OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

A Review of Techniques used in spoken-word Recognition System

Kamalvir Kaur Pannu[1], Neelu Jain[2]

[1]*ME Student, Department of Electronics and Communication Engineering, PEC University of Technology, Chandigarh, India*

[2]*Associate Professor, Department of Electronics and Communication Engineering, PEC University of Technology, Chandigarh, India*

ABSTRACT: *Humans feel most comfortable to interact among themselves via speech, rather than any other medium such as gestures, text. Computers have become an important part of the world today and we humans need to interact with it in our daily lives. But we do not feel at similar ease of interaction with computers as with our fellow beings. Human-computer interaction via speech is the solution to this impediment, hence an active area of research. Spoken word recognition system enables a computer to understand the words spoken by converting them in text form. This paper introduces the reader to the process and techniques of isolated word recognition systems. It gives a succinct review of techniques used during various stages of spoken word recognition system. Utility and Advantages of each technique are discussed briefly. Also, the generalized workflow followed for designing such a system is presented effectively.*

Keywords: *Isolated Word Recognition, Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Artificial Neural Network (ANN), Dynamic Time Warping(DTW)*

I. INTRODUCTION

A spoken word recognition system enables a computer to identify the words spoken by the user and convert them in textual form. It is accomplished by means of an implemented algorithm. The system tries to imitate human auditory system by understanding voice input to them. Such recognition systems have been developed in various languages such as English, French, Japanese, Chinese, Arabic, Hindi etc. But a fully efficient and accurate recognition system is still far from reach due to background noise and variability of dialects of different speakers [1]. Speech Recognition encompasses two main phases: Testing Phase and Training Phase. Training involves teaching the system by building its dictionary, an acoustic model (template) for each word that the system has to recognize, Testing phase compares the spoken word with all the templates available in dictionary and tries to find a close match with least mismatch[2].Feature Extraction and Classification are two major components of a speech recognition system. Both testing and training phases involve feature extraction and classification techniques. Feature Extraction is where speech signal is converted into a set of parameters called features. MFCC, LPC, PCA, DWT, PLP techniques are popularly used for feature extraction. Classification is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal.ANN, VQ, HMM, DTW are some of the popular classification techniques used. A spoken word recognition system is classified in two ways. If the system can recognize the word spoken by any speaker, then it is speaker independent system. But if its recognition is limited to a specific speaker, it is speaker dependant system. Such a system differentiates between the same word spoken by different speakers. Also the recognition method can be word based or phoneme based. Phonemes are sub parts of a word and a recognition system can be trained with either phonemes or complete words [3].Speech recognition systems have many potential applications including command and control, dictation, making transcripts and searching audio documents. Improving recognition accuracy, reducing memory requirement and improving speed are the main challenges faced by researchers in this field.

II. RELATED WORKS

Y. X. Zou et-al had taken into account MFCC, Linear Predictive Coding (LPC), the hybrid PCA-MFCC, and Linear Predictive Cepstral Coding (LPCC)[4]. Purva Kulkarni et-al proposed a modified MFCC technique in which DFT calculation step in MFCC is replaced by Wavelet packet transform. Support Vector Machine (SVM) was used as a classifier here[5]. S.D. Dhingra et-al described an approach of isolated speech recognition by using MFCC and DTW [6]. I.M. El-Henawy et-al carried out recognition of phonetic Arabic figures via Wavelet based Mel Frequency Cepstrum using HMM. Four Speech Recognition techniques were used for feature extraction i.e. MFCC, Short Time Energy, Cepstrum and LPC [7]. C. Ittichaichareon et-al has

used along with MFCC, PCA as the supplement in feature dimensional reduction prior to training and testing speech samples via Maximum Likelihood Classifier and SVM [8]. Jing Bai et-al proposed an anti noise speech recognition system based on improved MFCC features and wavelet kernel SVM.[9] Ch. Ramaiah and Dr. Srinivasa Rao developed Speech Recognition System by using MFCC and Vector Quantization(VQ)[10]. S.Sunny et-al developed a speech recognition system for recognizing isolated words of Malayalam. Two wavelet techniques DWT and Wavelet Packet Decomposition (WPD) were used for extracting the features. The performances of these systems were tested using the SVM classifier [11]. N.S. Nehe and R.S. Holambe proposed Wavelet Decomposition and reduced order LPC Coefficients. The proposed method provided effective, efficient and practical features [12]. S. Sunny et-al had done a comparative study on LPC and WPD for recognizing spoken words in Malayalam. Back propagation method was used to train ANN. WPD was found to be more suitable for Malayalam [13]. S.Ranjan had proposed a new scheme in which LPC is performed over DWT coefficients for Hindi Speech Recognition. K Means Algorithm was used to train HMM [14]. Hong-yanLi et-al proposed a Speech Enhancement Algorithm Based on Independent Component Analysis (ICA). Simulation result shows that much better Denoising effect and signal-noise ration can be obtained by using the proposed algorithm [15].

