ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

A Novel Approach for Speech Recognition using Vector Quantization through LBG Algorithm

Manjot Kaur[1], Lakhvir Garcha[2]
M.Tech Student[1], Professor[2]
Department of Computer Science and Engineering
North West Institute of Engineering and Technology, India

Abstract:
Speech is one of the important requirements and the well-situated method of communication between people. Real time speech to text is an accurate conversion of uttered words after speaking.STT is very useful tool to interact with people in counseling interviews or conference. Thus the conversion of speech to written language must be accurate and fast so that it can be easily understand by people**.** The fundamental approach of this paper is to develop an algorithm to convert speech to text using Punjabi phonetics. This paper introduces and discusses two popular and different noise reduction techniques (Auto Spectral Subtraction, LBG and MFCC) and presents our simulation result of a noise reduction system. It is shown that the system reduces the noise almost completely by finding the minimum Euclidean distance and keeps the enhanced speech signal very similar to the original speech signal. This paper presents a method to design a speech to text conversion module using Java. This method is simple to use and take less use of memory space.

Keywords: Auto Spectral Subtraction, Linde-Buzo -Gray algorithm, Mel Frequency Cepstral Coefficient, Speech Recognition.

I. INTRODUCTION

Speech is the most important form of communication in everyday life in order to make the interaction easier and faster. Speech is the vocalized form of human communication and it is considered to be the primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech. Automatic Speech Recognition provides a path for natural communication between man and machine. A simple alternative to a hardware interface is a software interface i.e. a Speech to Text system. Speech to Text Conversion or Speech Recognition allows a computer to identify the words that a person speaks into a mike or any other similar hardware and convert it into written words. Basically, the mode of communication between humans takes place in several ways such as facial expressions, gestures, eye contact and speech. Speech to text conversion is very advantageous and used in various applications areas. Human interact with each other in several ways such as facial expression, eye contact, gesture, mainly speech. The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech. The recognition of speech is one the most challenges in speech processing. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means of Algorithm implemented as a computer program. Speech to text conversion (STT) system is distinguished into two types, such as speaker dependent and speaker independent systems. The main difficulties in implementation of an ASR system are due to different speaking styles of human beings and environmental disturbances. So the main aim of an ASR system is to transform a speech signal into text message independent of the device, speaker or the surroundings in an accurate and efficient manner. On the basis of way to recognize speech recognition may be isolated word recognizes utterance to have quiet on both sides of sample windows i.e. only one word at a time and word is preceded and followed by silence. This is having "Listen and Non Listen state". Connected word system are same as isolated words but recognizes speech having one or more than one word and these words are divided or separated by small sound to be "run together minimum pause between them. Continuous speech recognizers allows user to talk almost naturally. Thus system recognizes more than one word and words are connected without any silence. Spontaneous Word speech Recognition recognizes speech that is natural sounding and not be rehearsed. An ASR System with impulsive speech should be able to handle a dissimilar words and mixture of natural speech feature such as words being run together like ums, ahs and others. Motive of Spontaneous Word Speech Recognition is to recognize natural speech. This paper gives a description of implementation of Speech to Text Conversion System using Auto Spectral technique. The system goes through different steps to accomplish the task of speech to text conversion that are signal preparation, acoustical analysis, training and testing. For the purpose of improving accuracy of the system, the system uses a noise reduction technique named Auto Spectral Subtraction. Auto Spectral Subtraction [8] is a simple and efficient noise reduction technique. In this technique, an average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved. The algorithm for the design of optimal VQ is commonly referred to as the Linde-Buzo-Gray (LBG) algorithm and it is based on minimization of the squared-error measure.

II. VECTOR QUANTIZATION

The main objective of data compression is to reduce the bit rate for transmission or data storage while maintaining the necessary fidelity of the data. The feature vector may represent a number of different possible speech coding parameters including linear predictive coding (LPC) coefficients, cepstral

coefficients. The VQ can be consideredas a generalization of scalar quantization to the quantization of a vector. The VQ encoder encodes a given set of k-dimensional data vectors with a much smaller subset. The subset *C* is called a codebook and its elements $C_i$arecalled codewords, codevectors, reproducing vectors, prototypes or design samples. Only the index *i* is transmitted to the decoder. The decoder has the same codebook as the encoder, and decoding is operated by table look-up procedure. The commonly used vector quantizers are based on nearest neighbour called Voronoi or nearest neighbour vector quantizer. Both the classical K-means algorithm and the LBG algorithm belong to the class of nearest neighbour quantizers.

