Sādhanā Vol. 38, Part 6, December 2013, pp. 1339–1346. © Indian Academy of Sciences

Connected digit speech recognition system for Malayalam language

CINI KURIAN and KANNAN BALAKRISHNAN

Department of Computer Applications, Cochin University of Science & Technology,
Cochin 682 022, India
e-mail: cinikurian@gmail.com; bkannan@cusat.ac.in

MS received 16 November 2011; revised 26 May 2013; accepted 18 June 2013

Abstract. A connected digit speech recognition is important in many applications such as automated banking system, catalogue-dialing, automatic data entry, automated banking system, etc. This paper presents an optimum speaker-independent connected digit recognizer for Malayalam language. The system employs Perceptual Linear Predictive (PLP) cepstral coefficient for speech parameterization and continuous density Hidden Markov Model (HMM) in the recognition process. Viterbi algorithm is used for decoding. The training data base has the utterance of 21 speakers from the age group of 20 to 40 years and the sound is recorded in the normal office environment where each speaker is asked to read 20 set of continuous digits. The system obtained an accuracy of 99.5 % with the unseen data.

Keywords. Malayalam speech recognition; perceptual linear predictive (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 (Sukhminder & Dinesh Kumar 2002). 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 (Jurasky & Martin 2007).

Keyboard is a popular medium which 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 (Lawrence & Biing-Hwang 2005).

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.

ASR is a branch of Artificial Intelligence (AI) and is related with number of fields of knowledge such as acoustics, linguistics, pattern recognition, etc. (Gold & Morgan 2005). 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 (Cini & Kannan 2009). 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 (Jyoti & Singhai Rakesh 2007).

In most of the current speech recognition systems, the acoustic component of the recognizer is exclusively based on HMM (Felinek 1997). 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) (Behrman *et al* 2000) and Support Vector machines (SVM) (Burges 1998) are other techniques which are being applied to speech recognition problems. In ANN and SVM, temporal variation of speech cannot be effectively represented compared to HMM.

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) (Davis & Mermelstein 1980) and Mel-Frequency Cepstrum Coefficients (MFCC) (Huang *et al* 2001) and Perceptual Linear Predictive (PLP) coefficient (Felinek 1997). Since PLP is more adapted to human hearing, PLP parameterization technique is used in this work.

In this work, an HMM-based public domain speech recognition development toolkit CMU sphinx (Hyassat & Abu Zitar 2008) is used for signal processing and acoustic modelling.

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 coefficient 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, robust and capable of capturing change of the spectrum with time.

The perceptual linear prediction (PLP) (Hermansky 1990) method 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 figure 1.

The primary step in any feature extraction process is blocking the frame. Here audio signals which are basically non-stationary are cut into fragments called frames. During windowing these frames are passed through Hamming Window. In spectral analysis, signal is passed though Fourier Transform process and then power spectrum of the signal is computed. Then during critical band analysis these power spectrum is fed into filter banks. The resulting spectrum is multiplied by the equal loudness curve and at the intensity loudness conversion stage, it is raised to the power of 0.33 to simulate the power law of hearing.

After spectrum analysis and inverse discrete Fourier transform, the all-pole model of LPC is applied to give a smooth and compact approximation. The post processing involves the computation of cepstral coefficients as well as the first and the second time differences between parameter values over successive frames–delta, and delta–delta coefficients, etc.

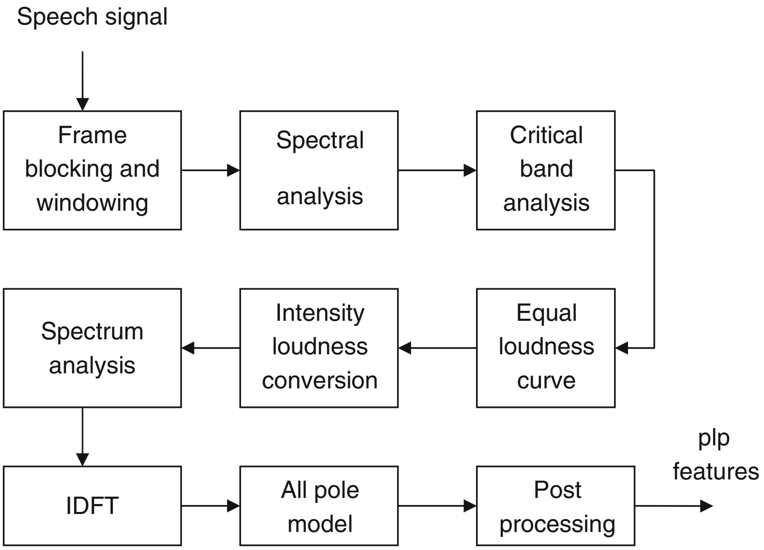

Figure 1. Block diagram of PLP extraction.

