بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

Sudan University of Science &Technology

Faculty of Computer Scinese & Infromation Technology

Department of Software Engineering

Speech To Text Conversion

تحويل الكلام العربي إلى نص مكتوب

A thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor in Software Engineering

Prepared by:

Alaa Hassan Mahmoud

Salma Alzaki Ali

Supersior by:

Dr. Howida Ali Abdul Gadir

2014

الآية

قال تعالى:-

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ
ٱقۡرَأۡ بِٱسۡمِ رَبِّكَ ٱلَّذِي خَلَقَ ١ خَلَقَ ٱلۡإِنسَٰنَ مِنۡ عَلَقٍ ٢ ٱقۡرَأۡ
وَرَبُّكَ ٱلۡأَكۡرَمُ ٣ ٱلَّذِي عَلَّمَ بِٱلۡقَلَمِ ٤ عَلَّمَ ٱلۡإِنسَٰنَ
مَا لَمۡ يَعۡلَمۡ ٥

صدق الله العظيم

سُورَةُ العَلَقِ، الآيات (١-٥)

DEDICATION

To our mothers …… Our first Teachers
To our fathers …….. Our Heroes
To Our Brothers, Sisters
&
To our Friends
We dedicate this research

Acknowledgment

First of all , thank to Allah who owed us with courage and ability accomplish this study .

Second we are deeply thankful to our university which gave us the chance to conduct our study.

Third, we are really grateful to our supervisor:

Dr. Howida Ali Abdul Gadir who exerted all possible efforts to us from the beginning of the study until its final stage and we benefited a lot from his valuable instruction.

ABSTRACT:

Though Arabic language is a widely spoken language, research done in the area of Arabic Speech Recognition is limited when compared to other similar languages. This paper concerns with convert Arabic spoken word into text using Mel-frequency Cepstrum Coefficient **(MFCC)** and Vector Quantization **(VQ)**.
This has been realized by first recording teachers's voices for each word in a noisy environment. Secondly these words have been used to extract their features using the Mel Frequency Cepstral Coefficients **(MFCC)** technique which are taken as input data to the Vector Quantization to construct codeword for each word . Finally ,in the conversion stage each codeword was indexed with the corresponding text.
The system targeting deaf students to help them solve some of the problems which face them in the university environment.
The system Word Error Rate was 20%.

مستخلص البحث

على الرغم من اللغة العربية من اللغات واسعة الإنتشار ،إلا أن الأبحاث التي أجريت في مجال التعرف على الكلام محدودة مقارنة
باللغات المماثله.
في صدد الإضافه إلى تطبيقات التعرف على الكلام العربي قمنا بتقديم تطبيق يحول الكلمه العربيه المنطوقه الى نص مكتوب.
يستهدف النظام فئة الطلاب الصم لمساعدتهم في حل بعض المشاكل التي تواجههم في المحيط الجامعي.
في أولى مراحل التطوير قمنا بتسجيل أصوات الأساتذه ,ثم استخرجت خصائص هذه الأصوات باستخدام تقنية ميل معاملات التردد Mel Frequency Cepstral Coefficients.
وأخذ الناتج من مرحلة استخراج الخصاائص إلى مرحلة التدريب والتي استخدمت فيها تقنية تكميم الإتجاه (Vector Quantization)
واخيرا تم عمل فهرسه لبيانات التدريب مع النص المقابل لكل كلمه للمساعده في عملية التحويل.
تم التوصل إلى معدل خطأ قدره 20%.

Chapter one

Introduction

CHAPTER ONE

INTROUDUCTION

1.1 Introduction:-

The advancement of Information and Communication Technology has effected in all aspects of our lives . so we can use it to improve our communication ,work and learning. The world population has just touched 7 Billion in 2012. According to the World Federation of the Deaf ,the total number of deaf people worldwide is around 70 million.

A lot of applications were developed in order to help people who have learning disabilites and improve thier lives's quality and solve some of their proplems.

In order to contribute in development of the applications that help deaf people,we developed a speech to text conversion system , as a result it will realy effect their communication with their enviroment.

1.2 The problem:-

the deaf people communicate with others using sign language instead of spoken language.

There are 25 countries where Arabic is an official language. In some countries Arabic is spoken by a minority of the people. Some sources put the number at 22-26 countries.

but learner deaf can use a written language to communicate ,and they face a lot of problem while they try to continue learn, specialy deaf who in university because the teachers in universities communicate with spoken language ,and the cost of come up with sign teacher to translate speech into sign will be high. And also other student will keep their attention with sign teacher .

In our project We triedto solve these proplems .so that the deaf people can communicate in effective way with the normal people.

1.3 The goal of the project:-

The system aims to help deaf people to know what is going on around them in the classroom because God did not give them by the grace of the hearing and it was necessary for us to help them and provide them with help.

The project also aims to spread knowledge and assistance in learning the generations, whatever their disability and their ability to acquire knowledge.

1.4 Importance of the project :-

The importance lies in the following: -

1 - Developing and teaching methods in universities and institutes of higher education.

2 - Assistance assigned people with special needs, especially the deaf**.**

3-suporting Arabic languages by adding new application for this language.

1.5 Scope and limitations of the project :-

Yes it was God's grace we have the science and application ,and Quest to seek knowledge was recommended by the Holy Messenger said : (Asking for knowledge from the cradle to the grave) ,But each has been deprived of some of the senses that help in seeking more of science ,and it is our duty as Muslims to help them toward each other, which urged us to develop this project, which Targeting the deaf community, especially the students of the Faculty of Fine Arts at the University of Sudan for Science and Technology (western branch) , So the message is sent for each educational category in university community .

in our project we concern with convert ten arabic spoken word into text .

the words are :

"سلام " ، " عليكم " ، " محاضره " ، " سابقه " ، "تحدثنا" ، " في " ، "رحمه" ، " الله" ، "عن" ، "موضوع".

1.6 Tools:-

MATLAB:

MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation.

Using the MATLAB product, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran.

1.7 The system overview :-

We try to do something for deaf people to help them and improve their life by ngissa some applications or something special to make the communications with external world easy and comfortable , and we use Arabic languages in our project because this is our main languages .

Our project is application that convert the speech to text that means if the teacher who train the system on his voice ,speak his voice signal , will be processed to find the Word matching and the result should be shown as a text .

We hope our project expanded in the future and work at all Arabic languages (not just ten words) that used in education environment this help students attain a better understanding of their curriculum , allow for simultaneous note-taking during the lecture, and help students to complete homework after the lecture.

1.8 Content:

In this section, we present brief information about the rest of this thesis. The remainder part of this thesis is:

- **Chapter 2: Literature Review:** This chapter intends to discuss how the speech signal is processed, the basics of speech recognition and the methods used in this field.
- **Chapter 3: System implementation:** This chapter describes the project implementation steps.
- **Chapter 4**: **Conclusion and Future work:** This chapter shows a conclusion for the results obtained and the recommendation of this research.
- **References:** Here are the used citations indexed by numbers.
- **Appendix A:** This appendix contains the project user interface.