A key component of pattern matching is the measurement of dissimilarity between two feature vectors. The measurement of dissimilarity satisfies three metric properties such as Positive definiteness property, Symmetry property and Triangular inequality property. Each metric has three main characteristics such as computational complexity, analytical tractability and feature evaluation reliability. The metrics used in speech processing are derived from the Minkowski metric [J. S. Pan et al. 1996]. The Minkowski metric can be expressed as

$$D_p(X,Y) = \sqrt[p]{\sum_{i=1}^{k} |x^i - y^i|^p},$$

Where $X = \{x^1, x^2, ..., x^k\}$ and $Y = \{y^1, y^2, ..., y^k\}$

are vectors and p is the order of the metric.

The City block metric, Euclidean metric and Manhattan metric are the special cases of Minkowski metric. These metrics are very essential in the distortion measure computation functions.

The distortion measure is one which satisfies only the positive definiteness property of the measurement of dissimilarity. There were many kinds of distortion measures including Euclidean distance, the Itakura distortion measure and the likelihood distortion measure, andso on. The Euclidean metric [Tzu-Chuen Lu et al., 2010] is commonly used because it fits the physical meaning of distance or distortion. In some applications division calculations are not required. To avoid calculating the divisions, the squared Euclidean metric is employed instead of the Euclidean metric in pattern matching.

The quadratic metric [Marcel R. Ackermann et al., 2010] is an important generalization of the Euclidean metric. The weighted cepstral distortion measure is a kind of quadratec metric. The weighted cepstral distortion key feature is that it equalizes the importance in each dimension of cepstrum coefficients. In the speech recognition, the weighted cepstral distortion can be used to equalize the performance of the recognizer across different talkers. The Itakura-Saito distortion [Arindam Banerjee et al., 2005] measure computes a distortion between two input vectors by using their spectral densities. The performance of the vector quantizer can be evaluated by a distortion measure D which is a non-negative cost $D(X_j,X_j)$ associated with quantizing any input vector *j X* with a reproduction vector *j X* ˆ . Usually, the Euclidean distortion measure is used. The performance of a quantizer is always qualified by an average distortion $D_v=E[D(X_j,X_j)]$ between the input vectors and the final reproduction vectors, where E represents the expectation operator. Normally, the performance of the quantizer will be good if the average distortion is small.

Another important factor in VQ is the codeword search problem. As the vector dimension increases accordingly the search complexity increases exponentially, this is a major limitation of VQ codeword search. It limits the fidelity of coding for real time transmission. A full search algorithm is applied in VQ encoding and recognition. It is a time consuming process when the codebook size is large.

III. LINDE–BUZO–GRAY (LBG) ALGORITHM

The LBG algorithm is also known as the Generalised Lloyd algorithm (GLA). It is an easy and rapid algorithm used as an iterative nonvariational technique for designing the scalar quantizer. It is a vector quantization algorithm to derive a good codebook by finding the centroids of partitioned sets and the minimum distortion partitions. In LBG, the initial centroids are generated from all of the training data by applying the splitting procedure. All the training vectors are incorporated to the training procedure at each iteration. The GLA algorithm is applied to generate the centroids and the centroids cannot change with time. The GLA algorithm starts from one cluster and then separates this cluster to two clusters, four clusters, and so on until N clusters are generated, where N is the desired number ofclusters or codebook size. Therefore, the GLA algorithm is a divisive clustering approach. The classification at eachstage uses the full-search algorithm to find the nearest centroid to each vector. The LBG is a local optimization procedure and solved through various approaches such as directed search binary-splitting, mean-distance-ordered partial codebook search [Linde et al., 1980, Modha et al., 2003], enhance LBG, GA-based algorithm [Tzu-Chuen Lu et al., 2010, Chin-Chen Chang et al. 2006], evolution-based tabu search approach [Shih-Ming Pan et al., 2007], and codebook generation algorithm [Buzo et al., 1980]. In speech processing, vector quantization is used for instance of bit stream reduction in coding or in the tasks based on HMM. Initialization is an important step in the codebook estimation. Two approaches used for initialization are Random initialization, where L vectors are randomly chosen from the training vector set and Initialization from a smaller coding book by splitting the chosen vectors.

IV. WORKFLOW

Speech Recognition undergoes various steps. There are usually two phases namely preprocessing and post processing. In preprocessing phase, speech is given as input signal. This signal is mapped in digital form i.e. sampled and quantized. Sampling frequency of speech varies from one format to another. After capturing audio signal, framing is performed followed by windowing. After this, feature extraction techniques are applied based on which features are being extracted. Usually, for Punjabi speech, MFCC and LPC features are being extracted. With this, different models are being trained. Recently work is being carried out in HMM, DTW and Neural Networks. After being trained, the models are set as persistent data. Also, we map the speech signals onto dictionary in order to obtain its grammar.

However, in post- processing phase, speech signal is given as testing data. In this, we extract feature vectors and further map it with the trained model. Further, we calculate distance between the testing file feature vectors and most resembling word vectors found in modeled file. This is done by using dictionary and grammar framed during preprocessing phase. If the distance is more than threshold, it is rejected else it isaccepted. Finally, performance is calculated on the basis of percentage of correct observations as follows:

WER=(S+D+I)/N Where S means number of substitutions; Dindicates number of deletions; I identifies number of insertions and N corresponds to number of words of reference.