III. FEATURE EXTRACTION TECHNIQUES

PURPOSE OF FEATURE EXTRACTION

When a speech signal is digitized (automatically when stored in computer), a large number of sample values are obtained. Such a large number is difficult to manage and requires large memory space**.** Features are extracted by taking these sample values as input and "summarizing" these values in small set.

For Speech Recognition, some of its characteristics in time/frequency or in some other domain must be known. So a basic requirement of a speech recognition system will be able to extract a set of features for each of the basic units. Features are just a more efficient and compact representation of sample values present in the signal**.**

The feature vector extracted should possess the following properties [16]:

- Consistent over a long period of time
- Can be easily measured from the input speech samples
- Should be small in dimension
- Should be insensitive to the irrelevant variation in the speech
- Should not have correlation with other features

Selection of feature extraction technique plays an important role in recognition accuracy. Some of them are explained below

3.2 Principal Component Analysis (PCA)

PCA is mainly used for dimensionality reduction in speech recognition area. It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables [8], hence achieving reduction in number of features.

3.3 Linear Predictive Coding (LPC)

LPC is one of the most powerful speech analysis techniques and is a useful method for encoding quality speech at a low bit rate. The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples[19].The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech signal and the estimated speech signal over a finite duration.

3.4 Relative Spectra Filtering Of Log Domain Coefficients (Rasta)

RASTA filtering technique compensates for linear channel distortions. Linear channel distortions appear as an additive constant in both the log spectral and the cepstral domains [19]. The high-pass portion of the equivalent band pass filter alleviates the effect of convolutional noise introduced in the channel. The low-pass filtering helps in smoothing frame to frame spectral changes.

3.5 Perceptual Linear Prediction (PLP)

PLP is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction modifies the short-term spectrum of the speech by several psychophysically

based transformations [19]. The goal of the original PLP model is to describe the psychophysics of human hearing more accurately in the feature extraction process.

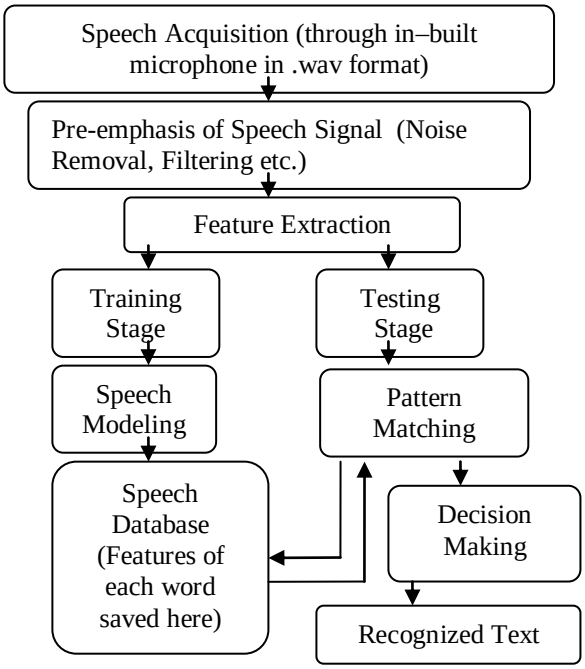

Fig.1: Generalized Methodology of Word Recognition Systems

3.6 Mel Frequency Cepstral Coefficients (MFCC)

MFCC Technique enables the word identification to be done in the same way as is done by natural human hearing mechanism. Our hearing and recognition mechanism does not respond to all frequencies equally. So frequency response is a non linear function of frequency. Now the question arises as to how MFCC technique models the natural recognition system. The answer is Mel scale. On Mel scale, frequencies below 1000 Hz are linearly spaced while frequencies above 1000 Hz are spaced logarithmically [5][20].

3.7 Discrete Wavelet Transform (DWT)

Wavelets have the ability to analyze different parts of a signal at different scales. This technique replaces the fixed bandwidth of Fourier Transform with one proportional to frequency which allows better time resolution at higher frequencies than Fourier Transform. This technique has Multiresolution Capability [9]. It provides high frequency resolution at low frequencies and high time resolution at high frequencies. Number of features obtained is much less that obtained in MFCC. So memory required for implementation of this technique is also less.

