Arabic Speech Recognition System through VQLBG and Euclidean Distance Algorithms using Matlab

Mowaffak O. A. Albaraq
Dean faculty of Science and Engineering,
National University, Sana'a, Yemen
Professor of Alhajr (Qubitah) Community College, Yemen

ABSTRACT
The goal of this paper is to create an Arabic Speech Recognition System, and apply it to a speech of an unknown words. The system has been developed for introducing a unique technique making interaction of human with a computer for natural language processing. In this paper 100 Arabic samples were recorded through a microphone and MFCC features of speech sample were calculated, Vector Quantization for mapping large feature vectors to finite cluster codewords, build trained codebook model for each word and VQLBG with Euclidean Distances used for recognition word according to distortions associated with features. This system provides a high accuracy in case of Arabic speech words.

General Terms
A rectangular window, ASR System, Feature Extraction, Arabic word, NLP, DTW, FFT, Mel frequency scale and Mel power spectrum.

Keywords
Arabic Speech Recognition System, MFCC, Codebook, VQLBG and Euclidean Distances Algorithms.

1. INTRODUCTION
Arabic Speech Recognition System is useful in a large variety of applications banking business applications, postal zip code reading, security affairs, controlling machine, operate equipment, robotics conversation, communication, expert systems and data entry applications….etc. Arabic is a language spoken by Arabs in over to 22 countries, and roughly associated with the geographic region of the Middle East and North Africa as first language (Mother Tongue). It is also spoken as a second language by several Asian countries, (e.g. Iran, Indonesia, India, Malaysia, Pakistan, ..etc), in which Islam is the principle religion [1], [9].

Arabic is a Semitic language, and it is one of the oldest languages in the world. It is the fifth widely used language nowadays. Non-Semitic Languages such as Farsi, Urdu, Malay, and some West African languages such as Hausa have also adopted the Arabic alphabet for writing. Due to the cursive nature of the script, there are several characteristics that make recognition of Arabic distinct from the recognition of Latin scripts or Chinese [9].

Due to the cursive nature of the script, there are several characteristics that make recognition of Arabic distinct from the recognition of Latin scripts or Chinese. There are difficulties in this script [1], [2], [3], [4] ,[5] ,[6], [7], [8], **[9]**. Not much work has been done in Arabic Speech Recognition as compared to other languages.

The main problems in Arabic Speech Recognition as follows:

1- Standard Arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants.

2- Arabic has fewer vowels than English. It has three long and three short vowels, while American English has at least 12 vowels [1].

3- Arabic phonemes contain two distinctive classes, which are named pharyngeal(بلعومي) and emphatic phonemes. These two classes can be found only in Semitic languages like Hebrew, Urdu, Malay, and some West African languages such as Hausa, [9], [10], [11], [12], [16], [19] .

Table 1: Arabic Digits From(0 to 9) with Corresponding Words and English Meaning

Arabic Digits	0	1	2	3	4	5	6	7	8	9	10
Latin Digits	0	1	2	3	4	5	6	7	8	9	10
Arabic Words	صفر	واحد	اثنان	ثلاثة	اربعة	خمسة	ستة	سبعة	ثمانية	تسعة	عشرة
English Words	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten
In English	Safer	wahad	Ethnan	..	..	..					Ashrah

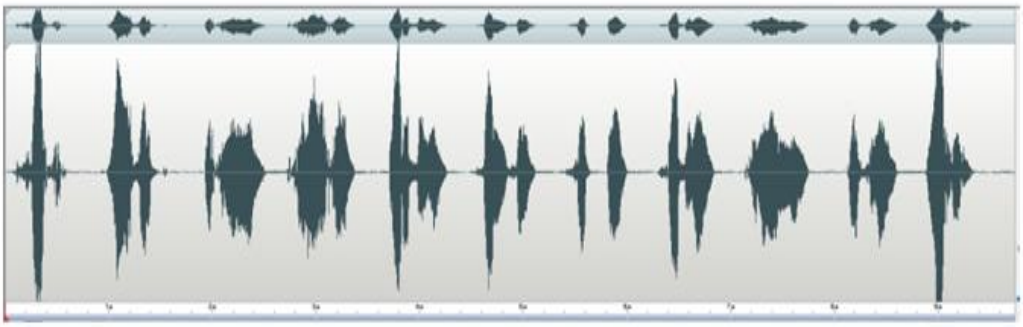

Figure 1: Arabic Speaker for Arabic Speech Words from (Safer to Ashrah) (Zero to Ten)

