Speech Recognition with Dynamic Time Warping using MATLAB

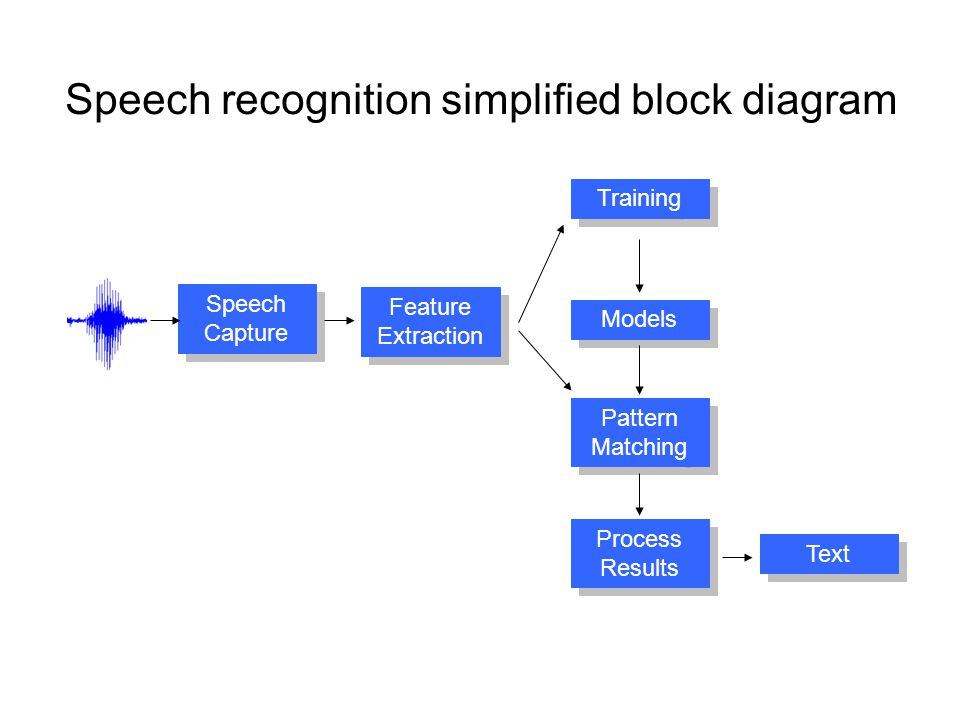
**Abstract**

Speech recognition has found its application on various aspects of our daily lives from automatic phone answering service to dictating text and issuing voice commands to computers, we present the historical background and technological advances in speech recognition technology over the past few decades. More importantly, we present the steps involved in the design of a speaker-independent speech recognition system. We focus mainly on the pre-processing stage that extracts salient features of a speech signal and a technique called Dynamic Time Warping commonly used to compare the feature vectors of speech signals. These techniques are applied for recognition of isolated as well as connected words spoken. We conduct experiments on MATLAB to verify these technique.

**INTRODUCTION**

Language is man's most important means of communication and speech its primary medium. Speech provides an international forum for communication among researchers in the disciplines that contribute to our understanding of the production, perception, processing, learning and use. Spoken interaction both between human interlocutors and between humans and machines is inescapably embedded in the laws and conditions of Communication, which comprise the encoding and decoding of meaning as well as the mere transmission of messages over an acoustical channel. Here we deal with this interaction between the man and machine through synthesis and recognition applications. The paper dwells on the speech technology and conversion of speech into analog and digital waveforms which is understood by the machines. Speech recognition, or speech-to-text, involves capturing and digitizing the sound waves, converting them to basic language units or phonemes, constructing words from phonemes, and contextually analysing the words to ensure correct spelling for words that sound alike. Speech Recognition is the ability of a computer to recognize general, naturally flowing utterances from a wide variety of users. It recognizes the caller's answers to move along the flow of the call. Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic phonetics, which describes the phonetic elements of speech (the basic sounds of the language) and tries to explain how they are acoustically realized in a spoken utterance. These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts

