

# Robotic Automation through Speech Recognition

Ms. Neerja S. Dharmale\*, Mr. Rupesh S. Mahamune\*\*

\* Assistant Professor, Electronics Engineering Department, Dr Babasaheb Ambedkar College Of Engineering & Research, Nagpur.

\*\* Assistant Professor, Electronics & Telecommunication Department, St. Vincent Pallotti College of Engineering and Technology, Nagpur.

**Abstract-** This paper is based on Digital signal processing. Controlling through human speech is one of the fascinating application of Digital Signal Processing (DSP) and Automation systems. The speech recognition can be defined as a technique of determining what is being spoken by a particular speaker. Speech recognition is one of the commonplace applications of the speech processing technology. The term automation refer to do the thing without hands-on human interaction. The proposed project is based on the speech recognition and it's application in control mechanism. This project involves the establishment of speech recognition system. After speech recognition, particular code related to spoken word is generated. Then this code is sent to microcontroller via wireless transceiver. Then microcontroller takes the necessary action according to the command signal. The wireless communication is provided via RS-232. The RS-232 cable provides data in the form of RS-232 level and Microcontroller is able to recognize the TTL level, so the data is converted to TTL level using MAX-232 IC.

**Index Terms-** DSP,DTW,KNN,MFCC, microcontroller , Zero-crossing, etc.

## I. INTRODUCTION

The system consist of microphone though which input in the form of speech signal is applied. The data acquisition system of the speech processor acquires the output from the microphone and then it will detect the exact word spoken. Speech processor is matlab based. Speech signal is saved as '.wav' file because it is unqiely decipherable to matlab. Matlab 7.9 is used to develop a speech recognition system. The command signal from the speech processor is generated accordingly which is then send to microcontroller via wireless transceiver. Microcontroller is used for controlling the ROBOT. Microcontroller takes necessary action according to the command signal. The generalized block diagram of system is as follow

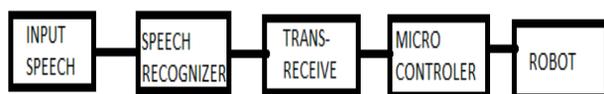


Fig. 1 System plan

## II. SPEECH RECOGNITION SYSTEM

### A. Classification Of Speech Recognition System

The goal of speech recognition is for a machine to be able to "hear," "understand," and "act upon" spoken information. This goal for now remains in the distant future. However, new advances in the field of speech recognition have shown considerable progress towards that goal. Modern speech recognition systems typically can be classified in three ways:

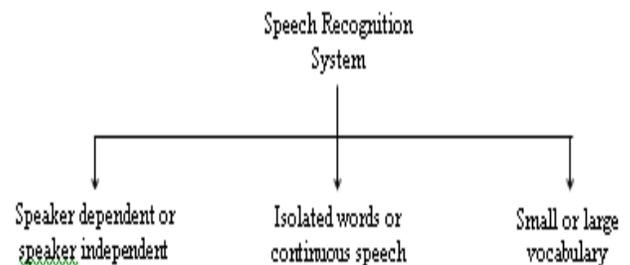


Fig 2. Classification of speech recognition system

- 1) Speaker dependent system or speaker independent system:
- 2) Isolated words or continuous speech:
- 3) Small or large vocabulary:

### B. Stages of Speech Recognition

#### 1) Training stage

Training stage involves "teaching" the system by building it's dictionary, an acoustic model for each word that the system needs to recognize.

#### 2) Testing stage

In testing stage, we will use acoustic models of these words to recognize isolated words using classification algorithm.