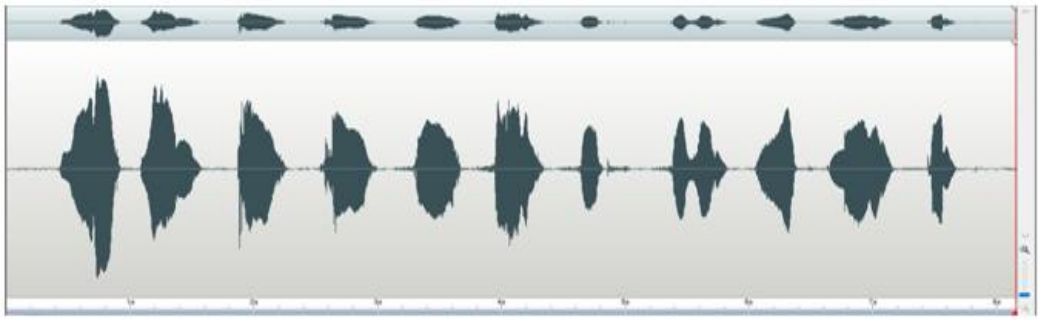

Figure 2: Arabic Speaker for English Words from (Zero to Ten)

4- The allowed syllables in Arabic language are: CV, CVC, and CVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant. All Arabic syllables must contain at least one vowel. Also Arabic vowels cannot be initials and can occur either between two consonants or final in a word. Arabic syllables can be classified as short or long [6], [7].

5- The CV type is a short one while all others are long. Syllables can also be classified as open or closed. An open syllable ends with a vowel, while a closed syllable ends with a consonant. For Arabic, a vowel always forms a syllable nucleus, and there are as many syllables in a word as vowels in it [12] .

6- Great difficulties occur when several speakers with different dialects are to be recognized. Homophone is a word that is pronounced the same as another word but differs in meaning. For example: The word (علم) mean flag, the word (علم) that mean understood and the word علم that means science. The word (ساعة) that means clock or time and the same word (ساعة) that means a day of the judgment [6], [7].

7- Arabic language is morphologically rich which causes a high vocabulary growth rate. This high growth rate is problematic for language models by causing a large number of out-of-vocabulary words [8], [9].

8- Arabic Language is cursive in general spoken and written from right to left. Arabic letters are normally connected to each other on the baseline and recognition must consider this aspect [17], [18], [19].

9- Isolate Arabic alphabet pronunciation is different from pronunciation the same alphabet connected in words [1], [2].

10- Arabic speech is more difficult than English speech as comparison in fig. 1, Fig. 2 above which appear from utterance Arabic energy and long which need more preprocessing for fixing and adapting for longer processing [3], [4], [5].

11- These problems can be minimized by restricting the number of speakers, words and working with good acoustic condition. Also, by avoiding the complexities of fluent speech and working on modern standard Arabic to overcome different dialects. Different approaches can be used in speech recognition such as HMM, ANN, SVM, GMM, Bayes, Fuzzy Logic, hybrid systems and Combined Classifiers [1], [10], [11], [20].

2. SYSTEM OVERVIEW

Speech recognition is an important application of Natural Language Processing (NLP). Speech is the most important part of communication. The study expressed important ideas through a specific language. Speech or word by word recognition is the process of automatically extracting and determining linguistic information conveyed by a speech wave using computers. Linguistic information, the most important information in a speech wave, is called phonetic information. The term speech recognition means the recognizing the speech words only. However, the recognition system has no idea what those words mean. It only knows that they are words and what words they are. To be of any use, these words must be passed on to higher-level software for syntactic and semantic analysis. It is a technique of pattern recognition, where acoustic signals are tested and framed into phonetics (number of words, phrases and sentences) [1],[2],[3],[5],[7], [8], [9].

