Comparison of HMM and DTW for Isolated Word Recognition System of Punjabi Language

Kumar Ravinder

Department of Computer Science & Engineering,
Thapar Univeristy, Patiala – 147004 (India)
ravinder@thapar.edu

Abstract. Issue of speech interface to computer has been capturing the global attention because of convenience put forth by it. Although speech recognition is not a new phenomenon in existing developments of user-machine interface studies but the highlighted facts only provide promising solutions for widely accepted language English. This paper presents development of an experimental, speaker-dependent, real-time, isolated word recognizer for Indian regional language Punjabi. Research is further extended to comparison of speech recognition system for small vocabulary of speaker dependent isolated spoken words in Indian regional language (Punjabi) using the Hidden Markov Model (HMM) and Dynamic Time Warp (DTW) technique. Punjabi language gives immense changes between consecutive phonemes. Thus, end point detection becomes highly difficult. The presented work emphasizes on template-based recognizer approach using linear predictive coding with dynamic programming computation and vector quantization with Hidden Markov Model based recognizers in isolated word recognition tasks, which also significantly reduces the computational costs. The parametric variation gives enhancement in the feature vector for recognition of 500-isolated word vocabulary on Punjabi language, as the Hidden Marko Model and Dynamic Time Warp technique gives 91.3% and 94.0% accuracy respectively.

Keywords: Dynamic programming (DP), dynamic time warp (DTW), hidden markov model (HMM), linear predictive coding (LPC), Punjabi language, vector quantization (VQ).

1 Introduction

Research in automatic speech recognition has been known for many years, various researchers have tried to analyze the different aspects of speech in Indian languages. Existing literature reveals that Punjabi (ਪੰਜਾਬੀ in Gurmukhi script) is highly phonetic language as compared to other national and international languages. Punjabi language is one of the popular and used north western India and in Pakistan. Gurmukhi script alphabet consists of 41 consonants and 10 vowels (laga matra), two symbols for nasal sounds (bindī and ṭippī), and one symbol which duplicates the sound of any consonant (addak) with writing style from left to right. Three consonants are used in Punjabi as conjuncts.

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 244–252, 2010.

Recently "British researchers have used the Punjabi language to help narrow down the identity of writers and develop a technique that could profile criminal authors of documents [1]. In consecutive phoneme that has immense variation, end point detection is highly difficult. However in time-domain equation the deviation of preemphasis filter enhances filtering of end point detection.

As indicated in figure 1, the speech recognition system contains four components: end point detection, linear predictive coding (LPC) processor, statistical-pattern-recognition techniques HMM /DTW and recognition process [2] [3]. The speech recognition has two algorithmic procedures to deal with the non-stationary speech signals: the temporal alignment technique and markov modeling. The time warping technique is combined with linear predictive coding analysis in DTW approach. In HMM approach, well known techniques of vector quantization and hidden markov modeling are combined with a linear predictive coding analysis. The DTW approach uses the nearest neighbor (NN) decision rule and HMM uses the maximum likelihood (ML) decision rule.