V. PERFORMANCE PARAMETERS

Accuracy and Speed are the criterion for measuring the performance of an automatic speech recognition system which are described below:

1. Accuracy Parameters

Word Error Rate (WER): The WER is calculated by comparing the test set to the computer-generated document and then counting the number of substitutions (S), deletions (D), and insertions (I) and dividing by the total number of words in the test set.

Formula WER=(S+D+I)/N

Word Recognition Rate **(**WRR): It is another parameter for determining accuracy.

Formula: WRR=1-WER

e. g. REF: Misunderstandings | usually | develop. S1: Misunderstandings | using | develop. The substitution error of the word *using* for the word *usually* would be scored as one substitution error, as opposed to one error of deletion (*usually*) and one error of insertion (*using*). *Single Word Error Rate* (SWER) and *Command Success Rate* (CSR) are two more parameters to determine accuracy of a speech recognition system.

2. Speed Parameter

Real Time Factor is parameter to evaluate speed of automatic speech recognition.

Formula: RTF = P/I

where P: Time taken to process an input Duration of input I e. g. RTF= 3 when it takes 6 hours of computation time to process a recording of duration 2 hours. RTF ≤ 1 implies real time processing.

D. Performance Degradation

Automatic speech recognition suffers degradation in recognition performance due to following inevitable factors:

i. Prosodic and phonetic context
ii. Speaking behaviour
iii. Accent & Dialect
iv. Transducer variability and distortions
v. Adverse speaking conditions
vi. Pronunciation
vii.Transmission channel variability and distortions
viii. Noisy acoustic environment
ix. Vocabulary Size and domain

VI. COMPARISON OF EXISTING APPROACHES

OCR	**NLP**	**SR**	**PR**	**DTW**	**HMM**
Optical Character Recognition is conversion software used to convert and extract text from image	Natural language processing (NLP) is a branch of computer science which deals which the human (natural) language with computer	Speech recognition (SR) is to translate an audio or voice into the text.	Pattern recognition is a branch of pattern recognition for machine learning.	Dynamic time warping (DTW) is an algorithm for used for pattern matching between the two who may vary in speed or time	A hidden Markov model (HMM) is a model which is to use to study the hidden or unobserved states.
It is mainly used in the passport documents, invoices, bank statements, computerized receipts and business cards.	It is used widely in the development of Artificial intelligence and is the key feature to solve the central artificial intelligence problem	It is widely used in the car systems, therapeutic use , military programs, education, home automation etc	It is used in license plate recognition, fingerprint analysis and face detection/verification	DTW has been applied to temporal sequences of video, audio, and graphics data.	HMM is widely used in the application areas of speech, handwriting, gesture, POS tagging
Its method is used in digitized printed texts so that it can be electronically edited, searched, stored more compactly	Its method is used of converting human language into another, like machine understandable language etc.	Its method is used in the telephony has shown benefits to short-term-memory re-strengthening in brain	Its method is commonly used in face recognition, to check spam or non spam email messages	The most common applications include speaker recognition and online signature recognition. Also it can be used in partial shape matching application	HMM provide a framework for modeling using mathematical computations.
OCR is a field of research area of AI and computer vision	It enables computers to derive to analyze all human language	It enables the humans to give verbal command to a machine to perform the particular task instead on controlling it manually.	PR main aim is to give useful and valid result on all possible likely matching inputs	Used in shape analysis and geometric shapes use in computers.	HMM pair is widely used in finding pair wise alignment of protein and DNA

VII. CONCLUSION

This Speech- to-Text conversion system is implemented by using the MFCC for feature extraction. In audio folder, 100 audio files are recorded and these are analyzed to get feature vectors. These features are initially modeling in the vector quantization. After that, the test spoken word is addressed by linde-buzo-gray algorithm. In this work, the performance of the system is more accurate and reliable by using linde-buzo-gray algorithm in preprocessing stage.

VIII. REFERENCES

[1]. SantoshK.Gaikwad, „A review on speech recognition techniques", International Journal of Computer Applications, Volume 10– No.3, November 2010

[2]. NishantAllawadi, "Speech-to-Text System for Phonebook Automation", Computer Science And Engineering Department Thapar University, June 2012.

[3]. SanjivaniS.Bhabad, „An overview of technical progress in speech recognition", International Journal of advanced research in computer science and software Engineering, Volume 3, Issuse 3, March 2013

[4]. Su Myat Mon, HlaMyoTun, „Speech-To-Text(STT) System Using Hidden Markov Model(HMM)", International Journal Of Scientific & Technology Research Volume 4,Issue 06, June 2015.

[5]. S. Furui [1986], "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustic, Speech, Signal Processing, Vol. 34, No. 1, pp. 52-59.