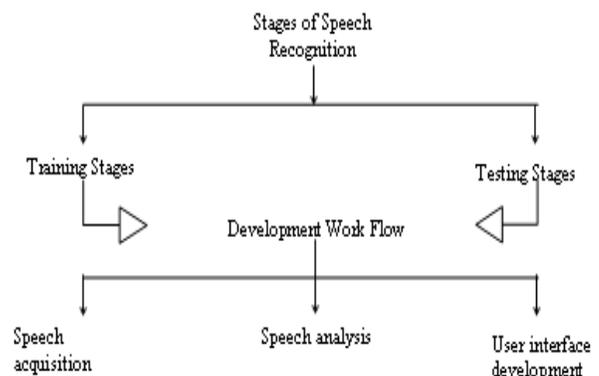


Fig 3. Development Work Flow

### III. SPEECH ACQUISITION

In training stage speech is acquire from microphone to create dictionary or words with repeated utterance of each words. GUI is created to acquire a speech. This GUI is created using matlab and data acquisition toolbox while in testing stage speech will be needed to acquire continuously and buffer speech samples and at the same time, it is needed to process speech signal frame by frame, or in continuous group of samples.

### IV. SPEECH ANALYSIS

#### A. Procedure and Algorithm for Speech Analysis

In training stage, MFCC algorithm, zero-crossing and energy based speech recognition algorithms are used to analyze the speech while in testing stage, zero-crossing and energy based speech recognition, MFCC algorithm, DTW algorithm and KNN algorithm are used for speech analysis.

#### B. Procedure in Training Stage

- 1) Perform speech endpoint trimming using Zero crossing and energy based speech Recognition algorithm.
- 2) Compute test feature vector using MFCC algorithm.
- 3) Save the result in working directory.

#### C. Procedure in Testing Stage

- 1) DC offset elimination.
- 2) Apply spectral subtraction.
- 3) Perform speech endpoint trimming using zero crossing and energy based speech recognition algorithm.
- 4) Compute test feature vector using MFCC algorithm.
- 5) Remove convolutional noise.
- 6) Calculate DTW score using DTW algorithm.
- 7) Apply KNN algorithm.

#### D. Zero-crossing rate

Zero crossing occurs if successive samples have different algebraic signs. The rate at which zero-crossings occur is used to measure the frequency content of a signal. High frequency implies high zero crossing rates and low frequency imply low zero crossing rates. If the zero crossing rate is high, the speech signal is unvoiced, while if it is low the speech signal is voiced.

#### E. Energy Content Measure

It provides a basis for distinguishing voiced speech segments from unvoiced speech segments.

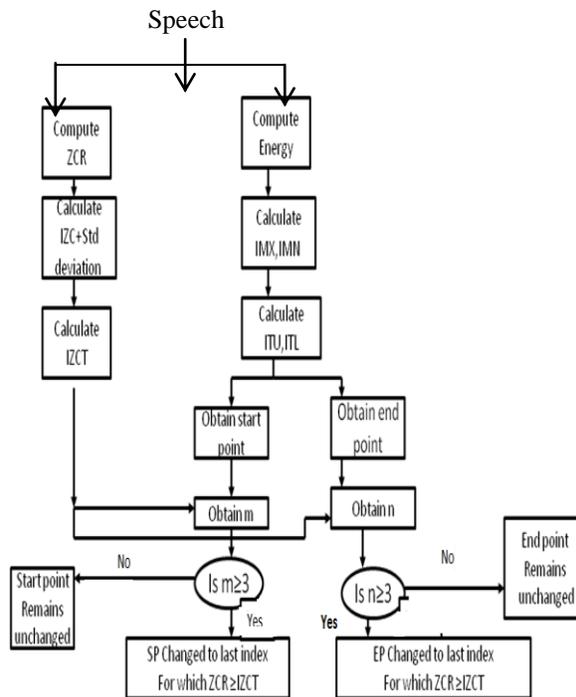


Fig 4. Flow Chart for Speech Endpoint Trimming

During the silence region of the word a zero crossing threshold,  $IZCT$ , is chosen as the minimum of a fixed threshold and the sum of the mean zero crossing rate during silence,  $IZC$ , plus twice the standard deviation of the zero crossing rate during silence, i.e.,

$$IZCT = \text{MIN}(IF, IZC + 2\sigma_{IZC})$$

The energy function for the entire interval  $E(n)$  is then computed. The peak energy,  $IMX$ , and the silence energy,  $IMN$ , are used to set two thresholds,  $ITL$  and  $ITU$ , according to the rule  $I1 = 0.03 * \{IMX - IMN\} - IMN$   
 $I2 = A * IMN$   
 $ITL = \text{MIN}\{I1, I2\}$   
 $ITU = 5 * ITL$