The speech wave itself contains linguistic information that includes the meaning the speaker wishes to impart, the speaker's vocal characteristics and the speaker's emotion. Speech recognition is the process of automatically extracting and determining linguistic information conveyed by a speech wave using computers or electronic circuits. Only the linguistic information is needed from the speech wave, while the rest of the information is used in other fields of signal processing. There are basically two types of speech they are:

1. Continuous speech

2. Discrete speech.

Discrete speech consists of isolated words that are separated by silences. The advantage of discrete speech is that word boundaries can be set exactly while with continuous speech; words will be spoken without silences[5], [6], [12], [18], [19].

The remaining of this paper will discuss the system architecture in section 2 and the phases of the ASR System Architecture are presented in section 3, section 4 deals with the results and experiment, conclusion is presented in section 5 and finally section 6 for references.

3. SYSTEM ARCHITECTURE

The block diagram of ASR System in fig. 3 has the following components: Data acquisition, Preprocessing, Features extraction, Features

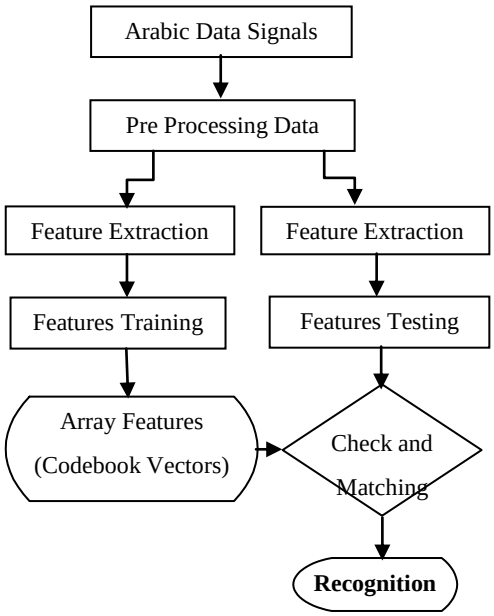

Figure 3: ASRS Architecture

Training, Codebook Vectors, as well as Features Testing, Check and Matching and Recognition for acoustics words. The steps will discuss in details as follows:

3.1 Data Acquisition

Own Data samples has been collected from Arabic speakers whose can speak Arabic language fluently and they recorded by the same recorder one by one to speaks Arabic words from (safer to ashrah) (zero to ten) . The data consists of about 100 samples with ten words and each speaker recorded five times for more efficiency. Though this might seem a lot, it is probably sufficient to obtain an ASR System. More over the speech database contains only male speakers. The speech samples are stored into (namefiles.wav) and the results in a discrete-time speech signal [17].

3.2Preprocessing

Pre-processing was used to make the discrete-time speech signal more amendable for the processes that followed. There are five pre-processing techniques that can be used to enhance feature extraction. These include DC offset removal, silence removal, pre-emphasis, windowing and autocorrelation [7], [10], [15],[18], [22].

3.2.1. Dynamic time warping (DTW)

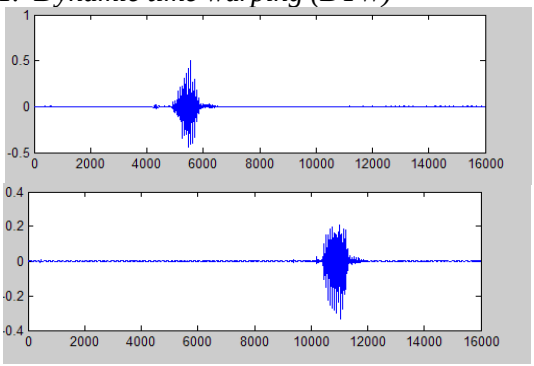

Figure 4: Utterances of the same word " Wahad " at different times

DTW is a technique that finds the optimal alignment between two time series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis as shown in fig 4.