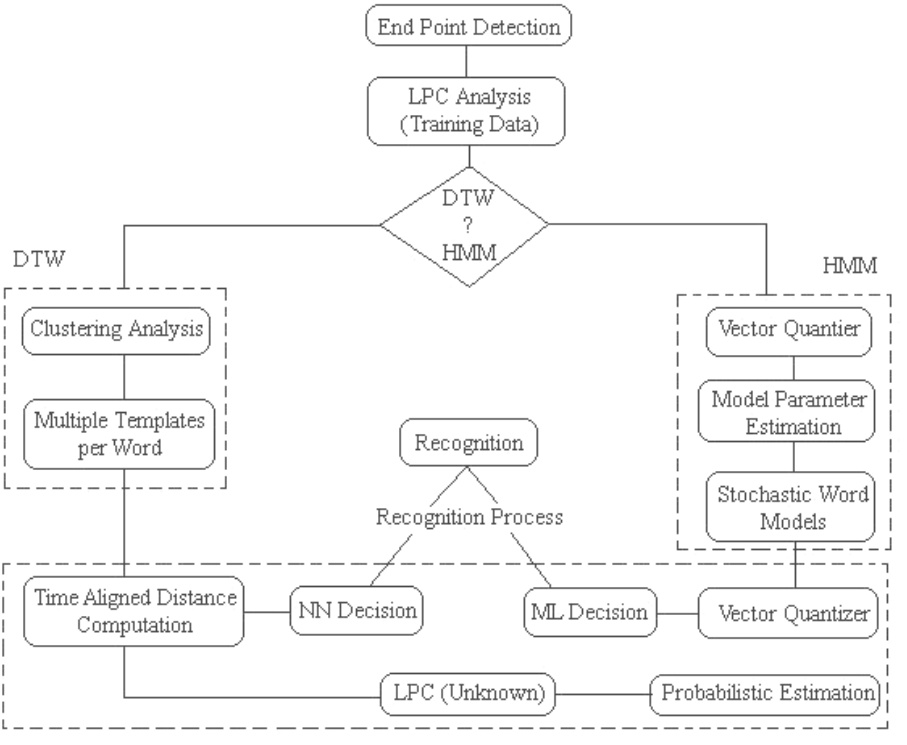

Fig. 1. Block diagram of Speech Recognition System

These methods display certain superficial similarities, as a result of which, it has occasionally been claimed that they are identical. From the simulation, it is clear that the methods are not identical. While their overall performances are comparable, they

appear to make different errors, and involve different amounts of computation and different complexities of training. This recognition system is implemented using Visual C++ with Multimedia API in Windows environment. The speech is recorded in the form of wave file using RIFF structure. The proposed system speech has to be sampled at 6.67 kHz, 16 kHz sampling frequency, sampling size 8 bit on mono channel and recording of a single word within limit 3000 milliseconds. Threshold energy 10.1917 dB is used in the word detection [4] [5] [6].

The organization of this paper is as follows. In section II, the system model is used for endpoint detection and feature extraction. Further, section III elaborates working of HMM and DTW approaches on Punjabi language. In section IV, experimental performance has been conducted and compared. Finally section V provides the concluding remarks.

2 System Model

2.1 End point Detection

The significant effort on the implementation of speech recognition is the problem of extricating background silence before and after the input speech. To find, the energy of the speech signal *Wave (s) s=48000*, it is formatted into blocks of 10 ms and each block, we define *Wave (n) is*, *n(1..160)* to be the n^{th} sample in the block. The log energy E_s of a block of length N samples is

$$E_s = 10 * \log 10(\in + \frac{1}{N}\sum_{n=1}^{N} Wave^2(n)) \tag{1}$$

Where $\in = 1.0\ e - 007$ is a small positive constant added to prevent the computing of log zero. Hence, log energy E_s is used for end point detection [7].

2.1 LPC Feature Analysis

To compute LPC feature analysis involves the following operation for each speech frames [2].

Preemphasis: The digitized speech signal is processed by a first order digital network in order to spectrally flatten the recorded signal *Wave(s)*.

$$pre(s) = Wave(s) - \alpha Wave(s-1) \tag{2}$$

$$Where\ \alpha = 0.9731$$

Blocking into frames: Here N is consecutive speech samples (we use N=320 corresponding to 20 ms of signal) are used as a single frame. Consecutive frames are spaced M sample apart (we use M = 160 corresponding to 10ms frame spacing or 20ms frame overlap) and L is the number of frames in voiced speech word.

$$X(\ell)(n) = pre(M\ell + n) \tag{3}$$

Frame Windowing: Each frame is multiplied by an N-samples window (we use Hamming window) so as to minimize the adverse effect of chopping N-sample out of the running speech signal.

$$W(\ell)(n) = (0.54 - 0.46\cos(\frac{2\pi n}{N-1}))X(\ell)(n), \tag{4}$$

Where $0 \le n \le N-1$ *and* $0 \le \ell \le L-1$

Autocorrelation Analysis: Each windowed set of speech samples is autocorrelated to a given a set of *(p + 1)* coefficient, where *p* is the order of the desired LPC analysis (we use *p =11*).

$$R(\ell)(k) = \sum_{n=0}^{N-1-k} W(\ell)(n)W(\ell)(n+k) \tag{5}$$

Where $0 \le k \le p$, $0 \le n \le N-1$ *and* $0 \le \ell \le L-1$

LPC/Cepstral Analysis: For each frame, a vector of LPC coefficients is computed from the autocorrelation vector using a Levinson-Durbin recursion method. An LPC derived cepstral vector is then computed up to the Q^{th} component, where Q is the order of the cepstral coefficients and Q > k, Q = 11 used in this work.