3.8 Wavelet Packet Decomposition (WPD)

WPD divides the speech signal into both high frequency and low frequency components. The approximation coefficients contain the useful information, but total elimination of high frequency components may result in loss of information. Discrete Wavelet Transform does not address this issue. This technique also replaces the fixed bandwidth of Fourier Transform with one proportional to frequency which allows better time resolution at higher frequencies than Fourier Transform like Discrete Wavelet Transform.

IV. FEATURE CLASSIFICATION TECHNIQUES

ANN, HMM, DTW and VQ are popularly used techniques for feature classification

4.1 Artificial Neural Networks (ANN)

ANNs are popularly used for classification due to their error stability and self organizing ability. They are based on principle of generalized learning, self adjustment and fault tolerance. The inputs presented to the network are the features derived from each spoken word. This input moves through the weights and non linear activation functions. At the output layer, and the error is calculated with help of targets. Targets are the words in text form. Then error is correlated in a backward direction and weights are adjusted accordingly. Neural Network provides three types of learning methods namely supervised, unsupervised and reinforced.

4.2 Hidden Markov Model (HMM)

HMMs statistically model the words. During training of the HMM, it encodes the observation sequence in such a way that if a observation sequence (e.g. Sequence of phonemes in a word) having many characteristics similar to the given one be encountered later (during testing), it should be able to identify it. K-Means Algorithm and Baum-Welch algorithm are popularly used for HMMs [21].

4.3 Dynamic Time Warping (DTW)

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed [4]. In isolated word recognition, the template for a word is a set of features derived for single utterance of that word. Template matching is achieved by pair-wise comparison of feature vectors. But, within a word, there could be a variation in the length of individual phonemes. The DTW matching process compensates for such length differences .The DTW algorithm finds an optimal match between two sequences of feature vectors which allows for stretched and compressed sections of the sequence [22].

4.4 Vector Quantization (VQ)

Feature Vectors can be mapped to discrete vectors by quantizing them. Vector quantization is a process of mapping feature vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a code word. The collection of all code words is called a code book. During Testing, the features derived from spoken words are used to assign the word to a particular cluster by finding minimum distance from each centroid.

V. CONCLUSION

In speech recognition domain new feature extraction methods are being developed using combination of traditional techniques (MFCC, PCA, PLP, Wavelet Decomposition etc.) in Speech Recognition Systems. Authors have claimed improvement in performance by using these hybrid techniques. Need is to optimize the combination of these existing feature extraction techniques to maximize the Recognition Accuracy, minimize the Training Time and Recognition Time and to increase the Noise Robustness of the features derived. Though MFCC technique is used in most of the works, but still more research is needed in reducing its dimensionality by exploring its combination with other techniques. To be able to use these systems in real world, more noise robust features need to be worked upon. Below is the utility/advantages of all the discussed techniques.

Table1: Popular Feature Extraction Techniques and their Utility

Technique	**Utility**
PCA	Good for Gaussian Data, used for dimensionality reduction[17]
LPC	Can better distinguish words having distinct vowel sounds
RASTA	Optimal for Noisy Speech [17]
PLP	Takes into account the effect of additive noise
MFCC	Good technique to be used for clean speech[5][18]
WPD, DWT	Wavelet Denoising, Can model the details of unvoiced sounds well

Table 2: Feature Classification Techniques and their advantages

Classification Technique	**Advantages**
ANN	Adaptable to new environments and self organizing ability
HMM	Length of input can be variable and simple to design
DTW	Compensates for length differences of phonemes within the words.
VQ	Very simple, easily modified and effective classification technique

REFERENCES

[1]. Hazrat Ali, Xianwei Zhou, Sun Tie, *Comparison of MFCC and DWT Features for Speech recognition of Urdu*, International Conference on Cyberspace Technology, Beijing, China, 23-23 Nov. 2013, pp 154-158

[2]. Nitin Trivedi, Dr. Vikesh Kumar , Saurabh Singh, *Speech Recognition by Wavelet Analysis*, International Journal of Advanced Computing, Vol. 15, Issue 8, February 2011, pp 27-32

[3]. Chaitanya Joshi, Kedar Kulkarni, Sushant Gosavi, Prof. S.B. Dhonde, *Feature Extraction Using MFCC Algorithm*, International Journal of Engineering and Technical Research, Vol. 2 Issue 4 , April 2014, pp 74-77

[4]. Y. X. Zou, W. Q. Zheng and Wei Shi, Hong Liu ,*Improved Voice Activity Detection based on support vector machine with high separable speech feature vectors* , 19th International Conference on Digital Signal Processing (DSP), Hong Kong, 20-23 August 2014, pp 763-767