3.2.2. Noise elimination

The biggest problem ever been in speech recognition systems is the noise in the environment. The pre-trained model for test might be inaccurate; the best result is got when we do the test in exactly the same room as the study recorded the training data. Speaking way during training also affects the result a little bit, since they does not make sure people say the same word always in the same way. It is necessary to record different kinds of utterances of Arabic words from the same person, to make sure the test utterance can still be recognized even if it is spoken in a weird way as well as in different times.

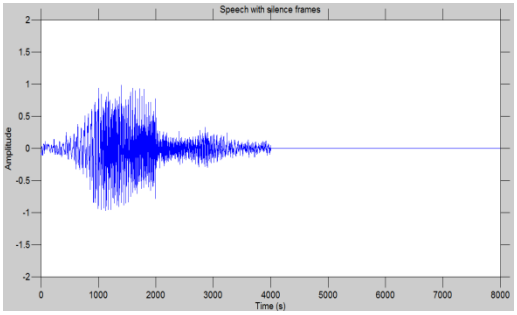

Figure 5: The word Wahad after detecting noises

3.2.3 Silence Removal

Voice Activity Detection (VAD), is the technique used to scan the speech signal from the end to its beginning to determine the presence or absence of speech signal.

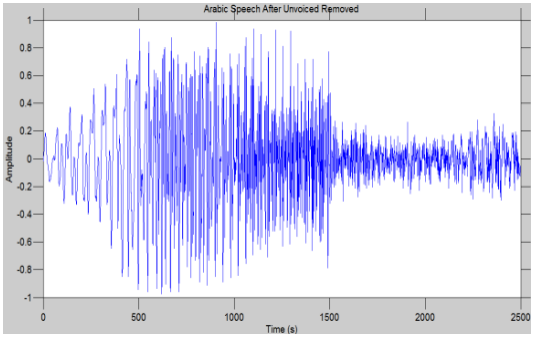

Figure 6 : The word whahad After Unvoiced Removed

The VAD may not just indicate the presence or absence of speech, but also whether the speech is voiced or unvoiced, sustained or early, etc. The technique is also used for deleting all values under some specified value which is the noise value.

3.3 Feature Extraction

There are a wide range techniques available for features extraction of speech signals. The most prominent ones are Linear Predictive Coding (LPC) and Fast Fourier transforms (FFT) and Mel Frequency Cepestral Coefficients (MFCC). MFCC is perhaps the best known and most popular, and will be described in this paper. The popularity of this method can be explained by the low computational cost compared to FFT and LPC based techniques [1], [2], [4], [6], [10], [18], [19], [23].

The main objective of this stage is to extract the important features that are enough for the Recognizer to recognize the acoustic words. Stages of feature extraction:

1) Rectangular window
2) Spectrum computation (Fast Fourier transform).
3) Power spectrum computation.
4) Mel frequency scale mapping.
5) Mel power spectrum computation.
6) Cepstrum computation.
7) Cepstral filtering.

3.3.1 Rectangular Window

Multiplication of the signal by a window function in the time domain is the same as convolving the signal in the frequency domain. Rectangular window gives maximum sharpness but large side-lobes (ripples) - hamming window blurs in frequency but produces much less leakage.

3.3.2 Spectrum computation

Compute the Fast Fourier Transform (FFT), which yields the discrete, complex–valued short term spectrum of the speech signal, the FFT is more faster and accurate than the discrete Fourier transfer.

3.3.3 Power spectrum computation

After computing the V(n) we take the absolute value |V(n)| and then take the power $|V(n)|^2$. Then we have the Power Spectrum = |V (n)|2.