Weighted Cepstral Coefficient: After computing cepstral coefficients, weight is given to them by multiplying them with cepstral weight. It will enhance the portion of the cepstrum in vocal tract information. The computation involved at this step is:

$$W_C(k) = 1 + \frac{Q}{2}\sin\left(\frac{\pi k}{Q}\right) \tag{6}$$

Where $0 \le k \le p$

$$WC_C(\ell)(k) = W_C(k) \cdot C_C(\ell)(k) \tag{7}$$

Where $0 \le k \le p$ *and* $0 \le \ell \le L-1$

Delta Cepstrum: The time derivative of the sequence of the weighted cepstral vectors is approximated by a first order orthogonal polynomial over a finite length window of *(2k+1)* frames centered on the current vector. The value of *k = 2* i.e., *5* frame window is used for computing the derivative). The cepstral derivative (i.e., the delta cepstral vector) is computed as

$$D_C(\ell)(m) = \left[\sum_{k=-K}^{K} kWC_C(\ell-k)(m)\right] \cdot G \tag{8}$$

Where $0 \le m \le Q$

Where *G* is a gain term chosen to make the variances of weighted cepstral coefficient and delta cepstral coefficient equal (A value of G of 0.375 was used.) [8] [9].

2.2 Vector Quantization

The VQ codebook is a discrete representation of speech. We will generate the codebook by using LBG algorithm. This algorithm has used two times, one is on training time and other is testing time. In training time we will generate codebook for delta cepstrum coefficients. In testing we will use stored codebook for getting the indices of codebook that give minimum distortion. The VQ will find a codebook index corresponding to the vector that best represents a given spectral vector, for an input vector sequence *V{v(1), v(2), v(3), ….., v(N)}, VQ* will calculate the vector distance between each vector in codebook *C{c(1), c(2), c(3), …, c(P)}* and each vector *v(n)*, and the codebook index with minimum distance will be chosen as output. After VQ, a sequence of codebook indexes *I{i(1), i(2), i(3),…., i(N)}* will be produced. The vector distance between an input vector *v(n)* and each vector in codebook are calculated as follows:

$$Distance\ (p) = \sum_{k=1}^{k=11} [v(n)(k) - c(p)(k)]^2 \tag{9}$$

$$i(n) = \arg\min_p (Distance\ (p)) \tag{10}$$

The vector quantization codebook of size $p = 256$ and vector length $k = 11$ are used. This size selection is based on the experimental results [8] [10].

3 HMM and DTW for Isolated Word Recognition

Dynamic Time Warp approach and Hidden Markov approach for isolated word recognition of small vocabulary are implemented. The time warping technique is combined with linear predictive coding analysis and HMM approach used with vector quantization and hidden markov modeling are combined with a linear predictive coding analysis.

3.1 Hidden Markov Model

LPC analysis followed by the vector quantization of the unit of speech, gives a sequence of symbols (VQ indices). HMM is one of the ways to capture the structure in this sequence of symbols. In order to use HMM in speech recognition, we should have some means to achieve the following.

Evaluation: Given the observation sequence $O=O_1, O_2, \ldots, O_T$, and the model $\lambda=(A, B, \pi)$, how we compute Pr $(O|\lambda)$, the probability of the observation sequence. The evaluation problem is efficient way of computing this probability using forward and backward algorithm.

Decoding: Given the observation sequence $O=O_1, O_2, \ldots, O_T$, how we choose a state sequence $I=i_1, i_2, \ldots, i_T$, which is optimal in some meaningful sense. The decoding helps to find the best matching state sequence given an observation sequence and solve using Viterbi algorithm.

Training: How we adjust the model parameters $\lambda=(A, B, \pi)$ to maximize Pr $(O|\lambda)$. The training problem resolved by Baum-Welch algorithm [11] [12][15].

3.2 Dynamic Time Warping

The DTW deals with features and distances (local and global) concept. To obtain a global distance between two speech patterns (represented as a sequence of vectors) a time alignment must be performed. Dynamic time warp alignment that simultaneously provides a distance score associated with the alignment. Consider a matrix with $N \times n$ *(N reference, n input signal)* and a local distance *l(i, j)* which returns a distance associated with *i, j* (where *i* template of reference, *j* of input signal). Compute the shortest path with minimum distance from each cell of $N \times n$ matrix. This problem exhibits optimal substructure. The solution to the entire problem relies on solutions to sub problems. Let us define a function *g (i, j)* as

$$g(i, j) = \textit{the minimum distance to reach at final } (i, j) \textit{ in } N \times n$$

$$g(i,j) = \begin{cases} \infty & j < 1 \quad or \quad j > n \\ l(i,j) & i = 1 \\ \min(g(i-1,j-1), g(i-1,j), g(i-1,j+1)) + l(i,j) & otherwise \end{cases}$$

This is recursive process and global distance is the value at top most cell of the last column in matrix. On training time the database of the features LPC Coefficients of the training data is created. In testing time, the input pattern (features vector of the test token) is compared with each reference pattern. The distance scores for all the reference patterns are sent to a decision rule, which gives the word with least distance as recognized word. The distance measure between two feature vectors is calculated using Euclidean distance metric [13] [14][15].