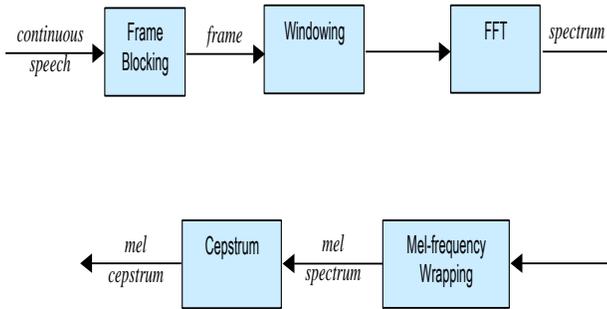
#### F. Mel-frequency cepstrum coefficients (MFCC)

MFCC (Mel Frequency Cepstral Coefficients) is the most common technique for feature extraction. MFCC tries to mimic the way our ears work by analyzing the speech waves linearly at low frequencies and logarithmically at high frequencies.

Psychophysical studies have shown that 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 a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Therefore we can use the following approximate formula to compute the mels for a given frequency  $f$  in Hz:

$$\text{Mel}(f) = 1127 \ln(1 + f/700)$$

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *mel-frequency* scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.



**Fig.5 Block diagram of the MFCC processor**

**G. Dynamic Time Warp (DTW)**

In this type of speech recognition technique the test data is converted to templates. The recognition process then consists of matching the incoming speech with stored templates. The template with the lowest distance measure from the input pattern is the recognized word. The best match (lowest distance measure) is based upon dynamic programming. This is called a Dynamic Time Warping (DTW) word recognizer.

In order to understand DTW, two concepts need to be dealt with,

- 1) **Features:** The information in each signal has to be represented in some manner.
- 2) **Distances:** some form of metric has been used in order to obtain a match path. There are two types:
  - Local:** a computational difference between a feature of one signal and a feature of the other.

**Global:** the overall computational difference between an entire signal and another signal of possibly different length.

Since the feature vectors could possibly have multiple elements, a means of calculating the local distance is required. The distance measure between two feature vectors is calculated using the *Euclidean* distance metric. Therefore the local distance between feature vector  $x$  of signal 1 and feature vector  $y$  of signal 2 is given by,

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

If  $D(i,j)$  is the global distance up to  $(i,j)$  and the local distance at  $(i,j)$  is given by  $d(i,j)$

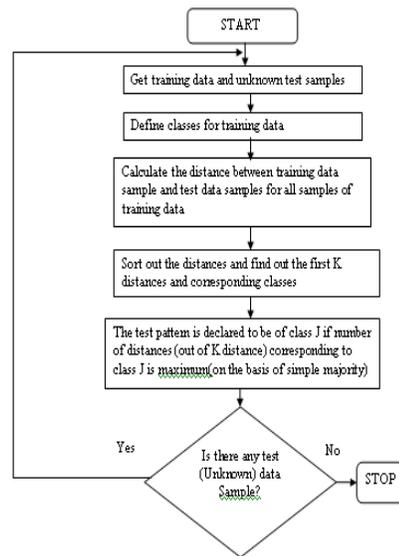
$$D(i,j) = \min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \tag{1}$$

Given that  $D(1,1) = d(1,1)$  (this is the initial condition), we have the basis for an efficient recursive algorithm for computing  $D(i,j)$ . The final global distance  $D(n,N)$  gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score.

Where,  
 $n$  is i/p speech length.  
 $N$  is o/p speech length.

**H. K Nearest Neighbor Rule**

This method consists of assigning to the unlabelled feature vector or the label of the training vector that is nearest to it in the feature space. In KNN a training set  $T$  is used to determine the class of a previously unseen sample  $X$ . A suitable distance is measured in the feature space between the unseen sample and all the samples of the training data. This distance is used to determine  $k$  element in  $T$  closest to  $X$ . and if most of these  $k$  nearest neighbours contain similar values, then  $X$  gets classified accordingly. These classification schemes clearly define nonlinear decision boundaries and thus improve the performance.