CHAPTER 2

MOTHODOLOGY

2.1 Introduction

In this chapter we will show how to design the speech recognition system using the algorithm of Vector Quantization (VQ) methods Also feature extraction and matching techniques used in the method above will be shown and discussed.

2.2Speech recognition

Speech recognition applications are becoming more and more useful now a days. Various interactive speech aware applications are available in the market. but they are usually meant for and executed on the traditional general purpose computers. with growth in the needs for embedded computing and the demand for emerging embedded platforms, it is required that the speech recognition systems (SRS) are telephones handheld devices which becoming more and more powerful an affordable as well. It has

become possible to run multimedia on these devices. speech recognition systems emerge as efficient alternatives for such devices where typing becomes difficult attributed to their small screen limitations.

We used this property for the development of this system that helps deaf people in the educational field, and this has led to support the development and increase speech recognition systems .

2.3 Voice recognition: -

Voice recognition is an ability of a computer, computer software program, or hardware device to decode the human voice into digitized speech that can be interpreted by the computer or hardware device. Voice recognition is commonly used to operate a device, perform commands, or write without having to operate a keyboard, mouse, or press any buttons. Today, this is done on a computer with automatic speech recognition (ASR) software programs. many ASR programs require the user to "train" the ASR program to recognize their voice so that it can more accurately convert the speech to text.

2.3.1 Types of voice recognition systems

There are two types of voice recognition systems:-

2.3.1.1 Speaker Dependence:-

Speaker dependence describes the degree to which a speech recognition system requires knowledge of a speaker's individual voice characteristics to successfully process speech. The speech recognition engine can "learn" how you speak words and phrases; it can be trained to your voice. Speech recognition systems that require a user to train the system to his/her voice are known as speaker-dependent systems. If you are familiar with desktop dictation systems, most are speaker dependent. Because they operate on very large vocabularies, dictation systems perform much better when the speaker has spent the time to train the system to his/her voice.

2.3.1.2 Speaker Independence :-

Speech recognition systems that do not require a user to train the system are known as speaker-independent systems. Speech recognition in the VoiceXML world must be speaker-independent. Think of how many users (hundreds, maybe thousands) may be calling into your web site. You cannot require that each caller train the system to his or her voice. The speech recognition system in a voice-enabled web application MUST successfully process the speech of many different callers without having to understand the individual voice characteristics of each caller.[1]

2.4 The process of speech recognition:-

The figure 1 below the process and the interaction between them .

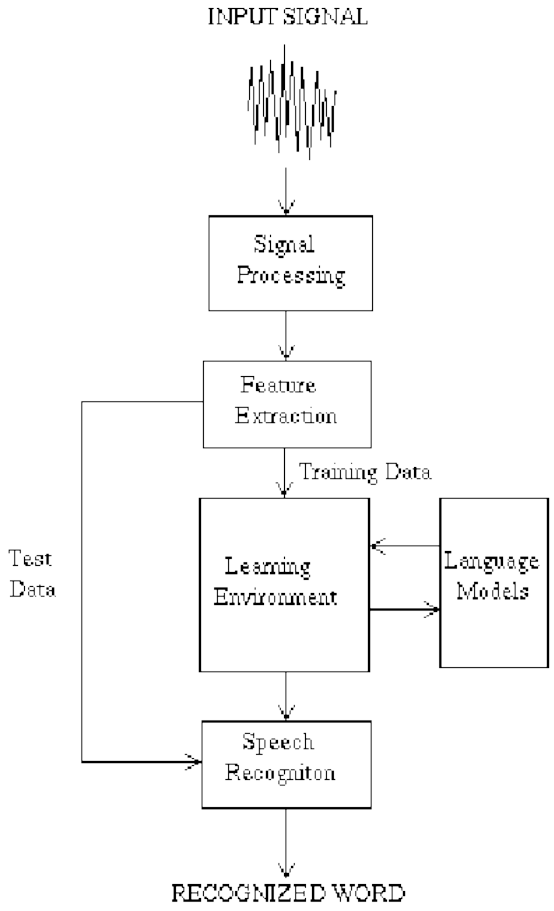


Figure 1 : the process of speech recognition

2.4.1 Feature Extraction:-

In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal.

The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectrum temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals.

The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage(which is not always present) transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognition stage.[3]

2.4.1.1 Feature extraction using MFCC:

In this project we are using the Mel Frequency Cepstral Coefficients (MFCC) technique to extract features.

The Mel-frequency Cepstrum Coefficient (MFCC) technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. It show human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale.

The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels , a block diagram of the MFCC processes is shown in figure 2 .[2]

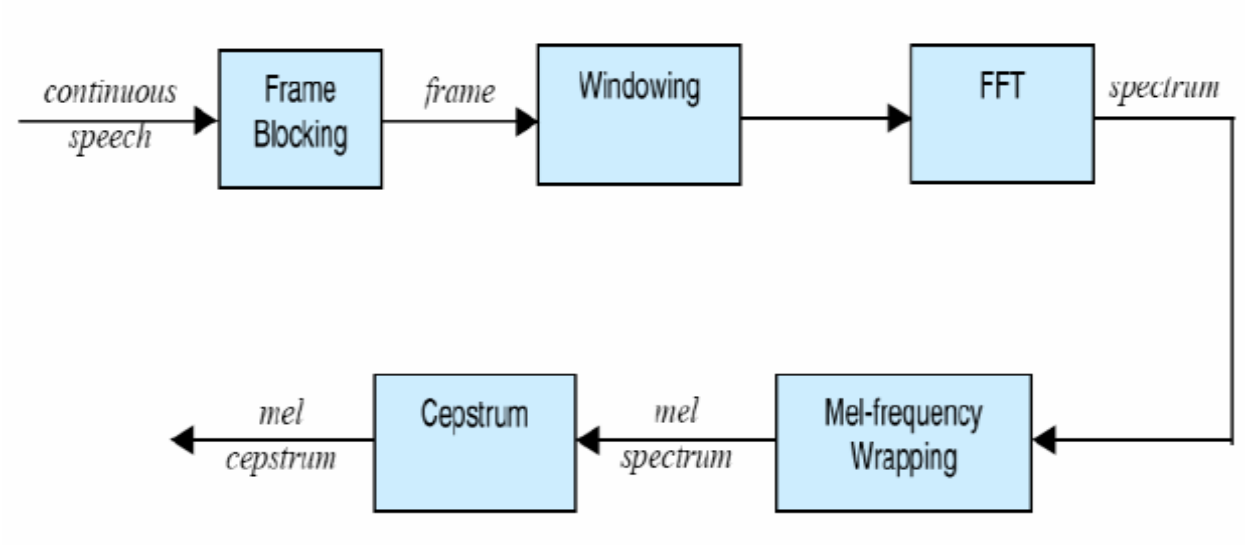


Figure 2:Block diagram of MFCC[1].