4 Experimental Performance

4.1 Performance Based on Punjabi Numerals

The recordings were done for one male speaker. The experimental result gives incremented accuracy as increases the size of the codebook, but constantly increasing the complexity and time of recognition increases. So the codebook size of 256 for the comparison between DTW and VQ/HMM techniques is used. (Tested various size of code book but give more accuracy/time results optimized on only size of 256). The train the system for Punjabi numerals is the five times. The recognition results of numerals are show in figure 2.

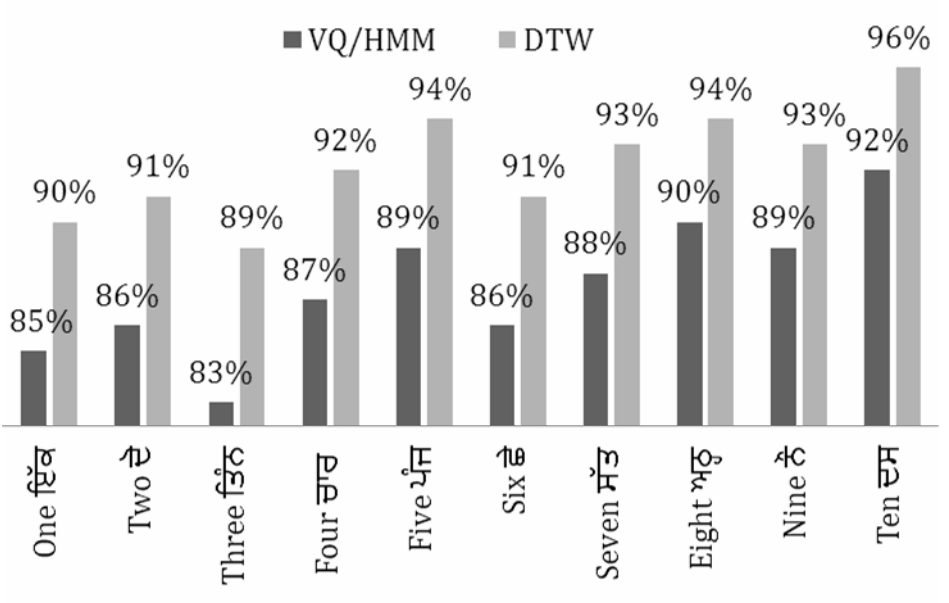

Fig. 2. Performance comparison of Punjabi numerals between DTW and VQ/HMM

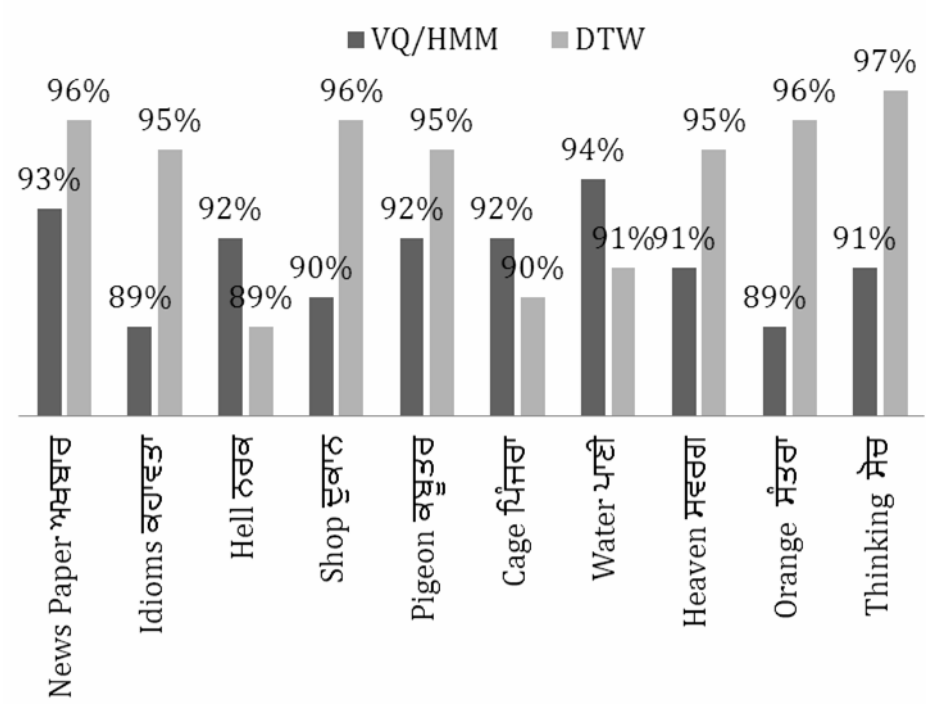

Fig. 3. Performance comparison of Punjabi words between DTW and VQ/HMM

4.2 Performance Based on Randomly Selected Words

The recognition vocabulary of Punjabi words (the English word "Newspaper" is translated and pronounced as "ਅਖਬਾਰ" in Punjabi language). The system for each of the words was trained with 10 utterances of the same word. The comparison for the performance of the VQ/HMM recognizer with DTW recognizer, a sub set of 10 randomly selected words from 500 trained words set was tested on both of the recognizers. The recognition accuracies of the VQ/HMM recognizer vs. DTW recognizer are as shown in figure 3.

5 Conclusion

Results carried out by above experiments reveal that performance of HMM recognizer is somewhat poorer than the DTW based recognizer because of the insufficiency of the HMM training data. However with the increase in the size of the codebook, the accuracy of the HMM based recognizer may improve but that will increases the complexity of the system. The time and space complexity of the HMM approach is less as compared to the DTW approach for same size of codebook because during HMM testing we have to compute the probability of each model to produce that observed sequence. In DTW testing, the distance of the input pattern from every reference pattern is computed, which is computationally more expensive. In experimental result the system gives very impressive performance for Punjabi language numerals with DTW is 92.3% and with HMM is 87.5%. The experimental research reveals that DTW approach is more appropriate for Punjabi numerals and isolated spoken words. Further, the results have also shown that the errors made by the two recognizers are largely disjoint. Hence there exist the potential of DTW using some fairly standard techniques for Punjabi Language as compared to HMM approach for better accuracy because is more phonetic as compared to other languages.