**Block Diagram**

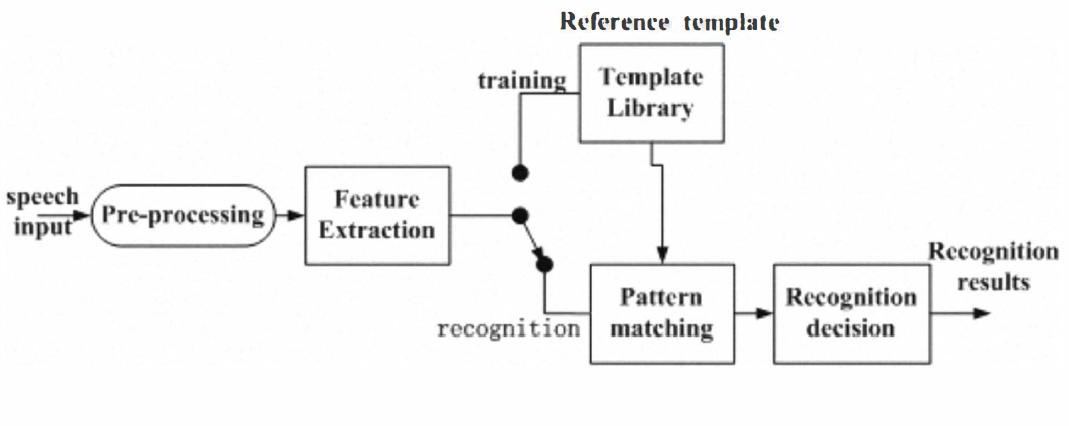


## SPEECH RECOGNITION TECHNIQUES

Speech-enabled applications over wireless networks and the World Wide Web (WWW) employing speech recognition are recently attracting more and more attention. At present google search is based on the words in text mode. ASR can be used to give the input words to google to be searched, through speech instead of text. Recently this application is also introduced in the google search. Managing the data of students, like entering the marks, attendance, and preparing progress reports is tedious process. If IWR is used for entering the data, the work load on the teacher will be reduced. Therefore ASR for IWR based HMM based word models is built for entering data of the students of a class and tested the performance of the system. However HHMs developed for word models can also used for connected word recognition and it is to spell the connected words as compared isolated words. Therefore, in this chapter the performance of the ASR system for Isolated words is given and next chapter will give the performance of connected word recognition.

Word Recognition is where a word uttered by the user has to be recognized by the speech recognition system. This is possible by the Reference Models stored in the database corresponding to each word intended to be recognized. Thus while performing the recognition an uttered word is compared to each of these models. Only the words having the Models in the system, can be recognized. If a new word is spelt for recognition, it will recognize as one of the word having model or simply give as a new word. The inputs to the Word Recognition system are stored MODELS and the MFCC features of the word uttered (TestFeatures).

The recognition process is simply matching the incoming speech with the stored Models In the recognition process, Forward Algorithm of Dynamic Time Warping, is used for calculating the Cost. All the MODELS (Reference Features) are given as Reference Features to the DTW, one after the other along with the features of the word uttered. The MFCC features of the word uttered for recognition are the test features applied to the DTW algorithm. Thus the DTW algorithm gives a cost for each model and the test features. The Model with the lowest distance measure (cost) is the recognized word. The word corresponding to the model with lowest cost is the recognized word. Hence the best match (lowest distance measure) is obtained from dynamic programming.



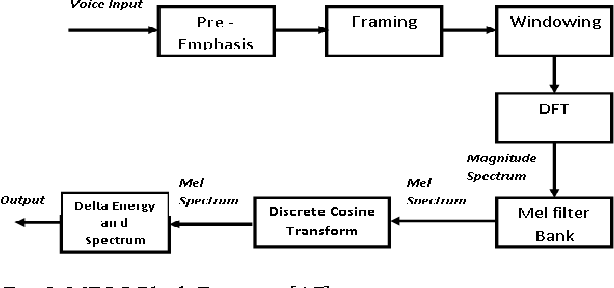
Block diagram of speech recogination

MFCC Technique

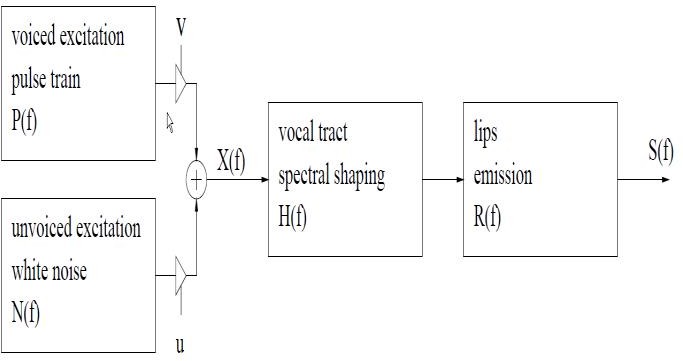
Mel Frequency Cepstral Coefficients (MFCC):