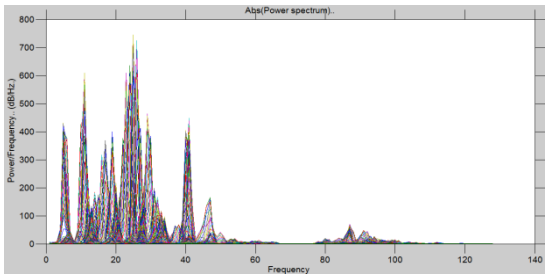

Figure 7: Power Spectrum Features

3.3.4 Mel frequency scale mapping.

Because the human ear does not show a linear frequency resolution but builds several groups of frequencies and integrates the spectral energies within a given group as shown in fig 8. Furthermore, the mid-frequency and bandwidth of these groups are non-linearly distributed. The non–linear warping of the frequency axis can be modeled by the so–called Mel-scale. The frequency groups are assumed to be linearly distributed along the Mel-scale.

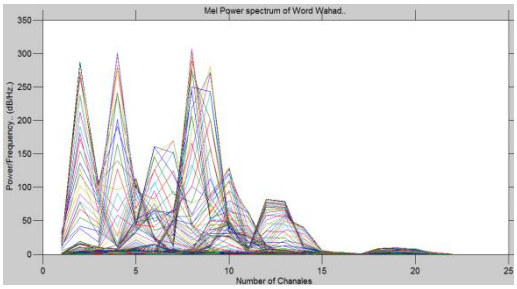

Figure 8: Mel-sclale for the Word wahad

3.3.5 Mel power spectrum computation

Mel-filter coefficients: Construction of filter channels with center frequencies linearly distributed along Mel scale. We are going to deal with Mel power spectrum not the original power spectrum, so we will multiply the original power spectrum by the Mel filter coefficients to get Mel power spectrum.

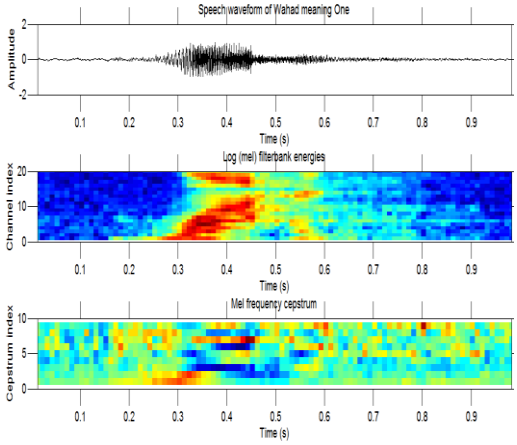

Figure 9: MFCC Features for word Wahad

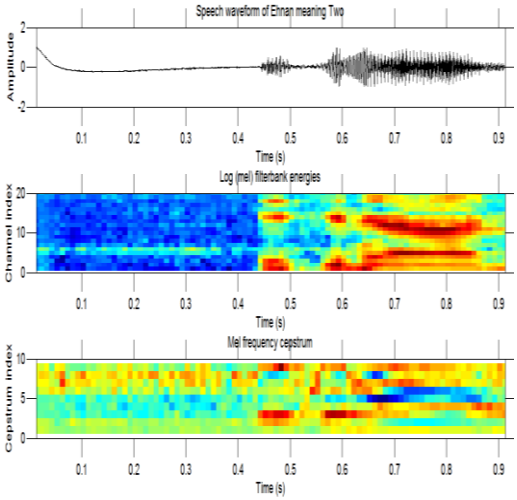

Figure 10: MFCC Features for word Ethnan

3.3.6 Cepstrum computation

Cepstrum is the inverse of spectrum, Cepstrum is in time domain, it is obtained by inverse Fourier transform to power spectrum. To obtained the Mel Cepstrum, so the study toke the log of Mel power spectrum instead of the power spectrum itself and transform it. Since the Mel power spectrum is symmetric

"as explained earlier", the Fourier Transform can be replaced by a simple cosine transform as shown blue.

$$C(q) = \sum_{k=0}^{k-1} \log(G(k)) * \cos(\frac{\pi * q(2k+1)}{2k})$$

Where C(q) is cepstrum in time domain

3.3.7 Cepstral filtering "liftering"

The MFCC are used directly for further processing in the speech recognition system instead of transforming them back to the frequency domain, but the study first applied the liftering process. Liftering is cutting off the higher order coefficients "max 14 coefficients" [15].

