Speech Recognition of Malayalam Numbers

Cini Kurian, Kannan Balakrishnan
Department of Computer Applications
Cochin University of Science & Technology, Cochin
Kerala, India
cinikurian@gmail.com, bkannan@cusat.ac.in

Abstract - Digit speech recognition is important in many applications such as automatic data entry, PIN entry, voice dialing telephone, automated banking system, etc. This paper presents speaker independent speech recognition system for Malayalam digits. The system employs Mel frequency cepstrum coefficient (MFCC) as feature for signal processing and Hidden Markov model (HMM) for recognition. The system is trained with 21 male and female voices in the age group of 20 to 40 years and there was 98.5% word recognition accuracy (94.8% sentence recognition accuracy) on a test set of continuous digit recognition task.

I. INTRODUCTION

Speech recognition is the most promising technology of the future. Speech has the potential to be a better interface other than keyboard and pointing devices [10]. A speech interface would support many valuable applications, i.e., telephone directory assistance, spoken database querying for novice users, "hands busy " applications in medicine, office dictation devices, or automatic voice translation into foreign languages. Such tantalizing application has motivated research in Automatic Speech Recognition (ASR) since 1950's. Since then, there are several commercial ASR systems developed, the most popular among them are: Dragon Naturally Speaking, IBM Viva voice and Microsoft SAPI.

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 even though the literary dialect throughout Kerala is almost uniform.

Numerous research and development have been taken place in various Indian languages during the recent years [9][12]. However, very less work has been reported in Malayalam language. A phonetic recognizer [11] and a wavelet based ASR [17] are the reported works in Malayalam.

In this paper we describe research on the speaker independent speech recognition of a continuous set of Malayalam digits without deliberate pause between each word.

Speech recognition is a highly complex task. The basic issue in speech recognition is dealing with two kinds of variability: acoustic and temporal [15]. Acoustic variability covers different accents, pronunciation, pitches, volume, and so on, while temporal variability covers different speaking rates. Development of a better acoustic modeling is the main task in Speech recognition research

In most of the current speech recognition systems, the acoustic modeling components of the recognizer are almost exclusively based on HMM [4][6][13]. HMM provides an elegant statistical framework for modeling speech patterns using a Markov process [6] that can be represented as a state machine. 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. The probability distribution associated with each state in an HMM, models the variability which occurs in speech across speakers or even different speech contexts.

Artificial Neural Networks (ANN) [1][5] are discriminative techniques that have been applied to speech recognition. One of the important draw backs of this model is that the temporal variations of the speech data can not be properly represented in ANN. Other difficulties encountered are; design of optimal model topologies, slow convergence during training and tendency to overfit the data.

Support Vector Machine (SVM) [18] is a classifier for machine learning with several applications [2]. The SVMs are effective discriminative classifiers with good generalizations and convergence property. Nevertheless, the application of these kernel-based machines to speech recognition is not cent percent perfect. The main draw backs are : i) SVMs, being a static classifiers, adaptation of the variability of duration of speech utterances is very difficult ; ii) ASR faces multiclass issues while SVMs are originally formulated as a binary classifier and iii) SVM training algorithms are very weak in managing huge databases typically used in ASR.

In this work, a public domain speech recognition development toolkit (CMU sphinx [7] is used for training and decoding. Phoneme based trigram model with 5 state HMM and left-to-right Bakis topology [6] are being used for this work.

978-1-4244-5612-3/09/$26.00 ©2009 IEEE

II. DESIGN AND DEVELOPMENT OF THE SYTEM

The design of the system is as shown in figure 1. Input signal is preprocessed and fed into the feature extraction module, where features which are useful for recognition are extracted. During training using the given database, parameters for acoustic and language modules were estimated and models created. For testing, the features of test speech are matched with the trained models.

