ISOLATED-WORD AUTOMATIC SPEECH RECOGNITION (IWASR) TECHNIQUES AND ALGORITHMS

Dr. Charansing N.Kayte , Dr. V.P.Pawar , Miss. A.A.Kale

Abstract— This system receives speech inputs from users, analyzes the speech inputs, searches and matches the input speech with the pre-recorded and stored speeches in the trained database/codebook, and returns the matching result to the users [1]. Developing this system is meant to assist customers calling a university's telephone operator to respond to their enquiries in a fast and convenient way using their natural speech. Callers are assisted using their own speech inputs to select their language preference, faculty in a university and finally select the staff name they wish to contact [2]. To extract features from the speech signals the Mel-Frequency Cepstral Coefficients (MFCC) algorithm was applied. Subsequently, Vector Quantization (VQ) algorithm based on the principle of block coding was used for all feature vectors generated from the MFCC algorithm.

Index Terms— Mel-Frequency Cepstral Coefficients(MFCC) , Vector Quantization (VQ), Human-computer, interaction, speech recognition , application.

Introduction:

It has been noticed that the success of isolated-word automatic speech recognition systems requires a combination of various techniques and algorithms, each of which performs a specific task for achieving the main goal of the system. Therefore, a combination of related algorithms improves the accuracy or the recognition rate of such applications. Thus, this Paper presents the techniques and algorithms used for the development and implementation of the Isolated-Word Automatic Speech Recognition (IWASR) system. In fact, this Paper provides a step-by step MATLAB implementation of the features extraction, features classification and features matching processes used in developing the IWASR.

The Mel-Frequency Cepstral Coefficients (MFCC) algorithm as the main algorithm used for the features extraction of all the set of distinct words[3]. IWASR also implemented the Vector Quantization (VQ) algorithm for the features classification/matching and pattern recognition[5]. In addition, this Paper explains the Euclidean distance measure as the similarity or distortion measure which was implemented in the IWASR.

2.Speech Samples Collection (Speech Recording)

The first factor is the profile of the talkers/speakers. The IWASR had five different speakers out of whom their speech samples were collected. Those five speakers include three male and two female speakers belonging to different ages, genders and races. Table 1 summarizes the first factor in more details about the profiles of talkers/speakers.

Table 1

Summary of Talkers' Profiles

Range of Age	Gender	Races
Adults ranging from 22 to 40 years old	1. Three male speakers 2. Two female speakers	Two main categories: 1. Marathi 2. Non- Marathi

The second factor is the speaking conditions in which the speech samples were collected from, which basically refer to the environment of the recording stage. The IWASR speech samples collection was done in a noisy environment. The speech samples were recorded in the faculty's speech technology lab; however, that lab is not a quiet or noise proof lab, meaning that all the speech samples were interrupted with noise. The rationale behind collecting the speech samples from noisy environments is to represent a real world speech samples collection, because most speech recognition systems are meant to be used in different environments and spheres.

Therefore, collecting speech samples from noisy environments was purposely done.

The third factor is the transducers and transmission systems. Speech samples were recorded and collected using a normal microphone. The fourth factor is speech units. The IWASR main speech units are specific isolated words. In other words, the purpose of IWASR is to recognize words that belong to isolated word recognition category of applications. A set of nine distinct words were recorded, which are (“Ek”, ”Don”, ”Tin”, ”Chaar”, ”Paach”, ”Saha”, ”saat”, ”Aat”, ”Nau”). each of which was recorded separately. Therefore, the IWASR is an isolated word recognizer rather a continuous or conversational word recognizer.

The IWASR used a simple Matlab function for recording speech samples. However, this function requires defining certain parameters which are the sampling rate in hertz and the time length in seconds. Figure 4.1 shows the Matlab code used for recording the speech samples.

```
R_fs = 16000; % Sampling rate

R_samp_len = 2; % Recording time length in seconds

% Record Function “wavrecord” is a Matlab function used for recording speech signals

ai = wavrecord(R_samp_len*R_fs,R_fs, 'double');
```

Fig 1

3.Features Extraction Using MFCC Algorithm

This stage emphasizes on the MFCC computational process, as the main algorithm used in this research for features extraction (front-end) analysis. It is also concerned with applying the MFCC algorithm against all collected speech samples in order to produce a features vector out of each collected speech sample.

There are certain parameters that need to be defined earlier in order to execute the MFCC algorithm and produce its coefficients. Table 2 shows the parameters and their defined values that are used in the entire MFCC Matlab code.

