Speech Recognition using Vector Quantization

H B Kekre
Department of Computer Science
Mukesh Patel School of Technology Management & Engineering, NMIMS University, Mumbai, India
+91-9323557897
hbkekre@yahoo.com

A A Athawale
Department of Computer Science
Thadomal Shahani College of Engineering, Mumbai University
India
+91-9226977842
athawalearchana@gmail.com

G J Sharma
Department of Computer Science
K J Somaiya College of Engineering
Mumbai University
India
+91-9324352009
grishmajsharma@yahoo.com

ABSTRACT

This paper presents a novel method for isolated English word recognition based on energy and zero crossing features with vector quantization. This isolated word recognition method consists of two phases, feature extraction phase and recognition phase. In feature extraction, end points are detected and noise is removed using end point detection algorithm, a feature vector is obtained by combining the energy and zero cross rate into a single vector of twenty dimensions. Recognition phase consists of two steps, feature training and testing, in feature training, codebooks for each reference samples are generated using LBG Vector Quantization algorithm. For testing Euclidean distance is calculated between test sample feature vector and codebook of all reference speech samples. Speech sample with minimum average distance is selected. Experimental results showed that the maximum recognition rate of 85% is obtained for codebook size of 4.

Categories and Subject Descriptors

I.2.7 Natural Language Processing.

General Terms

Algorithms, Measurement, Performance, Design, Experimentation, Verification.

Keywords

Isolated Speech Recognition, Vector Quantization, Codebook, Euclidean Distance.

1. INTRODUCTION

Speech recognition in a computer system domain may be defined as the ability of computer systems to accept spoken words in an audio format- such as wav or raw and then generates its content in text format. There have been many interesting advances and developments since the invention of the first speech recognizer at Bell Labs in the early 1950's. Besides inventing useful automated speech recognizers, scientists and researchers' contributions were to produce efficient algorithms that help to produce better quality automated speech recognition systems, and improve the accuracy and matching standards in order to make the systems more useful. The main goal of designing this isolated-word automatic speech recognition system is to automatically extract the spoken word from the input speech signal and finally return the correctly matched word to the user.

There are two main problems that have been considered for this research paper. The first is features extraction issue, whereas the second is speech classification/recognition issue.

According to a literature conducted, there are various speech features extraction techniques, including Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and Mel-Frequency Cepstral Coefficient (MFCC). However, MFCC has been the most frequently used technique especially in speech recognition and speaker verification applications [7]. MFCC features lead to a performance that is slightly superior to PLP and thus to LPC [6]. In addition, MFCC analysis gives better performance than the PLP derived cesptral in an unconstrained monophone test [6].

The problem of speech recognition belongs to a much broader scientific topic called pattern recognition or pattern matching/classification. Spoken language processing relies heavily on pattern recognition, which is one of the most challenging problems for machines [4].

Hidden Markov Model (HMM), Neural Networks (NN) and Vector Quantization (VQ) are the most frequently used pattern recognition techniques in speech recognition field.

The approach described in this paper is a speaker-independent, isolated word recognizer for English language. It uses Energy and Zero crossing features and Vector Quantization as pattern recognition technique.

The paper is organized as follows, section 2 is dealing with the recognition process, in this we discuss the feature extraction technique, vector quantization, Euclidean distance measure, section 3 describes the database development and recognition procedure and we discuss the performance analysis and results in section 4.conclusion and future scope is discussed in section 5.

2. THE RECOGNITION PROCESS

The General scheme for Speech Recognition is shown in Fig. 1.The isolated-word automatic speech recognition system operates in two phases. The first phase is known as training phase during which the system learns the reference patterns representing the different speech sounds (by words) that constitute the vocabulary of the application. Each reference is learnt from a set of spoken words and stored in form of database / codebook that acts as a template using statistical properties of the speech pattern. The second phase is known as testing phase where an unknown speech signal or new recorded speech signal is identified using the stored

ICWET'11, February 25–26, 2011, Mumbai, Maharashtra, India.