**Fig .6 Flow Chart for KNN**

**V. WIRELESS COMMUNICATION VIA SERIAL PORT**

A. Serial data is transmitted one bit at a time. Here for serial communication RS-232 serial cable is used.

B. The transferred bits include the start bit, the data bits, the parity bit (if defined), and the stop bits.

C. The communication via RS-232 line is asynchronous. This means that the transmitted byte must be identified by start and stop bits.

D. The data can be transferred as either binary data or ASCII data.

*E.* The data is then send to the microcontroller via wireless transceiver and the microcontroller takes necessary action according to the command signal.

*F.* The data bits coming from the RS-232 serial port will be in the form RS-232 level. So it will converted to TTL level by MAX-232 IC since the transceiver and microcontroller is able to recognize the TTL level.

## VI. TRANSCEIVER

*A.* It is a a combination transmitter/receiver in a single package. The term applies to wireless communication devices such as cellular telephones, cordless telephone sets, handheld two way radio, and mobile two way radio. Occasionally the term is used in reference to transmitter/receiver devices in cable or optical fiber systems.

*B.* Generally transceivers provides following three transmission modes

1) Simplex mode: In this only one device can send and other can only receive.

2) Half-duplex mode: Two devices can send and receive but not at same time.

3) Full-duplex mode: Two devices can transmit and receive at same time.

## VII. CONTROLLING THROUGH MICROCONTROLLER

*A.* Microcontroller will be used to control the system I.E ROBOT according to the command provided in form of speech signal.

*B.* The serial output coming from the receiver will be sent to microcontroller. It will accept the serial data, processes it and provide the output on one of its port accordingly.

*C.* The baud rate of microcontroller will set according to the baud rate of serial data send by computer.

*D.* Here it will be needed to configure the serial port's operation mode and baud rate.

*E.* For serial port configuration, it will be needed to write the data from serial port to serial buffer (SBUF), an special function register (SFR) dedicated to the serial port.

*F.* The interrupt service routine of microcontroller will automatically let the controller know about the reception of the serial data so that it will control the system according to the command send in the form of speech signal.

*G.* For configuring the baud rate compatible to the serial port, the timer register microcontrollers will be set according to the particular baud rate of the serial port of computer.

## VIII. LIMITATIONS

The present design of system has few limitations. Firstly the design is computer software based and it will be unable to be implemented without computer or laptap. Secondly the present system is speaker dependent so it will recognize the speech of only one speaker so it will limits its use for different person. Thirdly if we increase the no of words then it will require more no of commands which will result in higher occupation of hard disk of computer or laptap.

## IX. FUTURE ENHANCEMENTS

Recognized speech can be used to operate the distributed control systems without the interaction of the handwork and interfacing with spoken word only. Its wireless feature is able to make the control system less complicated. As it is speaker dependent the it can be used for security purpose. Also with the modification in the algorithm of the speech recognition, the system can be made speaker independent. This will makes the system flexible in its usage as everyone is able to operate the system. The implementation of the similar type of algorithm on Digital Signal Processor can make the system computer independent which can make the system portable and so the system can be used for driving cars without the interaction of the steering and clutch and gear system.

## ACKNOWLEDGMENT

I would like to acknowledge Prof. Mrs A. P. Khandait and Prof. Mrs. A. P. Rathkanthiwar. I also would like to acknowledge DBACER, Nagpur

## REFERENCES

- [1] Daryl Ning, "Developing an isolated word recognition system in MATLAB."
- [2] "Vibha Tiwari "and "Jyoti Singhai","Feature extraction technique for speech recognition" Proceeding of volume2 of international journal of electronics and computer ,January-June, 2010

## AUTHORS

**First Author** – Ms. Neerja Dharmale, M.Tech (VLSI), DBACER, Nagpur , n.dharmale@rediffmail.com.

**Second Author** – Prof. Rupesh S. Mahamune St. Vincent Pallotti College of Engineering and Technology, Nagpur – 440025., rupesh.mahamune@rediffmail.com