The trainer creates 183 total models, including 25 base and 155 triphones. Single Gaussian per state was used and context independent tying was employed. The number of context independent states reduced from 1089 to 819 after tying using a decision tree. Total states were 1089 and when tied, produced 819 total tied states. The default frame size and frame shift were 25msec and 10msec respectively. The 39 dimensional vectors used consists of 13 MFCCs, corresponding delta and acceleration coefficients. The system incorporates multiple pronunciations by making multiple entries in the pronunciation dictionary.

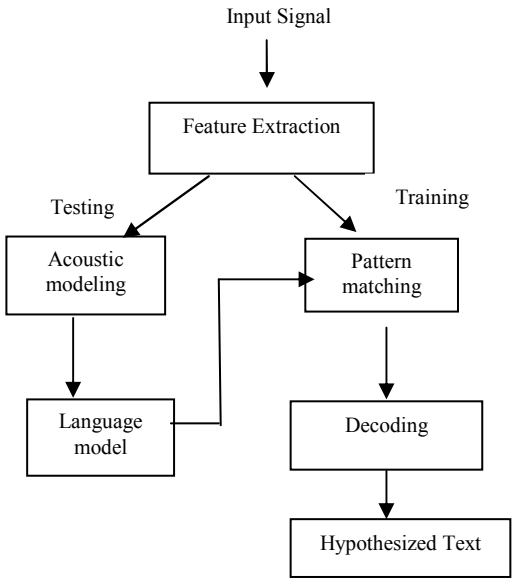


FIGURE 1. BLOCK DIAGRAM OF DIGIT SPEECH RECOGNTION SYSTEM

TABLE 1. MALAYALAM DIGITS USED IN THIS RESEARCH

Digit	Pronounciation	Malayalam Writing	IPA symbol	Pronunciation dictionary
0	puujyam'	പൂജ്യം	p u ˈ ʃ ˇ m	clp p uu j y a m
1	onnu'	ഒന്ന്	ɒ n ˈ	o n3 u’
2	ran't'u'	രണ്ട്	ɾ a ɳ t̪	r a n: vbd: d:
3	muunnu'	മൂന്ന്	m u ˈ n ˈ	m uu n3 u’
4	naalu'	നാല്	n a ː l ə̯	n aa l u’
5	anjchu'	അഞ്ച്	a ɲ c ˈ	a nj clc u’
6	aar’u’	ആറ്	a ː r ˈ	aa r ‘ u ‘
7	e'zu'	ഏഴ്	e ː ʒ ˈ	ee zh u’
8	et't'u'	എട്ട്	e d̪ ˈ	eclt tu
9	on_patu'	ഒന്പത്	ɒ n̪ p a t̪ ˈ	o m clp p a clt t u’ o n clp p a clt t u’

III. SPEECH DATABASE

Moreover, the system is designed to recognize any sequence (of any length) of Malayalam digits, therefore the size of the lexicon is eleven (including silence).

Speech was recorded in normal office environments. A headset which contains microphone with 70Hz to 16000 Hz of frequency range is used for recording. The recording is done with 16 kHz sampling frequency quantized by 16 bit, using a tool named CoolEdit in Microsoft wave format.

The database consists of 420 sentences. In order to capture all the acoustic variation across the boundaries of words, training database is designed to read small set of numbers which contain all possible pairs of digits. Accordingly, a sets of 7 digit numbers were generated, each set containing, 20 numbers capturing all distinct "word pairs". 21 speakers (10 male and 11 female) read 20 continuous strings of digits (7 digits) in normal manner.

The other database has utterance from five unknown speakers. The speakers were asked to utter any string of digits. Hence the database contained 25 sentences (5 speakers spoke five strings of digits). This database is exclusively used for online testing and to evaluate the speaker independence of the system.