Step 1:Frame Blocking:

The speech wave is cropped to remove silence or acoustical interfaces that may be present at the beginning or the end of the sound file ,the wave is divided into a small frame with the length within the range of 20 to40 msec. the voice signal is divided into frames of N samples. The first frame consists of the first N samples. the second frame begins M samples after the first frame, and overlaps it by N - M samples and so on. this process continues until all the speech is accounted for within one or more frames. Adjacent frames are being separated by:

M (M<N).

Typical values used are M = 100 and N= 256.

Step 2: windowing step:

The windowing step aim to minimize the signal discontinuities, the idea behind this step is to minimize the spectral distortion at the beginning and end of each frame, by tapering the beginning and the end of each frame to zero.

Step 3 : FFT(Fast Fourier Transform):

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain .[2]

Step 4: Mel Filter Bank Processing

Filterbank Analysis :-

The human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves recognition performance. A popular alternative to linear prediction based analysis is therefore filterbank analysis since this provides a much more straightforward route to obtaining the desired non-linear frequency resolution. However, filterbank amplitudes are highly correlated and hence, the use of a spectral transformation in this case is virtually mandatory if the data is to be used in a HMM based recogniser with diagonal covariances.

To implement this filterbank, the window of speech data is transformed using a fourier transform and the magnitude is taken. the magnitude coefficients are then binned by correlating them with each triangular filter. Here binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filterbank channel. As an alternative, the Boolean configuration parameter USEPOWER can be set true to use the power rather than the magnitude of the Fourier transform in the binning process.

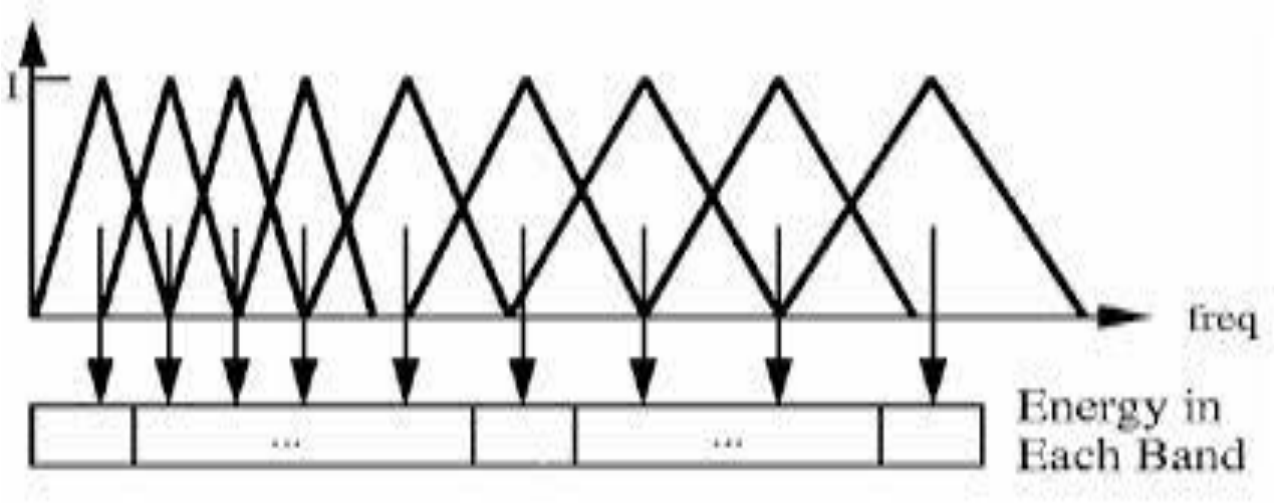


Figure 3:Meal-Scale of FilterBank

Normally the triangular filters are spread over the whole frequency range from zero up to the Nyquist frequency. However, band-limiting is often useful to reject unwanted frequencies or avoid allocating filters to frequency regions in which there is no useful signal energy.

Step5 :Mel-frequency wrapping:-

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz .As a reference point ,the pitch of a 1 KHz tone ,40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency fin Hz.

$$Mel(f) = 2595*\log10(1 + f/700) \quad \text{---------- (1)}$$

Ours approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular band pass

frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval , the mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. this corresponds to series of band pass filters with constant bandwidth and spacing on a mel frequency scale .

Step 6 :Cepstrum :-

Here we convert the log mel spectrum backto time. the result is called the Mel Frequency Cepstrum Coefficients (MFCC).The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time.

The result is called the Mel Frequency Cepstrum Coefficients (MFCC).the discrete cosine transform is done fortrans forming the mel coefficients back to time domain.[2]

2.5 Pattern Recognition:-

The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from aset of labeled training samples via a formal training algorithm.A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOVMODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns .[3]

2.5.1 Template Based Approach:-

Template based approach to speech recognition have provided a family of techniques that have advanced the field considerably during the last six decades. the underlying ideais simple. A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidates words.,recognition is then carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. Usually templates for entire words are constructed.

This has the advantage that, errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided. In turn, each word must have its own full reference template; template preparation and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred words. one key idea in template method is to derive a typical sequences of speech frames for a pattern (a word) via some averaging procedure, and to rely on the use of local spectral distance measures to compare patterns. Another key idea is to
use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across talkers as well as across repetitions of the word by the same talker .[3]

2.5.2 Stochastic Approach:-

Stochastic modeling entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability s, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state markov model and a set of output distributions. the transition parameters in the Markov chain models, temporal variabilities, while the parameters in the output distribution model, spectral variabilities. These two types of variabilites are the essence of speech recognition.[3]

2.6 Vector Quantization(VQ):-

Vector Quantization(VQ)is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction. Since transmission rate is not a major issue for ASR, the utility of VQ here lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. For IWR, each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure. In basic VQ, codebooks have no explicit time information, since codebook entries are not ordered and can come from any part of the training words. However, some indirect durational cues are preserved because the codebook entries are chosen to minimize average distance across all training frames, and frames, corresponding to longer acoustic.,segments are more frequent in the training data.

The VQ puts on speech transients can be an advantage over other ASR comparison methods for vocabularies of similar words. .[3]

2.6.1 Applications use VQ:-

Vector quantization is used in many applications such as image and voice compression, voice recognition (in general statistical pattern recognition).[4]

2.6.2 Vector Quantization detail :-

A vector quantizer maps k-dimensional vectors in the vector space into a finite set of vectors Y = {yi: i = 1, 2, ..., N}. Each vector yi is called a code vector or a codeword. And the set of all the codewords is called a codebook. Associated with each codeword, yi, is a nearest neighbor region called Voronoi region .