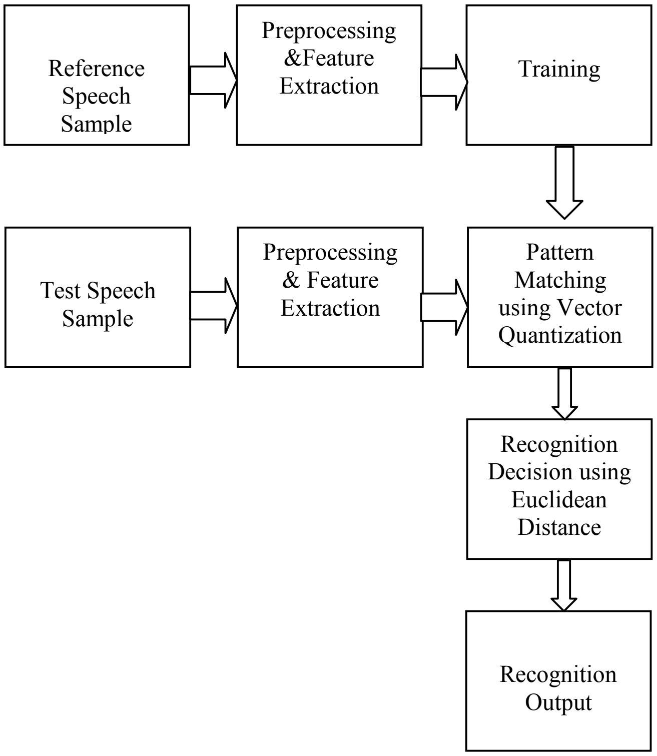
Reference Speech Sample
Preprocessing &Feature Extraction
Training
Test Speech Sample
Preprocessing & Feature Extraction
Pattern Matching using Vector Quantization
Recognition Decision using Euclidean Distance
Recognition Output

Figure 1. Speech Recognition System

Reference patterns or the template resulted from the training phase. Reference speech samples are selected from the database, feature vectors of all the reference speech samples are calculated. For feature training, codebooks for each reference samples are generated using LBG Vector Quantization algorithm. For feature matching unknown word is selected from database as test sample. Feature vector is calculated for this test sample. The Euclidean distance is calculated between test sample feature vector and codebook of all reference speech samples. Speech sample with minimum average distance is selected.

2.1 Preprocessing and Feature Extraction

The speech signal is stored in form of wave files and is read. The speech samples thus obtained are stored for further computation. It is important in a speech recognition system that beginning and ending of an utterance are accurately known. This is not only reduces the amount of data that needs to be processed but also discriminates the utterance against background noise. Endpoints are detected using energy and zero crossing measures.

The energy level of signal is calculated for each frame [3]. The average energy level and zero crossing rate of these frames are calculated and assigned as minimum threshold level and maximum zero crossing level. Only those frames which have an energy greater than minimum threshold and zero crossing rate less than the maximum set level are considered for further processing.

The main objective of feature extraction is to extract characteristics from speech signal that are unique, discriminative, robust and computationally efficient. The endpoints of speech signal are detected, using these endpoints silence region is eliminated, then this speech signal is divided into ten equally spaced sections, average energy and zero crossing is calculated for each section so final result is a feature vector of 20 dimension for any given input utterance.

2.2 Vector Quantization

Vector quantization (VQ) [1] is an efficient technique for data compression and has been successfully used in various applications involving VQ-based encoding and VQ-based recognition.

VQ can be defined as a mapping function that maps k-dimensional vector space to a finite set CB = {C1, C2, C3, ..…., CN}. The set CB is called codebook consisting of N number of code vectors and each code vector Ci = {ci1, ci2, ci3, ……, cik} is of dimension k. The key to VQ is the good codebook. Codebook can be generated in spatial domain by clustering algorithms or using transform domain techniques. The method most commonly used to generate codebook is the Linde-Buzo-Gray (LBG) algorithm [10].

Speech recognition is basically divided into two parts, namely features training and features Matching/testing. Features training is a process of enrolling or registering a new speech sample of a distinct word to the identification system database by constructing a model of the word based on the features extracted from the word's speech samples. Feature training is mainly concerned with randomly selecting feature vectors of the recorded speech samples and performs training for the codebook using the LBG vector quantization (VQ) algorithm.