3.4.Features Training

Vector Quantization must able to estimate of the computed feature vectors. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the features[1],[2]. The VQ technique is consists of extracting a small number of representative feature vectors as an efficient means of characterizing the word specific features. By means of VQ, storing every single vector that we generate from the training is impossible. By using these training data features are clustered to form a codebook for each acoustic of word [6], [10], [23]. Finally Saved trained features with specific intent for using and editing without the aid of any programming algorithms.

3.4.1 Feature Matching

The *patterns* in this case are sequences of acoustic vectors that are extracted from an input speech using many techniques described previously. The classes here refer to similar utterance words. Since the classification procedure is applied on extracted features, it can be also referred to as feature matching. *Training set* data used to derive trained algorithm [17]. The remaining *Testing set* used for testing and classification algorithms. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm. The VQ technique used for mapping 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 *codeword* and the collection is called a *codebook* [6], [14], [16], [17].

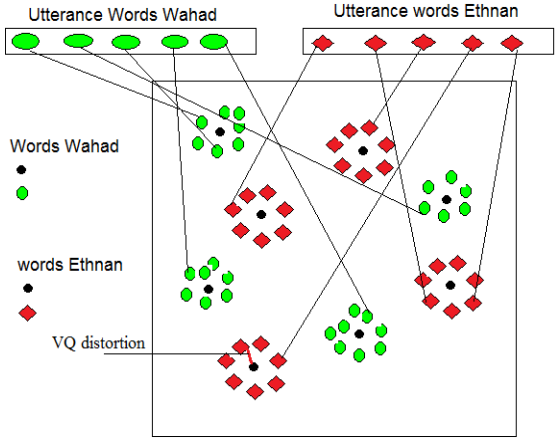

Figure 11: Vector Quantization

3.4.2 Clustering the Training Vectors

After the enrolment session, the acoustic vectors extracted from input speech of each features word provide a set of training vectors for that speech. As described above, the next important step is to build a word-specific VQ codebook for each word using those trained vectors. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of *Xi* trained vectors into a set of *Yi* codebook vectors [16], [18], [19].

3.4.3 Check and Matching

In this phase, an unknown words's voice is represented by a sequence of feature vectors (w1, w2 ….wi and it is compared with the codebooks database. In order to identify the unknown words, this can be done by measuring the distortion distance of two vector sets based on minimizing the distances. The Euclidean minimum distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn), is given by the formula as following:

$$\sqrt{\sum_{i=1}^{n}(pi-qi)^2} = \sqrt{(p1-q1)^2+(p2-q2)^2+.....(pn-qn)^2}$$

The word with the lowest distortion distance is chosen to be recognized as it is assigned to the known classes.

3.5.Recognition

It is very important to create an acoustical model for the detection of each uttered words. It is known that different Arabic words are produced by Arabic Speakers vocal cord and different sounds can have different frequencies, power spectral and MFCC which used to predict for computing different features on each class belong. Speech can be termed as short term stationary so MFCC features were again extracted and words pronounced were detected and classified according to each similarities classes of features the same words and differences of the feature for the different words. In this stage, the testing features is compared with codebook of each word and measure the differences. These differences are then used to make the recognition decision [23].

4. EXPERIMENT AND RESULT

The Matlab environment offers the most results as a matrix where each column is a frame of N samples from original speech signal [17] ,[18]. The study applied the pattern recognition technique to build words reference models for trained vectors and then can be recognize any sequences of acoustic vectors uttered by unknown uttered word. The Euclidean Distance method used to compute the pairwise ED between the code words and training vectors in the iterative process. Train and test programs used to simulate the training and testing procedure in Arabic speech recognition system, respectively [19], [23].