As an example we take vectors in the two dimensional case without loss of generality. Figure 4 shows some vectors in space. associated with each cluster of vectors is a representative codeword. Each codeword resides in its own voronoi region. these regions are separated with imaginary lines in figure 4 for illustration. given an input vector, the codeword that is chosen to represent it is the one in the same voronoi region.

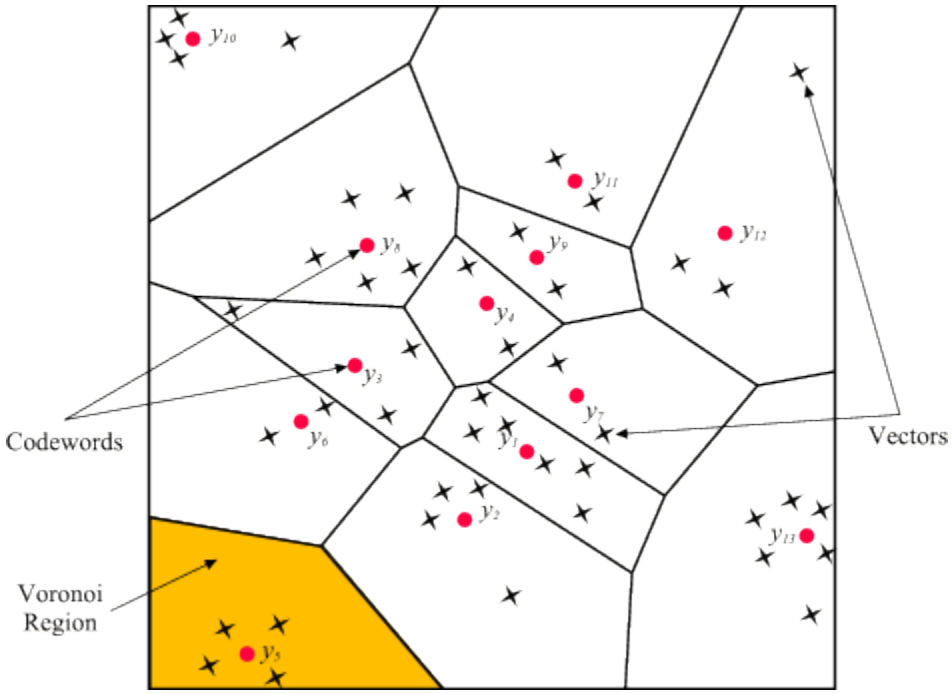


figure4 :vectors in space

The representative codeword is determined to be the closest in Euclidean distance from the input vector .where xj is the jth component of the input vector, and yij is the j this component of the codeword yi .[4]

2.6.3 CompressioninVQ:-

A vector quantizer is composed of two operations. The first is the encoder, and the second is the decoder. The encoder takes an input vector and outputs the index of the codeword that offers the lowest distortion. In this case the lowest distortion is found by evaluating the Euclidean distance between the input vector and each codeword in the codebook. Once the closest codeword is found, the index of that codeword is sent through a channel (the channel could be a computer storage, communications channel,

and so on). When the encoder receives the index of the codeword, it replaces the index with the associated codeword. Figure 5 shows a block diagram of the operation of the encoder and decoder. .[4]

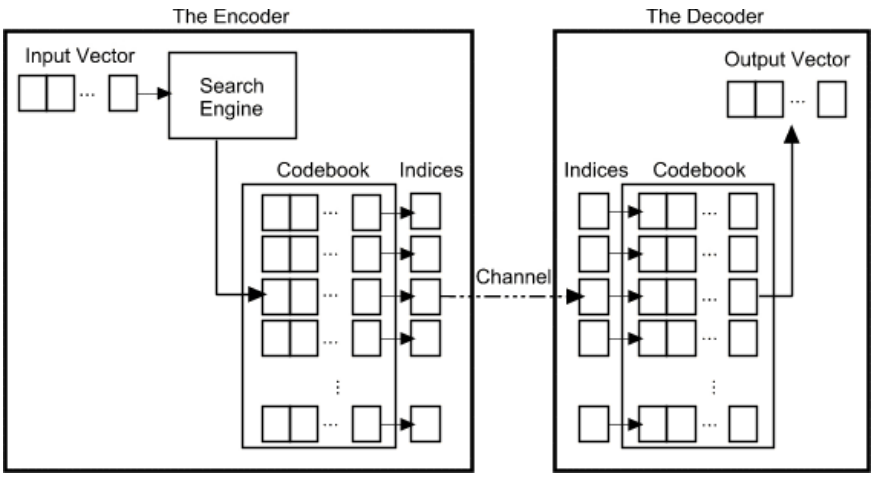


Figure 5: The Encoder and decoder in a vector quantizer.

Given an input vector, the closest codeword is found and the index of the codeword is sent through the channel. The decoder receives the index of the codeword, and outputs the codeword .[4]

2.7 Related Work :

Though Arabic language is a widely spoken language, research done in the area of Arabic Speech Recognition (ASR) is limited when compared to other similar languages. Also while the accuracy of speaker dependent speech recognizers has achieved best performance, the performance of speaker independent speech recognition system is still relatively poor. Some researches have been done in the area of Arabic speech recognition; we can mention some of relevant researches:

Kashyap Patel, R.K. Prasad ,Speech Recognition and Verification Using MFCC & VQ: The goal of the the project was to create a gender and speaker recognition system, and apply it to a speech of an unknown speaker.

By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify the unknown speaker. The feature extraction is done by using MFCC (Mel Frequency Cepstral Coefficients) and Vector Quantization (VQ) as classification algorithms.
The error rate of the system was about 13%. In second form. [5]

2. Mohammad Abushariah and Others, Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus:

The system was is based on the Carnegie Mellon University (CMU) Sphinx tools, and the Cambridge HTK tools were also used at some testing stages. The speech engine uses 3-emitting state Hidden Markov Models (HMM) for tri-phone based acoustic models.
The system obtained 91.23% and 92.54% word recognition accuracy with and without diacritical marks respectively.
The system obtained a word recognition accuracy of 95.92% and 96.29%, and a Word.
Error Rate (WER) of 5.78% and 5.45% with and without diacritical marks respectively.
On the other hand, for different speakers with different sentences, the system obtained a word recognition accuracy of 89.08% and 90.23%, and a WER of 15.59% and 14.44% with and without diacritical marks, respectively. 5.78% and 5.45% with and without diacritical marks respectively. On the other hand, for different speakers with different sentences, the system obtained a word recognition accuracy of 89.08% and 90.23%, and a WER of 15.59% and 14.44% with and without diacritical marks, respectively. [6]

3. Suma Swamy1 and K.V Ramakrishnan, An Efficient Speech Recognition
System:

The system developed using different techniques such as Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ) and Hidden Markov Model (HMM).
The coding of all the techniques mentioned above has been done using MATLAB. It has been found that the combination of MFCC and Distance Minimum algorithm gives the best performance and also accurate results in most of the cases with an overall efficiency of 95%. The study also reveals that the HMM algorithm is able to identify the most commonly used isolated word. As a result of this, speech recognition system achieves 98% efficiency. [7]

CHAPTER 3

IMPLEMENTATION AND RESULTS

3.1 Introduction

This chapter shows the implementation details of speech to text converting system. It also shows the steps required to achieve the complete voice recognition project. Also, it introduces the project testing using different testing environments and the results after testing.