The three important steps which make PLP features different from other feature extraction methods are explained below.

2.1a *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' namely bark frequency scale, represented as equation (1)

$$\mathrm{Bark} = 13\mathrm{atan}\,(0.76\mathrm{f}/1000) + 3.5\mathrm{atan}\left(\mathrm{f}^2/7500^2\right). \tag{1}$$

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

2.1b *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 40 dB level.

2.1c *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 (Stevens 1957), 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, $\mathrm{O} = o_1, o_2, o_3, \ldots, o_n$. Assuming that the utterance consists of sequence of words $W = w1, w2, w3, \ldots, wn$. The problem here is to determine the most probable word sequence, $\grave{\mathrm{W}}$ which matches O best. Hence $\grave{\mathrm{W}} = \mathrm{arg}w\mathrm{max}\ P(W/O)$.

Using Bayes' rule (Lawrence & Biing-Hwang 2005)

$$P\,(W/O) = P\,(O/W)\,P\,(W)\,/P(O),$$

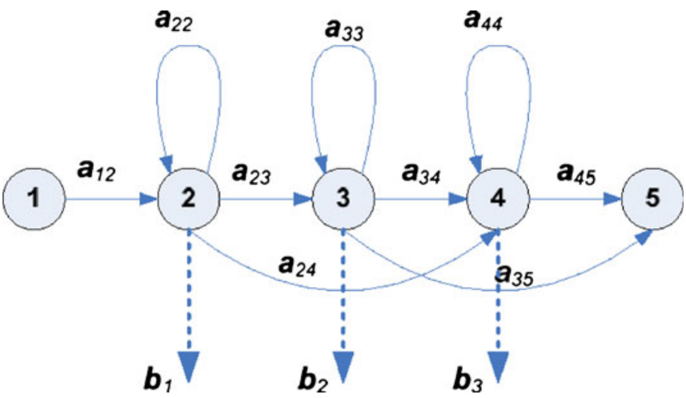

Figure 2. Topology of a 5 state HMM.

where $P(O)$ is the priori probability of the feature sequence. Since it is independent of the acoustic and language model, it can be ignored in the maximization operation. Hence

$$\grave{W} = \mathrm{arg}w\mathrm{max}\, \underset{Posterior}{P(O/W)}\, \underset{prior}{P(W)} . \tag{2}$$

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) 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 modelling of this speech recognition system is done using HMM. 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 (Felinek 1997; Huang *et al* 2001). These transition probabilities are denoted by matrix A. Here A is defined as $A = a_{ij}$, where $a_{ij} = P(t_{t+1} = j|j = i)$, 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

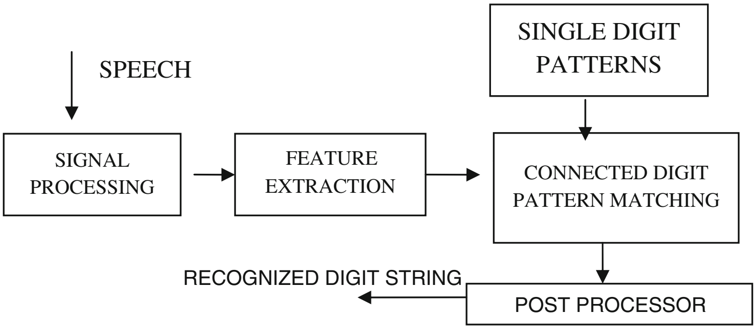

Figure 3. Block diagram of connected digit recognition method.

B. Here B is defined as $B = b_j(k)$, where $b_j(k) = P(v_k\ at\ t\ |i_t = j)$, 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 $\lambda = \{A, B, \pi\}$.

Basically, there are three ASR problems which can be well addressed with HMM. They are (i) recognition problem (decoding), (ii) optimization problem (scoring and evaluation) and (iii) training problem.

Problem (i) is associated with decoding or hidden state determination, where the best HMM state sequence is to be determined from the given observation sequence. The Viterbi algorithm (Jurasky & Martin 2007) is employed for solving the problem (i) as it is computationally efficient. Problem (ii) is Scoring and evaluation problem i.e., computing the likelihood of an observation sequence, given a particular HMM. This problem occurs in 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 (Lawrence & Biing-Hwang 2005). 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 (iii) 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–Welch algorithm (Jurasky & Martin 2007).