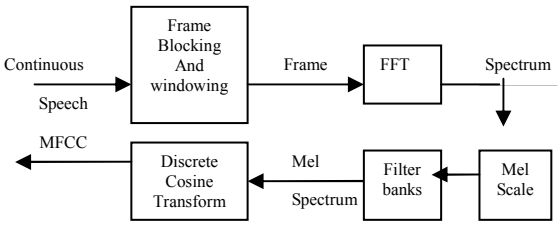
Transcription file is created for each utterance of the speaker and a language dictionary is created for each word in the string. These are stored in separate files. The vocabulary size of the language dictionary is 11. The phonetic dictionary contains 27 phonemes like units including silence.

IV. FEATURE EXTRACTION

For recognition of speech, the signals have to be represented with some specific features. Wide range of methods exists for parametric representation of sounds for the speech recognition tasks, such as Linear Prediction coding (LPC) [6], and Mel –Frequency Cepstrum coefficients (MFCC) [3]. MFCC is the well known popular method of feature extraction. To capture the phonetically important characteristics of speech, signal is expressed in Mel-Frequency Scale [14]. This scale has a linearly frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. MFCCs are less susceptible to the physical conditions of the speakers' vocal cord [13], compared to the speech wave forms. The block diagram of the feature extraction process is shown in figure 2.

.

FIGURE 2 . THE STEPS INVOLVED IN THE COMPUTATION OF MFCC



After the signal being preprocessed, it is fed into frame blocking and windowing process. Then the time domain signal is converted into frequency domain by applying Fast Fourier transform (FFT) on it. Then the spectrum is fed into Mel frequency wrapping. This involves two steps – mel-scale and filter banks. Here, for each tone of the signal, a subjective pitch is measured on the 'Mel'. For a given frequency f, measured in Hz, mels are calculated by (1)

$$mel(f) = 2595 \times \log_{10}(1 + f/700) \quad (1)$$

Mel Spectrum coefficients has to be converted to the time domain by applying Discrete cosine Transform (DCT) on it as in (2).

$$Cm = \sum_{k=1}^{N} \cos[m(k-0.5)\pi/N]Ek, m = 1,2..L \quad (2)$$

Where N is the number filters, L is the number of mel-scale cepstral coefficients, E_k is the log energy obtained from the filter. Applying the procedure described above, for each speech frame of about 25ms with overlap, a set of mel-frequency cepstrum coefficients are computed. These set of coefficients are called an acoustic vectors. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker [7]. Therefore each input utterance is transformed into a sequence of acoustic vectors.

V. STATISTICAL SPEECH RECOGNTION AND HMM MODEL

An unknown speech wave form is converted by a front-end signal processor into a sequence of acoustic vectors, $O = o_1\ o_2\ o_3\ o_4\ \ldots o_t$. The utterance consists of sequence of words $W = w_1, w_2, w_3\ \ldots . w_n$. In ASR it is required to determine the most probable word sequence, S, given the observed acoustic signal O. Applying Bayes' rule, [9]

$$S = \arg_w \max P(W/O) = \arg_w \max(P(O/W)P(W)/P(O))$$

$$S = \arg_w \max \underbrace{P(O/W)}_{posterior} \underbrace{P(W)}_{prior}$$

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. In this work, the acoustic modeling is computed by 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. 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), while the variables influenced by the states are visible. Each state has a probability distribution over the output. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states [6, 10, 13]. Thus HMM model can be defined as:

$\lambda = (Q, O, A, B, \Pi)$ Where ,

Q is $\{q_i\}$ (all possible states),

O is $\{v_i\}$ (all possible observations) ,

A is $\{a_{ij}\}$ where $a_{ij} = P(X_{t+1} = q_j \mid X_t = q_i)$ (Transition probabilities),

B is $\{b_i\}$ where $b_i(k) = P(O_t = v_k \mid X_t = q_{it})$ (Observation probabilities of observation k at state i),

$\Pi = \{\pi_i\}$ where $\pi_i = P(X_0 = q_i)$ (Initial state Probabilities), and

O_t Denote the observation at time t.