References

1. Prasun, S.: British experts use Gurmukhi to aid forensic research. In: Indo-Asian News Service, London. Hindustan Times (September 21, 2007)
2. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
3. George, M.W., Richard, B.N.: Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming. IEEE Transaction on Acoustics, Speech, and Signal Processing ASSP-24(2), 183–188 (1976)
4. Bovbel, E.L., Kheidorov, I.E.: Statistical recognition methods, application for isolated word recognition. IEEE Transaction on Digital Signal Processing, 821–823 (June 1997)
5. Guan, C., Zhu, C., Chen, Y., He, Z.: Performance Comparison of Several Speech Recognition Methods. In: 1994 International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, pp. 13–16 (April 1994)
6. Levinson, S.E., Rabiner, L.R., Sondhi, M.M.: Speaker Independent Isolated Digit Recognition Using Hidden Markov Models. In: International Conference on Acoustics, Speech, and Signal Processing, Paper 22.8, pp. 1049–1052 (April 1983)

7. Picone, J.W.: Signal Modeling Techniques in Speech Recognition. Proceedings of the IEEE 81(9), 1214–1245 (1993)
8. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice Hall PTR, Englewood Cliffs (1993)
9. Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., Picheny, M.A.: A Method for the Construction of Acoustic Markov Models for Words. IEEE Transaction on Speech and Audio Processing 1(4), 443–452 (1993)
10. Soong, F.K., Rosenberg, A.E., Rabiner, L.R., Juang, B.H.: A Vector Quantization Approach to Speaker Recognition. In: Conference Record 1985 IEEE International Conference on Acoustics, Speech, and Signal Processing, Paper 11.4.1, pp. 387–390 (March 1985)
11. Rabiner, L.R., Juang, B.H., Levinson, S.E., Sondhi, M.M.: Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities. Bell System Tech. Jour. 64(6), 1211–1234 (1985)
12. Rabiner, L.R., Juang, B.H.: An Introduction to Hidden Markov Models. IEEE ASSP Magazine 3(1), 4–16 (1986)
13. Rabiner, L.R., Schmidt, C.E.: Application of Dynamic Time Warping to Connected Digit Recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP 28(4), 377–388 (1980)
14. Myers, C.S., Rabiner, L.R., Rosenberg, A.E.: Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP 28(6), 623–635 (1980); Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
15. Axelrod, S., Maison, B.: Combination of hidden markov model with dynamic time warping for speech recognition. In: Proceedings of the ICASSP, pp. 173–176 (2004)