3.2 System implementation

In this section all the implementation details are presented including the software used associated with some facilitation figures.

MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. Using the MATLAB product, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran .

You can use MATLAB in a wide range of applications, including signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology. Add-on toolboxes (collections of special-purpose MATLAB functions, available separately) extend the MATLAB environment to solve particular classes of problems in these application areas.

MATLAB provides a number of features for documenting and sharing your work. You can integrate your MATLAB code with other languages and applications, and distribute your MATLAB algorithms and applications. Features like High-level language for technical computing , Development environment for managing code, files, and data

Interactive tools for iterative exploration, design, and problem solving , Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration , 2-D and 3-D graphics functions for visualizing data , Tools for building custom graphical user interfaces

3.3 Implementation

The first step of the development :

3.3.1 Voice recording

In this phase we recorded teachers's voices for each teacher10 words .
The next image is an example of an uttered voice signal, certainly the word "سلام".

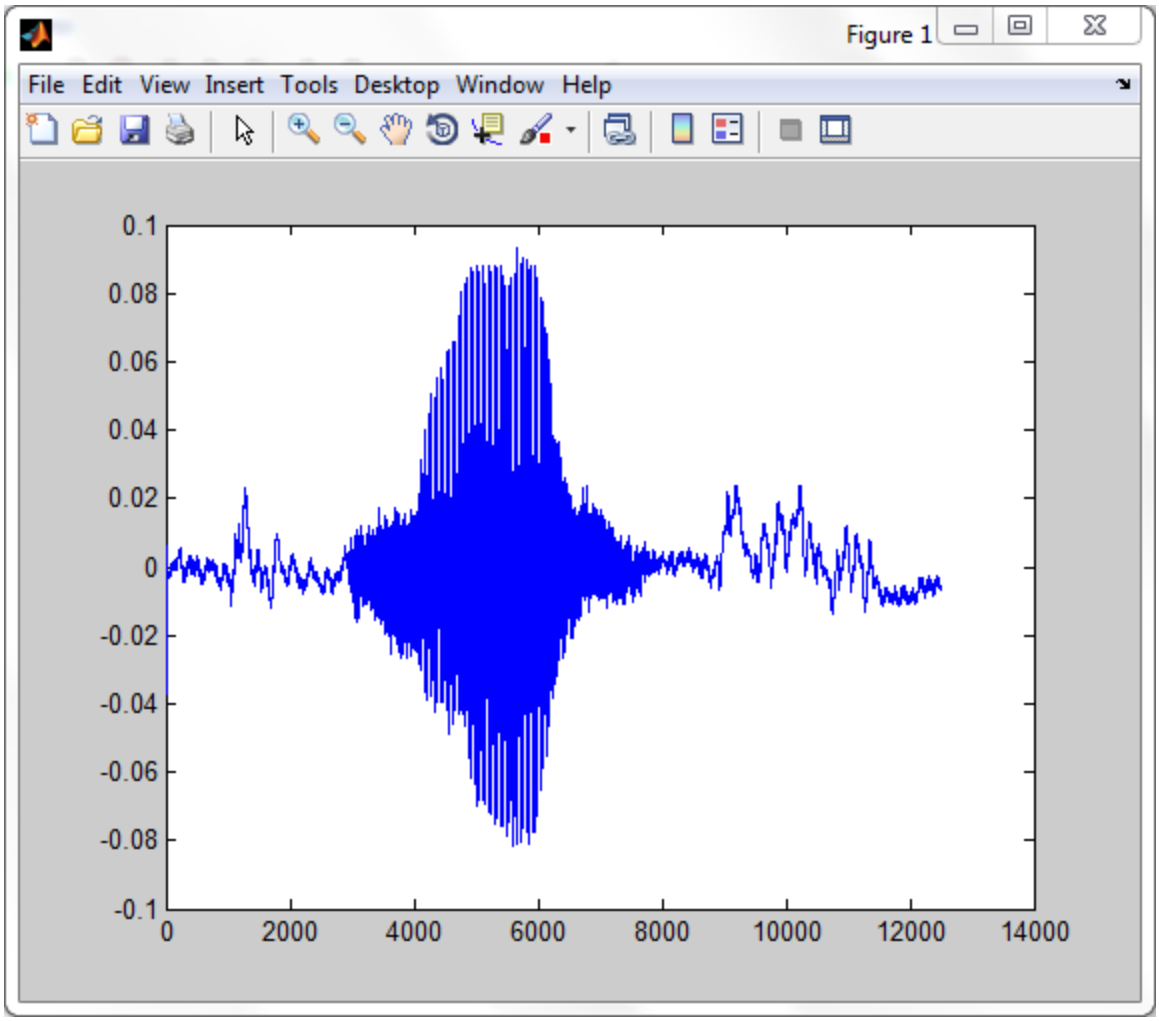


Figure 3.1 : "سلام" voice signal

3.3.2 Preprocessing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. There are two steps in Pre-processing.

1. Pre-emphasization.
2. Voice Activation Detection (VAD).

1- Pre-emphasization

The digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range and spectrally flatten the speech signal, pre-emphasis

is applied. First order high pass FIR filter is used to preemphasize the higher frequency components, the figure below show the Pre-emphasization .[7]