Feature matching/testing is a process of computing a matching score, which is the measure of Similarity of the features extracted from the unknown word and the stored word models in the Database. The unknown word is identified by having the minimum matching score in the database. For generating the codebooks, the LBG algorithm [3, 4] is used. This is also called as Generalized Lloyd Algorithm (GLA). The LBG algorithm steps are as follows,

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors.

2. Double the size of the codebook by splitting each current codebook yn according to the rule

$yn^{+} = yn\ (1+\varepsilon)$

$yn\ ^{+} = yn\ (1-\varepsilon)$

Where n varies from 1 to the current size of the codebook, and ε is a splitting parameter. Where ε is usually in the range of (0.01 ≤ ε ≤ 0.05).

3. Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).

4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold.

6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed.

Initial codebook is to serve as a starting codebook for training each selected feature vector against one another. For the squared distortion and for each speech vector, the nearest code vector in the current codebook is calculated using Euclidean distance. Then the vector is assigned to that nearest code vector. Update the centroid in each cell using the centroid of the training vectors assigned to that cell by taking the average of the speech vector in a cell to find the new value of the code vector.

2.3 Euclidean Distance Measure

Euclidean distance measure is applied in order to measure the similarity or the dissimilarity between two spoken words, which take place after quantizing a spoken word into its codebook.

The matching of an unknown word is performed by measuring the Euclidean distance between the features vector of the unknown word to the model (codebook) of the known words in the Database. The word with the smallest average minimum distance is picked as shown in the equation below

$$d(x,y) = \sqrt{\sum_{i=1}^{D} (xi - yi)}^{\,2} \qquad (1)$$

where xi is the ith input features vector, yi is the ith features vector in the codebook, d is the distance between xi and yi.

3. DATABASE DEVELOPMENT AND RECOGNITION

Table 1. Database Description

Parameter	Sample characteristics
Language	English
No. of Speakers	08 (4 Male, 4 Female)
Sampling frequency, quantization	16000 Hz, 16 bits
Average duration of training and testing utterance	1 – 2 sec
Total number of words	10
Number of sample utterances per word	20
Total number of utterances in database.	10*20 = 200.

Database of speech samples is prepared; the speech samples used in this project are recorded using Windows Sound Recorder 2010, 9.0.1. Table 1 shows the database description. The samples are collected from eight different speakers, so that speaker independent speech recognition can be done. The performance of speech is evaluated in terms of recognition rate, the following recognition measure for computing the recognition rate,

$$\text{Recognition Rate} = \frac{\text{No of Successful detection of word}}{\text{No of words in testing set}} \qquad (2)$$

4. RESULTS

Database used for training and testing is recorded. For each word, twenty utterances from different speakers are collected, samples are taken from each speaker in two sessions so that training model and testing data can be created. Ten samples from each word are used for training phase and remaining ten is used for testing phase.

The feature vector of all reference speech samples are calculated in the training phase, in the matching phase, the test sample that is to be identified is taken and similarly processed as in training phase to form feature vector. The stored feature vector which gives the minimum Euclidean distance with the input sample feature vector is declared as word identified. Ten computer related words are taken like *Go, Lift, Come, Up, Down, Start, Stop, Show, Throw and Turnleft* and performance is calculated in terms of recognition rate using Eq.(2).

Table 2. Recognition Performance for VQ codebook size of 64

Input Word	Recognition Rate
GO	78.70%
LIFT	69.9%
COME	80.48%
UP	73.79%
DOWN	79.36%
START	72.52%
STOP	87.66%
SHOW	80.82%
THROW	87.5%
TURNLEFT	66.37%

As shown in Table 2. different recognition rate is obtained for all ten words, due to presence of plosives at the beginning or end of some of the words, as in, ‘Start’ and ‘Up’ (ends in a plosive), which are misinterpreted as more than one word. Further the words ‘Go’, ‘Throw’ and ‘Show’ have the same vowel part and differ only in their unvoiced beginnings and endings. Similarly the word ‘Lift’ and ‘Turnleft’ have similar endings, so these two words are mostly misinterpreted with each other.

Different tests were conducted and analysis was carried out with a different codebook size of VQ and the results obtained are shown in Figure 2. Maximum recognition rate is obtained for code book size of 4 as 85%, recognition rate drops as codebook size is increased above 4, but there is a slight variation in recognition rate for 32, 64 and 128, also for less than codebook size of 4, recognition rate decreases.