# The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. For frequencies lower than 1 kHz, human ear hears tones with a linear scale instead of logarithmic scale for the frequencies higher that 1 kHz. In other words, MFCC is based on known variation of the human ear‟s critical bandwidth with frequency. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech

# MFCC is based on the human auditory system. The human perception of the intensity of speech or audio signal with respect to energy does not follow a linear scale. Similarly the human perception of distinguishing two speech or audio signals at different frequencies 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 logarithmic spacing above 1kHz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

The speech signals have most of their energy in the low frequencies. It is also very natural to use a Mel-spaced filter bank to analyze a speech.

When producing speech sounds, the air flow from a speaker lungs first passes the glottis and then throat and mouth. Depending on which speech sound you articulate, the speech signal can be excited in three possible ways:



A simple model of speech production.

 **voiced excitation** The glottis is closed. The air pressure forces the glottis to open and close periodically thus generating a periodic pulse train (triangle–shaped). This ”fundamental frequency” usually lies in the range from 80Hz to 350Hz.

 **unvoiced excitation** The glottis is open and the air passes a narrow passage in the throat or mouth. This results in a turbulence which generates a noise signal. The spectral shape of the noise is determined by the location of the narrowness.

 **transient excitation** A closure in the throat or mouth will raise the air pressure. By suddenly opening the closure the air pressure drops down immediately. (”plosive burst”)

With some speech sounds these three kinds of excitation occur in combination. The spectral shape of the speech signal is determined by the shape of the vocal tract (the pipe formed by your throat, tongue, teeth and lips). By changing the shape of the pipe (and in addition opening and closing the air flow through your nose) you change the spectral shape of the speech signal, thus articulating different speech sounds.

## A Simple Model of Speech Production

The production of speech can be separated into two parts: Producing the excitation signal and forming the spectral shape. Thus, we can draw a simplified model of speech production as shown in Figure 3. This model works as follows: Voiced excitation is modeled by a pulse generator which generates a pulse train (of triangle–shaped pulses) with its spectrum given by P(f). The unvoiced excitation is modeled by a white noise generator with spectrum N(f). To mix voiced and unvoiced excitation, one can adjust the signal amplitude of the impulse generator (v) and the noise generator (u). The output of both generators is then added and fed into the box modeling the vocal tract and performing the spectral shaping with the transmission function H(f). The emission characteristics of the lips is modeled by R(f). Hence, the spectrum S(f) of the speech signal is given as:

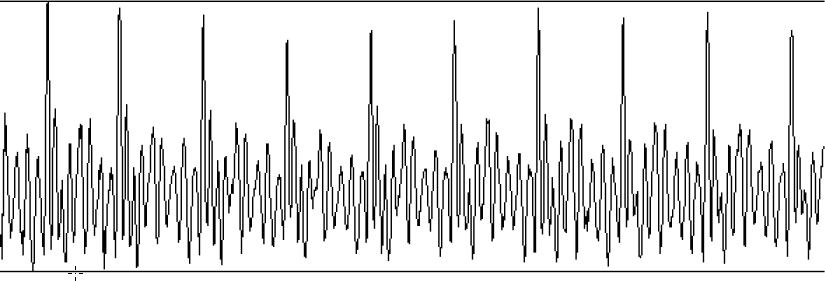
To influence the speech sound, we have the following parameters in our speech production model:

 the mixture between voiced and unvoiced excitation (determined by v and u)

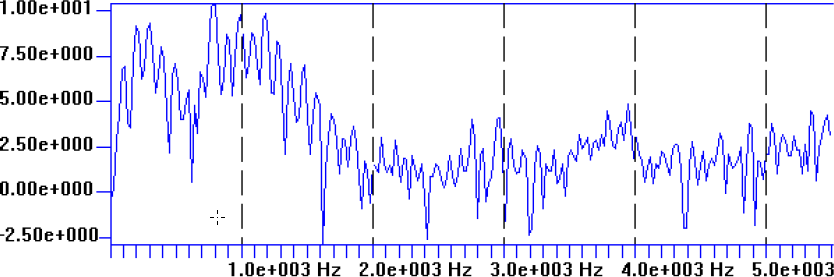
the fundamental frequency (determined by P(f))  the spectral shaping (determined by H(f))

the signal amplitude (depending on v and u)