5. CONCLUSION

The results obtained using MFCC, VQLBG are considerable for training and building model for each speech word features. Euclidean algorithm computed distortions used for recognition Arabic Speech and accuracy obtained was 82.5%. It can improved by taking voice samples using high quality audio devices in a noise free environment. As well as use more numbers of centroids increases the performance factor but degrades the computational efficiency. Hence an economical trade-off between code vectors and number of computation is required for optimizing performance of VQLBG with Euclidean distances algorithms.

6. REFERENCES

[1] Nisha N. Nichat and P. C. Latane, "Real Time Speaker Recognition using Mel- Frequency Cepstral Coefficients

(MFCC),VQLBG & GMM Techniques", Vol. 5, Issue 6, June 2016.

[2] Dr. R.K. Prasad and Mr. K.Patel, "Speech Recognition and Verification Using MFCC & VQ", IJAR in CS and SE, Vol.3, Issue 5, 2013.

[3] Ms. Vrinda ,Mr. Chander Shekhar, "Speech Recognition System For English Language" IJAR in CCE , January 2013.

[4] S Chitode, Anuradha S. Nigade " Throat Microphone Signals for Isolated Word Recognition Using LPC ", IJAR in CS and SE, Volume 2, Issue 8, August 2012.

[5] Ms. Arundhati S. Mehendale and Mrs. M.R. Dixit "Speaker Identification" Signals and IP:, An IJ (SIPIJ) Vol. 2, June 2011.

[6] M. Shaneh and A. Taheri, " Voice Command Recognition System Based on MFCC and VQ algorithms" World Academy of SE&T 2009.

[7] Vergyri, D., K. Katrin, D. Kevin and A. Stolcke, " Morphology-based language modeling for Arabic speech recognition" Proc. ICSLP, Jeju, 2004, South Korea.

[8] Kirchhoff, K., et al., "Novel approaches to Arabic speech recognition", Final Report from the JHU Summer Workshop, 2002.

[9] Mowaffak Al_Barraq And S.C. Mehrotra "Handwritten Arabic Text Recognition System Using Window Based Moment Invariant Method", IJARCS, Volume 2, No. 1, 2011.

[10] Lazli, L. and M. Sellami, "Speaker independent isolated speech recognition for Arabic language using hybrid HMM-MLP-FCM system", AICCSA, Tunisia, 2003.

[11] Bahi, H. and M. Sellami, "A connectionist expert approach for speech recognition", The International Arabic Journal of IT, 2004.

[12] El Choubassi, M.M. et al., "Arabic speech recognition using recurrent neural networks", IEEE, Intl. Symp. Signal PI, 2003.

[13] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech, Signal Processing, Vol. ASSP-28, No. 4, August 1980.

[14] Y. Linde, A. Buzo& R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

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

[16] F.K. Song, A.E. Rosenberg and B.H. Juang, "A vector quantization approach to speaker recognition", AT&T Technical Journal, Vol. 66-2, pp. 14-26, March 1987.

[17] Jamel Price, S. Student, Dr. Ali Eydgahi "Design of an Automatic Speech Recognition System Using MATLAB" Chesapeake Information Based Aeronautics Cons. 2005.

[18] E. Darren. Ellis "Design of a Speaker Recognition Code using MATLAB ", Dep. of Computer and EE, University of Tennessee, Knoxville Tennessee 37996. 9th May 2001.

[19] Ramzi A. Haraty and Omar El A, " CASRA+: A Colloquial Arabic Speech Recognition Application", Lebanese A. Un., B., L., 2007.

[20] G.S. KUMAR, K.A.P. RAJU, Dr.Mohan R. C. and P.Satheesh, "Speaker Recognition Using GMM", IJE Science and Tech, Vol. 2(6), 2010.

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

[22] L.R Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Englewood Cliffs, N.J., 1978.

[23] H Mr. Kashyap Patel, "Speech Recognition and Verification Using MFCC & VQ", IJAR in CS and SE, Volume 3, Issue 5, May (2013).