Table 2

MFCC Parameters Definition

Parameter	Defined Value
Time Length (len)	2 seconds
Sampling Rate (R_fs)	16000 Hertz per second
Frame Size (N)	256
Overlap Size (M)	156
Number of Filters (nof)	40

4. Framing

In IWASR, framing is meant to frame the speech samples into segments small enough so that the speech segment shows quasi-stationary behavior. The length of each segment is 256 samples which is equivalent to [((256 / 16000) * 1000)] = 16 milliseconds. Figure 2 shows the Matlab code for performing the framing of speech samples, whereas Figure 3 shows a segmented speech signal of frame size equal to 256 samples.

```
s = data; % data generated from the Matlab function “wavread” of the speech signal

a(1:256,1:83)=0; % framing the signal with overlap

a(:,1)=s(1:256);

for j=2:83

a(:,j)=s((N-M)*j+1:(N-M)*j+256); ; % N is size of each frame and M is overlap size end;
```

Fig 2

Matlab Code for the Framing Stage of MFCC

- ***Dr.Charansing N. Kayte*** *received Ph.d degree from singaniya University Rajshthan Currently working as a assistant professor in Digital and Cyber Forensic, Govt. Institute of Forensic Science, Aurangabad in 2013.E-mail:charansing@yahoo.co.in*

- ***Dr. Vrushsen P. Pawar*** *received MS, Ph.D.(Computer) Degree from Dept .CS & IT, Dr. B. A. M. University & PDF from ES, University of Cambridge, UK. Also Received MCA (SMU), MBA (VMU) degrees respectively. He has received prestigious fellowship from DST, UGRF (UGC), Sakaal foundation, ES London, ABC (USA) etc. He has published 90 and more research papers in reputed national international Journals & conferences. He has recognized Ph. D Guide from University of Pune, S. R. T. M. University & Singhaniya University (India). He is senior IEEE member and other reputed society member. Currently working as a Associate Professor in CS Dept of SRTMU, Nanded. E-mail:vrushvijay@yahoo.co.in*

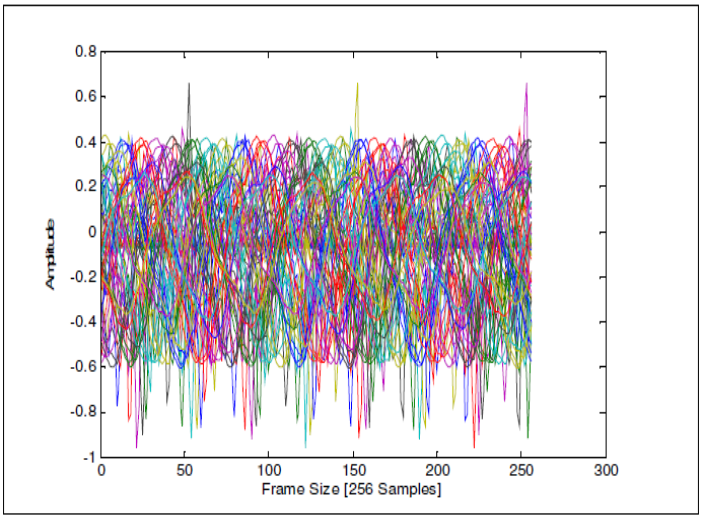

Fig 4

Segmented Speech Signal (Frame Size = 256 samples)

5.Windowing

In IWASR, this step is used to window the speech segment using the hamming window.Figure 5 shows the Matlab code for performing the windowing of the segmented speech samples, whereas Figure 6 shows the hamming window of the same segmented speech signal in Figure 4.

```
h= hamming(256); % windowing using hamming window

for j=1:83

b(:,j)= a(:,j).* h; % a is the segmented frame of size 256 samples and h is the

hamming window for the segmented frame
```

Fig 5

Matlab Code for the Windowing Stage of MFCC

The effect of windowing on the speech segment in Figure 4 can be seen clearly in Figure 7. There seems to be a smooth transition towards the edges of the frame.