These are the technical parameters describing a speech signal. To perform speech recognition, the parameters given above have to be computed from the time signal (this is called speech signal analysis or ”acoustic preprocessing”) and then forwarded to the speech recognizer. For the speech recognizer, the most valuable information is contained in the way the spectral shape of the speech signal changes in time. To reflect these dynamic changes, the spectral shape is determined in short intervals of time, e.g., every 10 ms. By directly computing the spectrum of the speech signal, the fundamental frequency would be implicitly contained in the measured spectrum (resulting in unwanted ”ripples” in the spectrum).



Time signal of the vowel /a:/ (fs = 11kHz, length= 100ms). The high peaks in the time signal are caused by the pulse train P(f) generated by voiced excitation.



Log power spectrum of the vowel (fs = 11kHz, N = 512). The ripples in the spectrum are caused by P(f).

## Cepstral Transformation

As shown above, the direct computation of the power spectrum from the speech signal results in a spectrum containing ”ripples” caused by the excitation spectrum X(f). Depending on the implementation of the acoustic preprocessing however, special transformations are used to separate the excitation spectrum X(f) from the spectral shaping of the vocal tract H(f). Thus, a smooth spectral shape (without the ripples), which represents H(f) can be estimated from the speech signal. Most speech recognition systems use the so–called mel frequency cepstral coefficients (MFCC) and its first (and sometimes second) derivative in time to better reflect dynamic changes.

Since the transmission function of the vocal tract H(f) is multiplied with the spectrum of the excitation signal X(f), we had those unwanted ”ripples” in the spectrum. For the speech recognition task, a smoothed spectrum is required which should represent H(f) but not X(f). To cope with this problem, cepstral analysis is used. If we look at Equation 1, we can separate the product of spectral functions into the interesting vocal tract spectrum and the part describing the excitation and emission properties

## Mel Cepstrum

As was shown in perception experiments, 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. 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 shown in Figure 3.6. The frequency groups are assumed to be linearly distributed along the mel- scale. The mel–frequency fmel can be computed from the frequency f as follows:

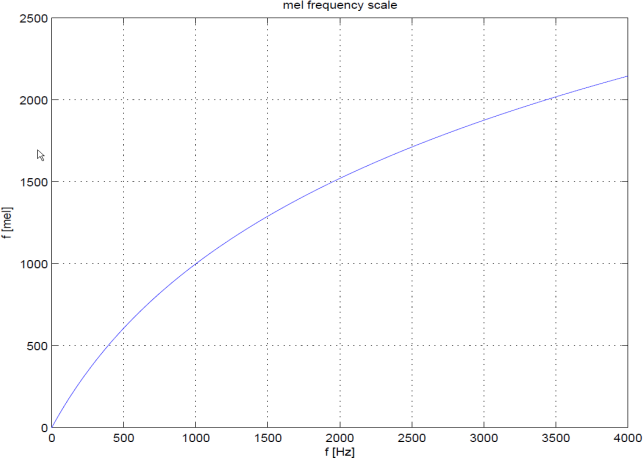
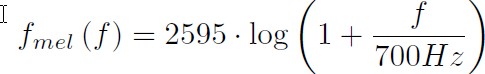


Fig. 3.6 The mel frequency scale.

The human ear has high frequency resolution in low–frequency parts of the spectrum and low frequency resolution in the high–frequency parts of the spectrum. The coefficients of the power spectrum of a speech signal can be transformed to reflect the frequency resolution of the human ear. A common way to do this is to use K triangle–shaped windows in the spectral domain to build a weighted sum over those power spectrum coefficients which lie within the window. In analogy to computing the cepstrum, we now take the logarithm of the mel power spectrum (instead of the power spectrum itself) and transform it into the quefrency domain to compute the mel cepstrum. The mel ceptral coefficients are used directly as feature vectors representing a speech signal for further processing in the speech recognition system instead of transforming them back to the frequency domain.