3. Speech database and system development

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 70 Hz 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. 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.

The block diagram of the connected digit recognition system is shown in figure 3. There are three basic steps in the recognition process. In spectral analysis, in which the speech signal, is converted to an appropriate spectral representation. Connected word pattern matching is the process, in which the sequence of spectral vectors of the unknown (test) connected digit string is matched against whole word (single digit) patterns. The output of this process is a set of candidate digit strings, generally of different lengths, ordered by distance score. In post processor, the candidate digit strings are subjected to further processing so as to eliminate unreasonable candidates. The post processor chooses the most likely digit string from the ordered list of candidates which passed the postprocessor tests.

Table 1. Performance evaluation of the system with training and testing data.

	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	100	0	1	0	99.57	0	1	0
2	99.57	0	1	0	99.29	0	2	0
3	100	0	0	0	99.57	0	1	0

4. Training and testing

Training is done by famous Baum–Welch algorithm (Felinek 1997) and testing by Viterbi algorithm (Felinek 1997). In training phase, knowledge models are created for the phonetic units. 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 the remaining 1/3rd of the data is used for testing. In 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.

5. Performance evaluation and discussion

Word Error Rate (WER) is the standard evaluation metric for speech recognition. It is computed by SCLITE (Jurasky & Martin 2007), a scoring and evaluating tool from National Institute of Standards and Technology (NIST). The inputs to the SCLITE are the reference text and the recognized text (output of the decoder). After each training and testing experiment, recognized text and the reference text are converted into the sclite compatible form and fed as input in sclite. Then detailed results are 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, $WER = ((S+D+I)N)/100$. Sentence Error Rate (SER) = (number of sentences with at least one word error/ total number of sentences) * 100.

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.

As detailed in table 1, the test experiment is conducted in three sessions as explained in the training and testing session above. In each experiment, speakers who are involved in the training session are excluded in the testing session. Hence the system obtained a word accuracy of 99.5% by averaging the word accuracy of test experiment 1, 2 and 3 as detailed in table 1.

6. Conclusion

This paper has illustrated the speaker-independent Malayalam connected digit speech recognition system using hidden Markov model and PLP cepstral coefficient. From the results of the experiment it can be concluded that PLP and HMM are ideal candidates for spoken connected digit recognition system. The system achieves almost cent percent (99.5) recognition accuracy.

References

Behrman L, Nash J, Chandrashekar S V and Skinner S 2000 Simulations of quantum neural networks. *Inf. Sci*. 128: 257–269

Burges C J C 1998 A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov*. 2(2): 995–974

Cini Kurian and Kannan Balakrishnan 2009 Speech recognition of Malayalam numbers. *IEEE Transaction on Nature and Biologically Inspired computing* (NaBIC-2009) 1475–1479

Davis S and Mermelstein P 1980 Comparison of parametric representations for Monosyllabic word Recognition in continuously spoken sentences. *IEEE Trans. On ASSP* 28(4)(2): 357–366

Felinek F 1997 Statistical methods for speech recognition. Cambridge Massachusetts, USA: MIT Press

Gold B and Morgan N 2005 Speech and audio signal processing. N.Y: John Wiley and Sons

Hermansky H 1990 Perceptual linear predictive PLP analysis of speech. *J. Acoust. Soc. Am*. 57(4): 1738–52

Huang X, Alex A and Hon H W 2001 ‘*Spoken Language Processing; A Guide to Theory, Algorithm and System Development*’, New Jersey: Prentice Hall, Upper Saddle River

Hyassat H and Abu Zitar R 2008 Arabic speech recognition using SPHINX engine. *International Journal of Speech Technology* Springer, 133–150

Jurasky D and Martin J H 2007 *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*, 2nd edition Prentice-Hall

Jyoti and Singhai Rakesh 2007 ‘Automatic speaker recognition 2007: An approach using DWT based feature extraction and vector quantization’. *IETE Tech. Rev*. 24(5): 395–402

Lawrence Rabiner and Biing-Hwang Juang 2005 Singapore ‘*Fundamentals of Speech Recognition*’, 2nd edn. pp 49 Pearson Education

Stevens S S 1957 On the psychophysical law. *Psychol. Rev*. 64(3): 153–11

Sukhminder Singh Grewal and Dinesh Kumar 2002 Isolated word Recognition System for English language. *International Journal of Information Technology and Knowledge Management* 2(2): 447–450