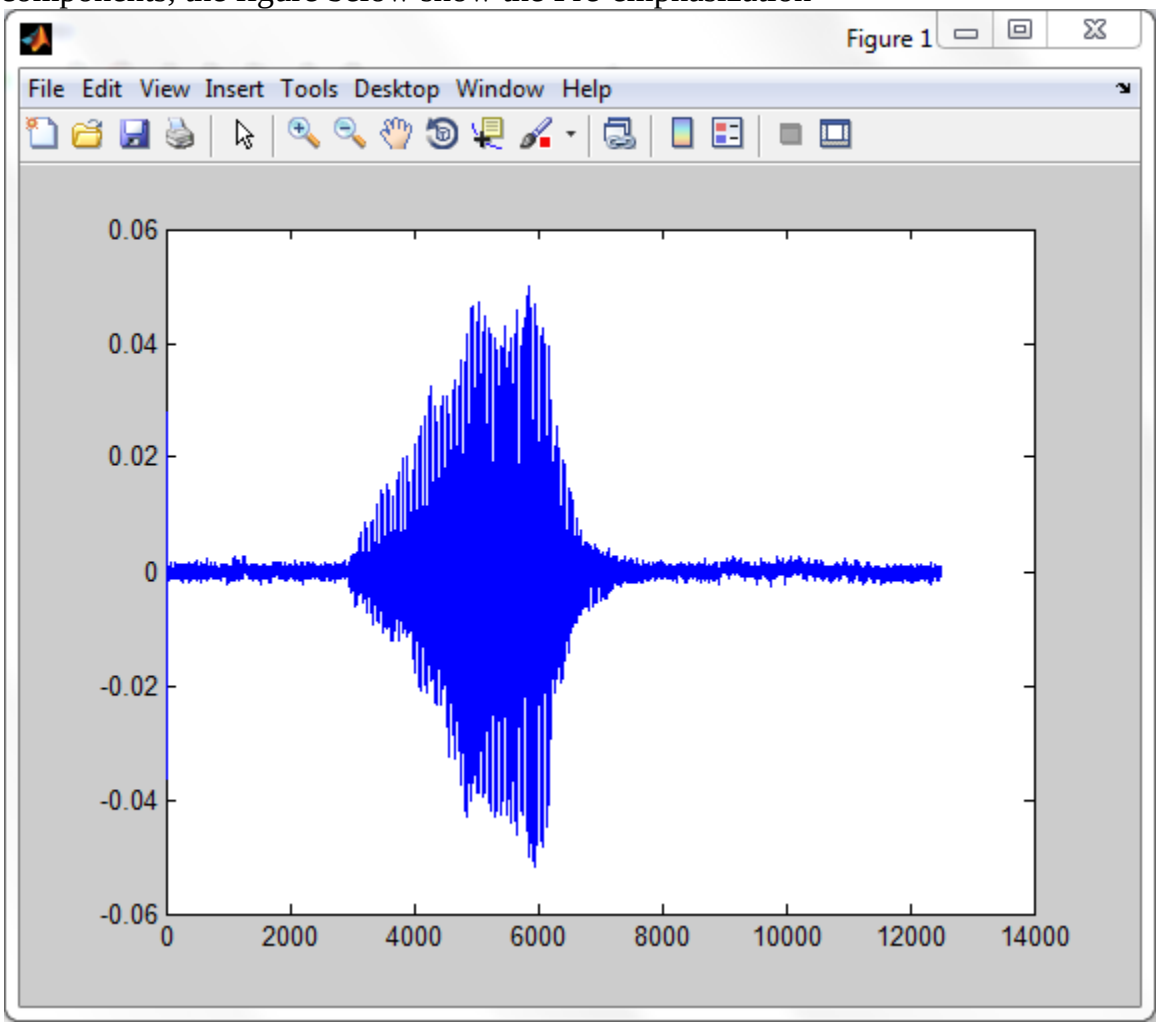


Figure 3.2 : Pre-emphasization Speech Signal

2. Voice Activation Detection (VAD)

VAD facilitates speech processing, and it is used to deactivate some processes during non-speech section of an audio sample. The speech sample is divided into non-overlapping blocks of 20ms. It differentiates the voice with silence and the voice without silence, the figure below show the Voice Activation Detection .[7]

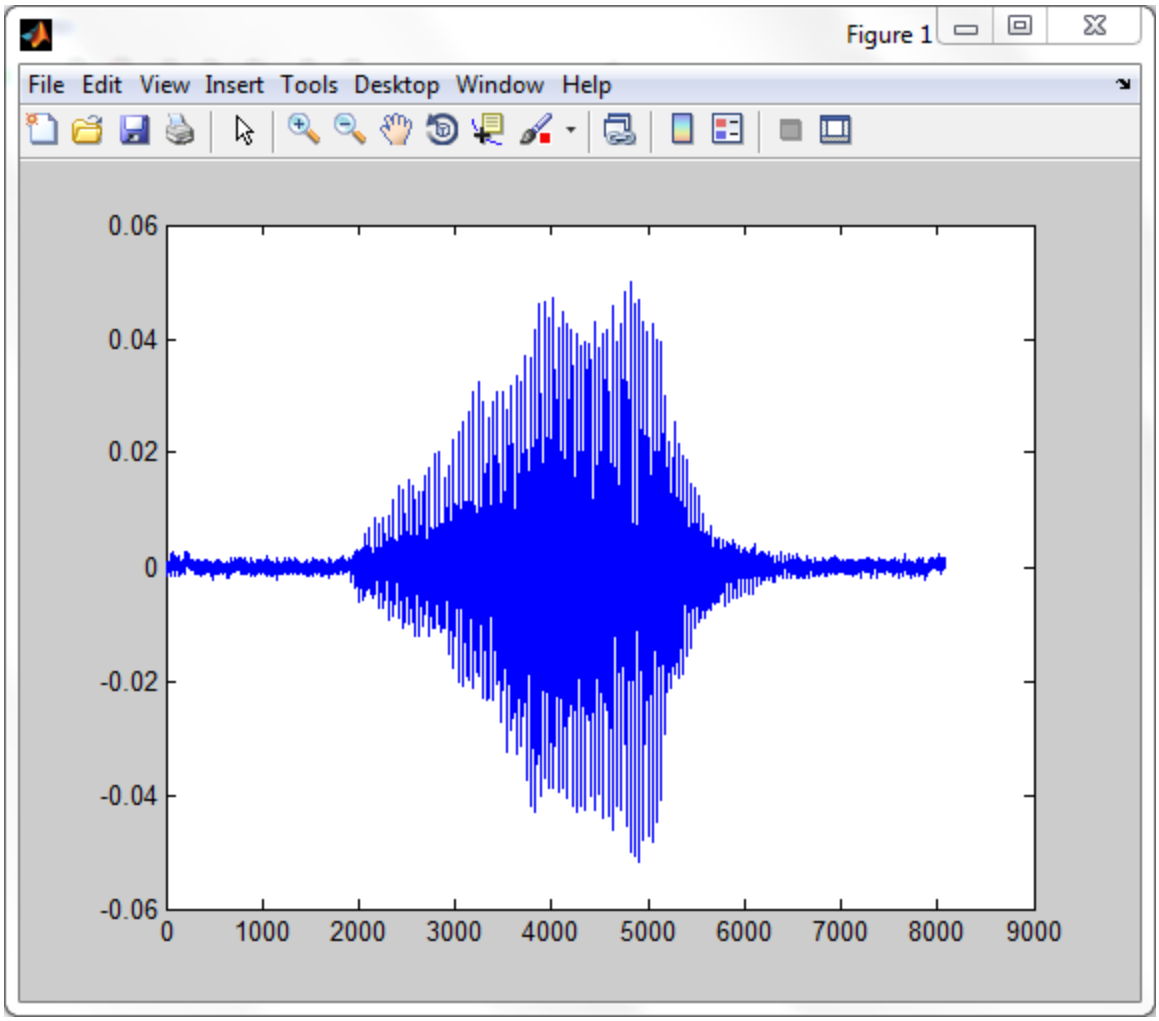


Figure 3.3.:VAD Speech Signal

3.3.3 Feature extraction

Several feature extraction algorithms can be used to do this task, such as - Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC), and Human Factor Cepstral Coefficient (HFCC), The MFCC algorithm is used to extract the features , MFCC are chosen for the following reasons:-