# DYNAMIC TIME WARPING

A speech signal is represented by a series of feature vectors which are computed every 10ms. A whole word will comprise dozens of those vectors, and we know that the number of vectors (the duration) of a word will depend on how fast a person is speaking. In speech recognition, we have to classify not only single vectors, but sequences of vectors. Lets assume we would want to recognize a few command words or digits. For an utterance of a word w which is TX vectors long, we will get a sequence of vectors X= {x0, x1, . . . , xTX−1} from the acoustic preprocessing stage. What we need here is a way to compute a ”distance” between this unknown sequence of vectors X and known sequences of vectors W = {w0,w1, . . .

,wTW} which are prototypes for the words we want to recognize.

The main problem is to find the optimal assignment between the individual vectors of unequal vector sequence X and W . In Fig. 4.1 we can see two sequences X and W which consist of six and eight vectors, respectively. The sequence W was rotated by 90 degrees, so the time index for this sequence runs from the bottom of the sequence to its top. The two sequences span a grid of possible assignments between the vectors. Each path through this grid (as the path shown in the figure) represents one possible assignment of the vector pairs. For example, the first vector of X is assigned the first vector of W, the second vector of X is assigned to the second vector of ˜W, and so on. Fig. 4.1 shows as an example the following path P given by the sequence of time index pairs of the vector sequences (or the grid point indices, respectively):



The length LP of path P is determined by the maximum of the number of vectors contained in X and W . The assignment between the time indices of W and X as given by P can be interpreted as ”time warping” between the time axes of Wand X . In our example, the vectors x2, x3 and x4 were all assigned to w2, thus warping the duration of w2 so that it lasts three time indices instead of one. By this kind of time warping, the different lengths of the vector sequences can be compensated. For the given path P, the distance measure between the vector sequences can now be computed as the sum of the distances between the individual vectors.

**MATLAB SOFTWARE**

**MATLAB** (matrix laboratory) is a [multi-paradigm](https://en.wikipedia.org/wiki/Multi-paradigm_programming_language) [numerical computing](https://en.wikipedia.org/wiki/Numerical_analysis) environment and [proprietary programming language](https://en.wikipedia.org/wiki/Proprietary_programming_language)developed by [MathWorks](https://en.wikipedia.org/wiki/MathWorks" \o "MathWorks). MATLAB allows [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) manipulations, plotting of [functions](https://en.wikipedia.org/wiki/Function_(mathematics)) and data, implementation of [algorithms](https://en.wikipedia.org/wiki/Algorithm), creation of [user interfaces](https://en.wikipedia.org/wiki/User_interface), and interfacing with programs written in other languages, including [C](https://en.wikipedia.org/wiki/C_(programming_language)), [C++](https://en.wikipedia.org/wiki/C%2B%2B), [C#](https://en.wikipedia.org/wiki/C_Sharp_(programming_language)), [Java](https://en.wikipedia.org/wiki/Java_(programming_language)), [Fortran](https://en.wikipedia.org/wiki/Fortran) and [Python](https://en.wikipedia.org/wiki/Python_(programming_language)).

Although MATLAB is intended primarily for numerical computing, an optional toolbox uses the [MuPAD](https://en.wikipedia.org/wiki/MuPAD" \o "MuPAD) [symbolic engine](https://en.wikipedia.org/wiki/Computer_algebra_system), allowing access to [symbolic computing](https://en.wikipedia.org/wiki/Symbolic_computing) abilities. An additional package, [Simulink](https://en.wikipedia.org/wiki/Simulink), adds graphical multi-domain simulation and [model-based design](https://en.wikipedia.org/wiki/Model-based_design) for [dynamic](https://en.wikipedia.org/wiki/Dynamical_system) and [embedded systems](https://en.wikipedia.org/wiki/Embedded_system).

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses include:

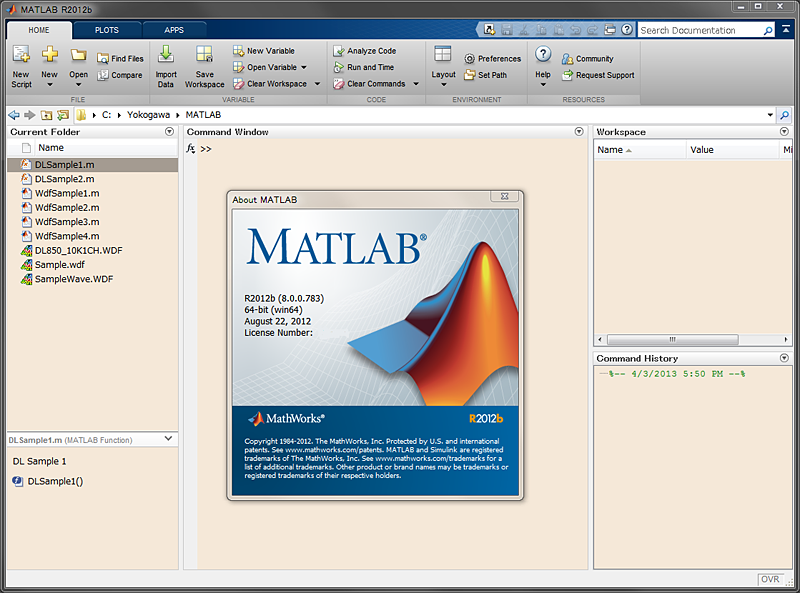
* Math and computation
* Algorithm development
* Modeling, simulation, and prototyping
* Data analysis, exploration, and visualization
* Scientific and engineering graphics
* Application development, including Graphical User Interface building

MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. This allows you to solve many technical computing problems, especially those with matrix and vector formulations, in a fraction of the time it would take to write a program in a scalar noninteractive language such as C or Fortran.

The name MATLAB stands for matrix laboratory. MATLAB was originally written to provide easy access to matrix software developed by the LINPACK and EISPACK projects, which together represent the state-of-the-art in software for matrix computation.

MATLAB has evolved over a period of years with input from many users. In university environments, it is the standard instructional tool for introductory and advanced courses in mathematics, engineering, and science. In industry, MATLAB is the tool of choice for high-productivity research, development, and analysis.

MATLAB features a family of application-specific solutions called toolboxes. Very important to most users of MATLAB, toolboxes allow you to learn and apply specialized technology. Toolboxes are comprehensive collections of MATLAB functions (M-files) that extend the MATLAB environment to solve particular classes of problems. Areas in which toolboxes are available include signal processing, control systems, neural networks, fuzzy logic, wavelets, simulation, and many others.



The MATLAB System

The MATLAB system consists of five main parts:

#### The MATLAB language.

This is a high-level matrix/array language with control flow statements, functions, data structures, input/output, and object-oriented programming features. It allows both "programming in the small" to rapidly create quick and dirty throw-away programs, and "programming in the large" to create complete large and complex application programs.

#### The MATLAB working environment.

This is the set of tools and facilities that you work with as the MATLAB user or programmer. It includes facilities for managing the variables in your workspace and importing and exporting data. It also includes tools for developing, managing, debugging, and profiling M-files, MATLAB's applications.

#### Handle Graphics.

This is the MATLAB graphics system. It includes high-level commands for two-dimensional and three-dimensional data visualization, image processing, animation, and presentation graphics. It also includes low-level commands that allow you to fully customize the appearance of graphics as well as to build complete Graphical User Interfaces on your MATLAB applications.

#### The MATLAB mathematical function library.

This is a vast collection of computational algorithms ranging from elementary functions like sum, sine, cosine, and complex arithmetic, to more sophisticated functions like matrix inverse, matrix eigenvalues, Bessel functions, and fast Fourier transforms.

#### The MATLAB Application Program Interface (API).

This is a library that allows you to write C and Fortran programs that interact with MATLAB. It include facilities for calling routines from MATLAB (dynamic linking), calling MATLAB as a computational engine, and for reading and writing MAT-files

**Key Features**

 High-level language for numerical computation, visualization, and application development

 Interactive environment for iterative exploration, design, and problem solving

 Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration, and

Solving ordinary differential equations

 Built-in graphics for visualizing data and tools for creating custom plots

 Development tools for improving code quality and maintainability and maximizing performance

 Tools for building applications with custom graphical interfaces

 Functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET, and Microsoft® Excel®