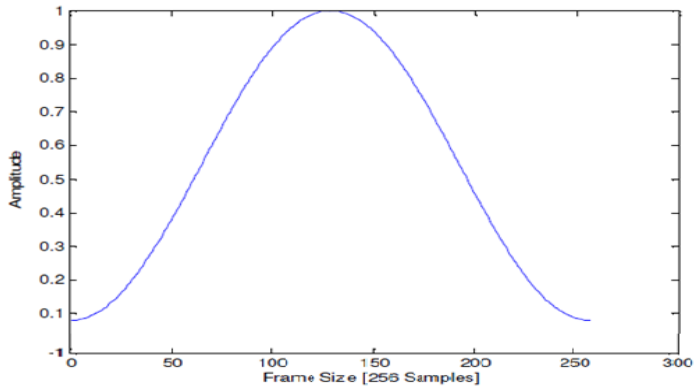

Fig 6

Hamming Window

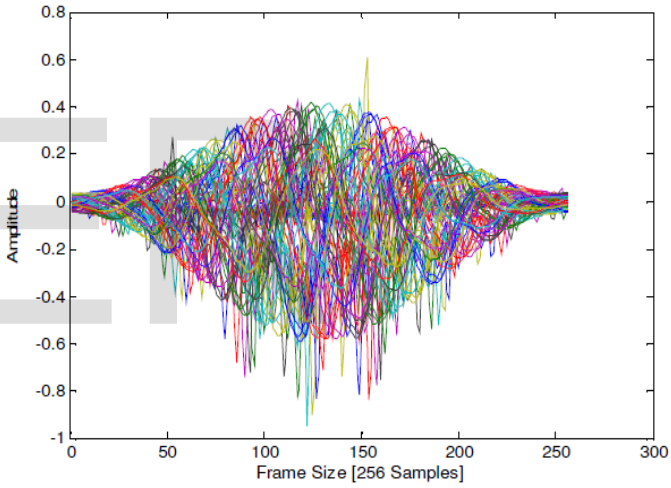

Fig 7

Windowed Speech Segment

6. Features Classification Using Vector Quantization (VQ) Algorithm

This step 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. Features training is mainly concerned with randomly selecting feature vectors of the recorded speech samples and perform training for the codebook using the LBG vector quantization (VQ) algorithm [5]. On the other hand, features 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.

Features Training Using Vector Quantization (VQ) Algorithm

Vector quantization (VQ) is the process of taking a large set of feature vectors and producing a smaller set of feature vectors that represent the centroids of the distribution, i.e. points spaced so as to minimize the average distance to every other point. Vector quantization has been used since it would be impractical to store every single feature that is generated from the speech utterance through MFCC algorithm.

7. Features Matching/Testing Using 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 goal is to find the codebook that has the minimum distance measurement in order to identify the unknown word[4],[5].

In IWASR, a simple Euclidean distance measure is applied on an unknown features vector compared against the trained codebook. Therefore, there must be an unknown speech signal and a trained codebook as inputs to this algorithm in order to measure their distance and test the entire performance of the IWASR[5]. The outputs of this algorithm are the ID numbers assigned for each features vector in the trained codebook as well as the distances or the squared error values. However, this algorithm picks up the ID number of the features vector which has the minimum distance to the unknown features vector. The most important purpose of performing this stage is to measure the accuracy/recognition level of IWASR in order to measure the validity of the algorithms used in this application.

8. Conclusion

This Paper has presented a detailed technical overview of MFCC and VQ, and how those two algorithms relate to each other. It was clearly mentioned that MFCC handles the features extraction process, which then produces outputs of speech feature vectors that are then considered as the training set used in the VQ algorithm to train the VQ codebook. Therefore, VQ works as a classification or pattern recognition technique that classifies different speech signals according to the classes. LBG VQ is the most commonly used VQ algorithm, which is divided into two phases. The first phase is the training, whereby a randomly selected speech signals form a training set of samples that are used as an initial codebook for training the VQ codebook. The second phase is the matching/testing that uses the Euclidean distance measure for comparing an unknown speech signal against the VQ codebook, which then selects the codeword in the codebook with the minimum distance. The combination of MFCC and VQ has been widely used in speaker recognition. Thus, this research studies the possibility of using this combination in telephony speech recognition systems.

It also emphasized on the fact that the IWASR is a standalone application categorized under the telephony applications of speech recognition.

References

[1] Kayte C.N. "Isolated Word Recognition for Marathi Language using VQ and HMM"., Science Research Reporter 2(2):161-165, April 2012

[2] Kayte C.N. " A Multi-HMM Marathi Isolated Word Recognizer", Science Research Reporter 2(2):175-177, April 2012

[3] Benoit Legrand, Nallasivam Palanisamy,"Chromosome classification Using Dynamic Time Warping", ScienceDirect Pattern Recognition Letters 29(2008) 215– 222.

[4] Cory Myers, Lawrence R. Rabiner, Aaron E. Rosenberg,"Performance Tradeoffs in Dynamic TimeWarping Algorithms for Isolated Word Recognition", Ieee Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp-28,No. 6, December 1980.

[5] Charansing N. Kayte, DrV.P.Pawar, Chandrashekhar

D. Sonawane(2012), "Human Computer Interaction Using Isolated-Words Speech Recognition System", Indian Streams Research Journal, Vol.1,Issue.V/June; 12pp.1-4.

[6] Charansing N. Kayte, DrV.P.Pawar, Chandrashekhar D. Sonawane "Isolated-Words Speech Recognition System Using Vq With MFCC", Vol.1,Issue.XII/June 2012pp.1-4.

IJSER