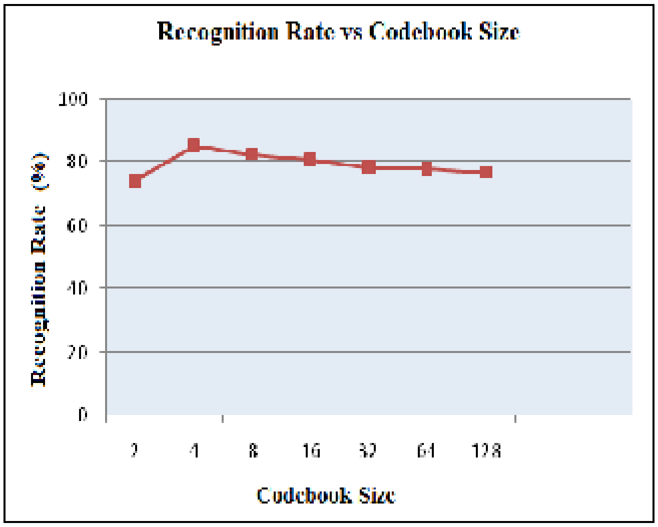

Figure 2. Variation in Recognition rate with different VQ codebook size

5. CONCLUSION [2]

A speaker independent isolated word recognition system for English language was implemented. The results were found to[4] be satisfactory for a vocabulary of English words. The accuracy of a system can be increased significantly by using an improved speech detection/noise elimination algorithm. Further improvement can be obtained by a better VQ codebook design with the training set including utterances from a large number of speakers with variation in ages and accents. As we can see from results different recognition rate was obtained for all ten words, because of different phonemes are used for different words, also[1] maximum recognition rate is obtained for codebook size of 4.

The accuracy of the identification process can be influenced by certain factors such as different level of surrounding noise during the recording session, the quality of the microphone used to input the speech signals, and many other factors. Even though it is difficult to avoid some of these factors, steps should be taken to minimize the effect.

6. REFERENCES

[1] H. B. Kekre, Tanuja K. Sarode, "Speech Data Compression using Vector Quantization", WASET International Journal of Computer and Information Science and Engineering (IJCISE), Fall 2008, Volume 2, Number 4, pp.: 251-254, 2008.

[2] H.B.Kekre, Ms Vaishali Kulkarni,'Speaker Identification by using Vector Quantization', International Journal of Engineering Science and Technology, Vol. 2(5), 2010, 1325-1331.

[3] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J.Prentice-Hall, 1993.

[4] Mohamed Debyeche, Jean-Paul Haton & Amrane Houacine, 'Improved Vector Quantization Approach for Discrete HMM Speech Recognition System', International Arab Journal of Information Technology – 2007.

[5] Navnath S. Nehe, Raghunath S. Holambe,' Isolated Word Recognition using Normalized Teager Energy Cepstral Features', Proceedings IEEE , International Conference on Advances in Computing, Control, and Telecommunication Technologies,2009, pp. 106-110.

[6] Poonam Bansal, Amita Dev, Shail Bala Jain, 'Optimum HMM Combined with Vector Quantization for Hindi Speech Word Recognition', Proceedings of IETE Journal of Research, vol-54, issue-4, July-Aug-2008.

[7] Poonam Bansal, Amita Dev, Shail Bala Jain,' Enhanced Feature Vector Set For VQ Recoginizer In Isolated Word Recoginition',Proceedings of International Conference Information Research & Applications, i.Tech-2007,Verna, Bulgeria , pp.390-395,June-2007.

[8] Rabiner, L. R. and Sambur, M. R., "An Algorithm for Determining the Endpoints of Isolated Utterances," The Bell System Technical Journal, Vol. 54, No. 2, February 1975.

[9] Shivesh Ranjan,' A Discrete Wavelet Transform Based Approach to Hindi Speech Recognition', Proceedings IEEE, International Conference on Signal Acquisition and Processing, 2010, pp. 345-348.

[10] Y. Linde, A. Buzo, and R. M. Gray.: 'An algorithm for vector quantizer design," IEEE Trans. Commun.' vol. COM-28, no. 1, pp. 84-95, 1980.