PROPOSED SCHEME

system for voice recognition using dynamic time wrapping algorithm, and comparing the entered voice signal of the speaker with pre-stored voice signals in the database for the purpose of verifying. By using good statistical method for the process of comparing. Then extracting the main features of the speaker voice signal by using Mel-Frequency Cepstral Coefficients, which is one of the most important factors in achieving high recognition accuracy. In order to solve the problem of extracting the components and features of the voice signals that entered into the computers and performing the comparison to getting the best results for maintaining the confidentiality, security and integrity of the information. The extraction of the feature done by creating source for each digital voice from a set of vocabulary that forming the sound database. Which it is a voice signal for the voice called source signal. Where each signal is divided into blocks of equal length samples from beginning to end. Then each template converted to vector attributes that extract the signal features in that template. These include vector in groups are called Features Vectors. This processing repeated for each digital voice in the vocabulary set. The Features matching is called “recognition process”, in this process the coming signal that to be recognized is transformed into a series of features vectors by using the conversion (begin – end), for processing of features extraction. These feature will be compared with all possible probability exist on the database by using pattern matching method. To give the recognition decision from matching quality by using Euclidean Distance between two series of features vectors, which one representing feature vectors of the source signal and the other feature vectors of the test signal. The proposed system moves through two phases:

# Training Phase

At this stage the system will trained by creating training groups consists of different sounds samples, from which the system can create its own sound database.



Taking the Voice Sample



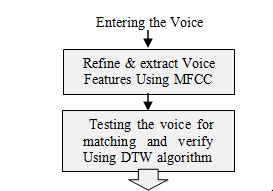
System Database

Create Training Group to all person

For the purpose of training, was taken 10 voice samples from every person, to be compared with all samples of the trainee person. Then taking the sample that have highest match with more purity, and store it on the system voice database.

# Testing Phase

At this stage the system will tested for recognizing the voice of the speaker after the matching process with samples taken from training stage. The system can take a decision whether the voice exist on the Database or not. As shown in the Fig 2



Display number

In this test phase system match with sample recorded in training folder. We record 10 samples per number up to 9.if number match with sample then shown on display output. If not match then other number display depending upon neighbour value.

For the purpose of testing recognition in the proposed scheme, they have been recorded 10 words sound for 1 person and stored in bit format with rate 11025 Hz and extension “.wav”.

Sample store in wave format.10 sample store for 1 number.so we recorded 100 sample for 0 to 9 number.this sample store automatically in train folder depending upon matlab programming.

**VERFICATION USING DTW**

This method uses distance measure to find the nonlinear matching between test signal and all source signal. Then selecting the source has less distance, and the voice that represented by that source, refers to the result of recognition. The system has been tested on 10 samples which represent the system database, and the verification result is in range from 0 to 9

ADVANTAGE

* speech recognition technology includes the dictation ability it provides. With the help of the technology users can easily control devices and create documents by speaking. Speech recognition can allow documents to be created faster because the software generally produces words as fast as they are spoken, which is generally much faster than a person can type. Dictation solutions are not only used by individuals but also by organizations which require heavy transcription tasks such as healthcare and legal.
* Businesses which provide customer services benefit from the technology in order to improve self-service in a way that enriches customer experience and reduces organizational costs
* Use for Security purpose.

**CONCLUSION**

The performance of recognition systems can be determined by voice sources including the speaker voice specification, the rate of voice issuance, delivery source, and recording media. The accuracy of recognition come from the behavior of speaker in issuing the voice. From the tests and training above we found that the DTW is flexible mathematical method, they gives high accuracy results. The performance can be improved by selecting the sources carefully, which have significant role in influencing the accuracy of recognition.

The signal analysis by using MFCC provide spectrum factors which represents the exact vocal system for stored words. MFCC provide a high level of perception of the human voice, where they work to remove all unimportant information, then give a better representation of the signal, which leads to a higher resolution in the performance of recognition.

The conclusion of spectrum factors that have high-specification show their importance depending on the speaker himself and the method of producing the voice and vocal pronunciation style which can be used in many

.

# REFERENCES

1. Shumaila Iqbal, Tahira Mahboob and Malik Sikandar Hayat Khiyal, “Voice Recognition using HMM with MFCC for Secure ATM”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011, ISSN (Online): 1694-0814, pp. 297- 303.
2. Debnath Bhattacharyya, Rahul Ranjan, Farkhod Alisherov A. and Minkyu Choi, “Biometric Authentication: A Review”, International Journal of u- and e- Service, Science and Technology Vol. 2, No. 3, September, 2009.
3. Judith A. Markowitz, “Voice Biometrics”, September 2000/Vol. 43, No. 9 Communications of the ACM.