VI. TESTING AND TRAINING

Training is done by famous Baum-Welch algorithm [6] and testing by Viterbi algorithm [6]. In training phase knowledge models are created for the phonetic units. The database is divided into three equal parts and for each experiments, 2/3 of the data is selected for training and the remaining 1/3 is selected for testing. From the test results word accuracy rate for each set is calculated. Using the above trained model the system has also tested with speech from unknown speakers.

VII. PERFOMANCE EVALUTAION AND DISCUSSION

Word Error Rate (WER) is the standard evaluation metric for speech recognition. It is computed by SCLITE [8], a scoring and evaluating tool from National Institute of Standards and Technology (NIST). Inputs to SCLITE are the reference text and the output of the decoder is the recognized text (hypothesized sentence). WER align a recognized word string against the correct word string. If N is the number of words in the correct transcript; S, the number of substitutions; and D, the number of Deletions, then,

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

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

A. Result1: Performance of the system for Training data

For training and testing, the database is divided into three equal parts. For each experiment 2/3rd of the data is taken for training and the remaining 1/3rd for testing. Table 2 gives the digit recognition accuracy and number ('sentence') recognition accuracy obtained. The most confusing pairs obtained was muunu' =>onnu' and the most falsely recognized digit was onnu'.

TABLE 2: PERFORMANCE OF THE SYSTEM FOR TRAINING DATA

Experiment Number	Word Recognition Accuracy		Sentence Recognition Accuracy	
	%		%	
	Train	Test	Train	Test
1	99.8	98.5	99.29	96.43
2	99.8	98.9	97.86	94.29
3	99.3	95.9	98.57	82.86
Average	99.63	97.76	98.57	91.19

B. Result 2: Performance of the system for test(unseen data)

In order to test the system for live application, five speakers whose voice was unknown to the system were selected. They were asked to utter any sequence of digits of any length. The test result gave an accuracy of 95.7%.

VIII. CONCLUSIONS

This paper has illustrated recognition system for Malayalam language numbers using Hidden Markov Models. It recognizes any combination of Malayalam digits pronounced without pause between the digits. Spoken number recognition system provides a user-friendly interface for feeding numeric data into computers. The accuracy of the system was found to be satisfactory. The accuracy can be further improved by using larger training data; including utterance from a large of speakers with variations in age and accent.

ACKNOWLEDGMENT

We would like to thank Dr. Samudravijaya K, School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, for his invaluable support, encouragement and guidance, without which this work would not have been completed.

REFERENCES

[1] A. Sperduti and A. Starita, "*Supervised Neural Networks for Classification of Structures*", IEEE Transactions on Neural Networks, 8(3): pp.714-735, May 1997.

[2] 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.

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

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

[5] E. 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.

[6] F.Felinek, "*Statistical Methods for Speech recognition"* MIT Press, Cambridge, Massachusetts, USA,1997.

[7] http://cmusphinx.sourceforge.net.

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

[9] 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.

[10] Jurafsky, D., and Martin, J.H (2007). "S*peech and Language processing: An introduction to natural Language processing, computational linguistics, and Speech recognition "*, 2nd edition, http://www.cs.colorado .edu/~martin/slp2.html.

[11] 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.

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

[13] Rabiner L R, "*A Tutorial on Hidden Markov Models and selected Application s in Speech Recognition*" Proc. IEEE, vol. 77, 1989, pp. 257 – 286.

[14] S.S Stevens and J. Volkman (1940), " The *relation of pitch to frequency"*, American Journal of Psychology, vol. 53(3), pp 329-353..

[15] Saumudravijaya K, "*Hindi Speech Recognition*" (2001), J. Acoustic Society India, 29(1), pp 385-395.

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

[17] Syama R, Suma Mary Idikkula (2008) " HMM Based Speech Recognition System For Malayalam", ICAI'08 – The 2008 International Conference on Artificial Intelligence, Monte Carlo Resort, Las Vegas, Nevada, USA (July 14-17, 2008) (Accepted).

[18] V. N. Vapnik, "*Statistical Learning Theory"*. New York: Wiley, 1998.

.