[5]. Purva Kulkarni, Saili Kulkarni, Sucheta Mulange, Aneri Dand, Alice N Cheeran, *Support Vector Machines for Isolated Word Recognition using Wavelet Packet Features*, International Journal of Engineering & Technology Research, Vol. 2, Issue-2, March-April, 2014, pp 31-37
[6]. S.D. Dhingra, Geeta Nijhawan, Poonam Pandit, *Isolated Speech Recognition Using MFCC and DTW*, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2 Issue 8, August 2013, pp 4085-4092
[7]. I.M. El-Henawy, W.I. Khedr, O.M.Elkomy, *Recognition of phonetic Arabic figures via wavelet based Mel Frequency Cepstrum using HMMs*, HBRC(Housing and Building National Research Center) Journal, ELSEVIER, Sep 2013, pp 49-54
[8]. C. Ittichaichareon ,S. Suksri and T.Yingthawornsuk, *Speech Recognition using MFCC, International Conference on Computer graphics, Simulation and modeling*, Pattaya(Thailand), 28-29 July 2012, pp 135-138
[9]. Jing Bai, Peiyun Xue, Xueying Zhang and Lihong Yang, *Anti-noise Speech Recognition System based on Improved MFCC Features and Wavelet Kernel SVM*, Journal of Advances in Information Science and service Scientist, Vol. 4 Issue 23, Dec 2012, pp 599-607
[10]. Ch. Ramaiah, Dr. V. Srinivasa Rao, *Speech samples recognition based on MFCC and Vector Quantization*, International Journal of Computer Science & Engineering Technology, Vol. 1 Issue 2, December 2012, pp 1-7
[11]. Sonia Sunny, David Peter, K.P. Jacob , *Design of a Novel Hybrid Algorithm for Improved Speech Recognition with SVM classifier*, International Journal of Emerging Technology and Advanced Engineering, Vol. 3 Issue 6, June 2013, pp 249-254
[12]. Navnath S Nehe, Raghunath S Holambe, *DWT and LPC based feature extraction methods for isolated word recognition*, EURSASIP Journal on Audio ,Speech and Music Processing, Vol. 2012 Issue 1, January 2012, 2012:7
[13]. S. Sunny, D. Peter and K. P. Jacob, *Feature Extraction Methods based on LPC and WPD for recognizing spoken words in Malayalam*, International Conference on Advances in Computing and Communications, Chennai, India, 3-5 August 2012, pp 27
[14]. Shivesh Ranjan, *A Discrete Wavelet Transform based approach to Hindi Speech recognition*, Proceedings IEEE International Conference on Signal Acquisition and Processing, Singapore, 26-28 Feb 2011, pp 345-348
[15]. Hong-yanLi , Qing-hua Zhao, Guang-long Ren ,Bao-jin Xiao, *Speech Enhancement Algorithm Based on Independent Component Analysis*, Fifth International Conference on Natural Computation, Tianjin, 14-16 August 2009, pp 598-602
[16]. 16.Nitin Trivedi, Dr. Vikesh Kumar , Saurabh Singh, *Speech Recognition by Wavelet Analysis*, International Journal of Advanced Computing, Vol. 15, Issue 8, February 2011, pp 27-32
[17]. Chaitanya Joshi, Kedar Kulkarni, Sushant Gosavi, Prof. S.B. Dhonde, *Feature Extraction Using MFCC Algorithm*, International Journal of Engineering and Technical Research, Vol. 2 Issue 4 , April 2014, pp 74-77
[18]. T.B. Adam, M.S. Salam, T.S. Gunawan , *Wavelet Cepstral Coefficients for Isolated Speech Recognition*, TELOMNIKA(Telecommunication, Computing, Electronics and Control) Indonesian Journal of Electrical Engineering, Vol. 11 Issue 5, May 2013, pp 2731-2738
[19]. Urmila Shrawankar , Dr. Vilas Thakare, *Techniques For Feature Extraction In Speech Recognition System : A Comparative Study*, International Journal of Computer Applications In Engineering, Technology and Sciences, Vol. 1, Issue 1, May 2013, pp 412-418
[20]. Mahmoud I. Abdalla , Haitham M. Abobakr, Tamer S. Gaafar, *DWT and MFCC based Feature Extraction Method for Isolated Word recognition*, International Journal of Computer Applications, Vol. 69, Issue 20, May 2013, pp 21-26
[21]. Rakesh Dugad, U.B. Desai, *A Tutorial On Hidden Markov Models*, May 1996
[22]. Steve Cassidy, *A Tutorial on Speech Recognition*, Department of Computing, Macquarie University