1. MFCC are the most important features, which are required among various kinds of speech applications
2. It gives high accuracy results for clean speech.
3. MFCC can be regarded as the "standard" features in speaker as well as speech Recognition .[7]

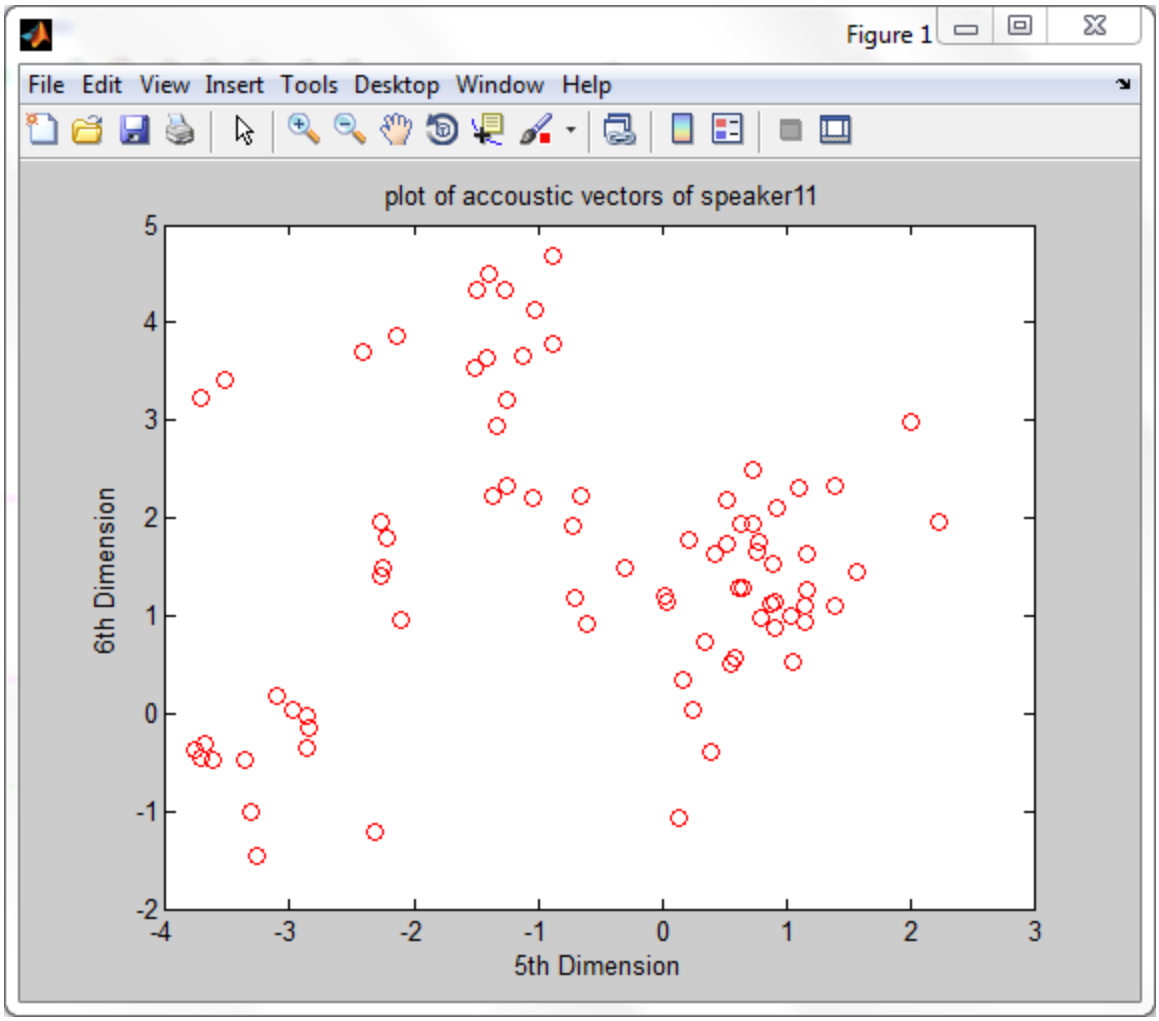
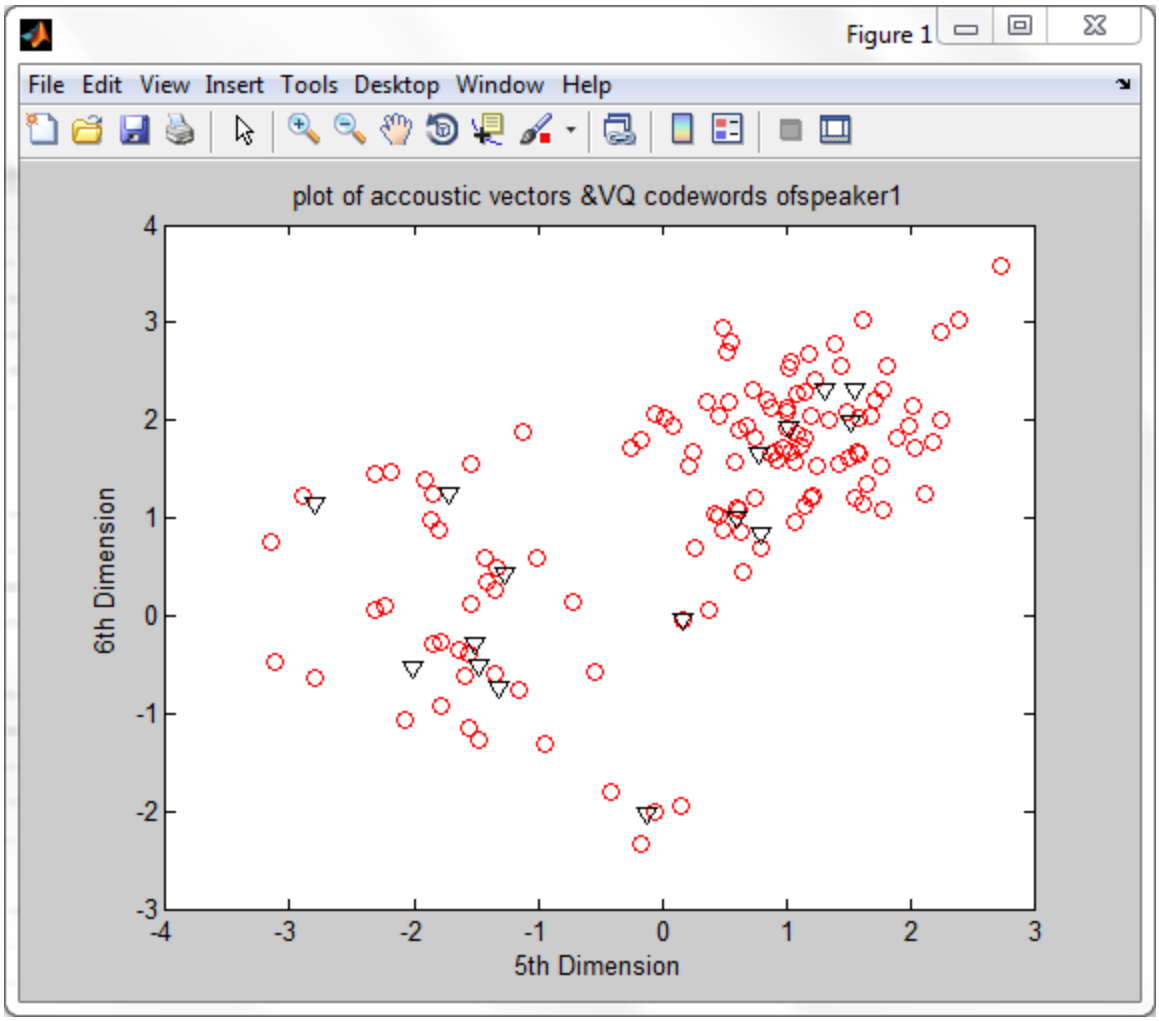


Figure 3.4 : Feature extraction

3.3.4 The Training

In this phase we took the recorded voice files one by one to create codebook using vector quantization algorithms for each files .

the next figure show the representation of vector quantization work when applying on the word "سلام".



Figurer3.5 : The traning out put

The following figure show Sample of the codebook data of the word "سلام":

-30.7271	-25.0531	-29.3751	-15.8206	-30.0231	-20.8630	-27.5050	-15.2664	-30.1930
4.8511	10.3677	5.8821	12.0396	4.8000	10.0813	7.2697	10.3662	4.7409
0.8024	4.0599	0.9979	-2.2841	-0.4296	-1.4262	1.7452	-5.3691	1.1583
0.2747	0.5100	0.5130	-3.2023	0.1365	-0.7133	1.2656	-2.2849	1.5072
0.7781	-1.4684	1.0188	-2.7915	1.5603	-0.1234	0.7922	-1.3096	1.5168
1.6634	-0.5153	1.9278	1.1473	2.3125	-2.0258	0.8454	-0.7287	1.9826
0.6076	-0.5573	0.6424	-1.1325	1.1223	-2.6315	-0.0936	-0.1549	0.4967
1.3548	0.3513	0.9876	1.2345	1.6296	1.3281	0.9597	2.1468	1.0013
0.3071	-0.8545	0.4197	0.9007	0.3339	-0.3991	-0.2062	-1.7472	0.1421
0.0947	-1.0386	0.2468	-1.3489	-0.0228	-2.0167	-0.9736	-3.1871	0.1691
0.1420	-1.0723	-0.2850	-0.1598	-0.4266	-0.2407	-0.0319	0.6214	-0.0844
0.8640	-0.0592	0.7235	0.0618	0.3922	1.2391	0.8762	0.7131	0.8452
0.2369	-0.6662	0.5521	-0.8189	0.2989	-1.2382	0.2834	-1.5095	0.3386
0.1629	0.0643	0.4853	0.7967	0.2298	-0.2932	0.2628	0.6306	0.3172
-0.1853	-0.2260	-0.0861	-0.0010	-0.2869	1.1411	-0.1281	-0.0616	-0.0113
-0.1791	-0.4285	-0.2359	-0.1718	-0.2273	1.2535	-0.0084	-0.1457	-0.0383
-0.0177	0.1720	0.0085	-0.0661	0.3020	0.2212	0.4696	0.4982	0.1062
0.1315	0.7249	0.1812	0.1798	0.0884	-0.0140	0.9097	0.4881	0.0862
0.1291	0.8590	-0.0362	0.4283	-0.0418	0.3697	0.3611	0.8947	-0.0207
0.1211	0.8658	0.0704	0.2970	0.1630	0.4138	0.0932	0.9612	0.0214

Figurer 3.6 : The codeword of the word" سلام"

Conversion stage

In this stage when the system capture a word from the user it is extract the voice feature ,then compare these feature with the codeword which stored in the training stage ,to find the suitable index .

After found the suitable index the system represent the corresponding text.

Testing phase

For our system we calculated Word Error Rate(WER) which is a common metric of the performance of a speech recognition or machine translation system.

Testing result

The system WER is 20% .

General issues should be considered to enhance system accuracy:

- Internal and external noise could affect the system accuracy.
- Internal noise should be avoided by using a high quality (fine) microphone for entering the voice.

CHAPTER 4

CONCLUSION AND FUTURE WORK

4.1 CONCLUSION:

This research has discussed an isolated word recognition system which developed using MFCC (Mel Frequency Cepstral Coefficients) as and VQ algorithms . The system was designed and implemented perfectly using matlab tool.

4.2 FUTURE WORK:

Noise is a really big deal; it can increase the error rate of speaker identification system. So, use of noise cancellation and normalization techniques to reduce the channel and the environment effects is recommended. Also, voice activity detection should be done. All of these can improve the recognition accuracy.

This system based on “isolated word recognition” ,but we hope it can be extended in the future to “Continuous Word Recognition” to make it more effective .

Some other aspects which can be looked into are:

- The system could be improved so that it can works in different training and testing environments.
- The size of the training data i.e. the code book can be increased in VQ as it is clearly proven that the greater the size of the training data, the greater the recognition accuracy .

This training data could incorporate aspects like the different ways via the accents in which a word can be spoken, the same words spoken by male/female speakers and the word being spoken under different conditions say under conditions in which the speaker may have a sore throat etc.

- The system could be designed as client-server Architectural to make the matlb work as a server to mobile phone which capture the speech and send it to the matlb to process it .

REFERENCES

1. **Kimberlee A. Kemble**:An Introduction to Speech Recognition.
2. :http://www.researchtrend.net/ijet/4_Vibha.pdf , access at: 15/6/2014 5:22 pm
3. **M.A.Anusuya**, Speech Recognition by Machine ,
4. http://www.mqasem.net/vectorquantization/vq.html,access at:15/6/2014 5:05 pm
5. **Kashyap Patel, R.K. Prasad,** Speech Recognition and Verification Using MFCC & VQ.
6. **Mohammad Abushariah and Others,** Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on Phonetically Rich and Balanced Speech Corpus.
7. **Suma Swamy1 and K.V ramakrishnan ,** An Efficient Speech Recognition system.