International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-12, Issue-6, (December 2022)
www.ijemr.net https://doi.org/10.31033/ijemr.12.6.28

Speech Recognition Robot using Endpoint Detection Algorithm

Prasanna Kumar M
Assistant Professor, Department of Electronics and Instrumentation, Dr. Ambedkar Institute of Technology, Bengaluru, INDIA

Corresponding Author: prasannakm13@gmail.com

ABSTRACT

Controlling the machines and environment with speech makes human life easier and comfortable. In this direction a robot has been designed which can easily be controlled through the speech commands given by an authorised person.

This work consists of two phases: Speech recognition and Robot control. Voice commands are given as an input, which is processed using the LabVIEW software. Speech processing is done using two algorithms: Endpoint detection algorithm and Silence removal algorithm. These algorithms differentiate the voice signal from the background noise, detect the word boundary and extract only the voiced part of the input signal and removing the background noise associated with it. The extracted voice command signal is then matched with the stored templates and on match, the code corresponding to a particular movement of robot is encoded and then transmitted to the robot controlling module via RF transmitter.

RF receiver in robot controlling module receives the transmitted signal, which is decoded and applied as an input to microcontroller. The microcontroller interprets the code and initiates the robot movement depending on the command given. By giving proper command, robot can be made to stop, move forward, backward, turn left, turn right etc.,

This robot can be deployed in hazardous environment and can be controlled by an authorised person. It may also assist disabled people to carry out their daily work with ease.

Keywords-- Speech Recognition, End Point Removal, Silence Removal, LabVIEW

I. INTRODUCTION

To be able to control and operate a device or robot by speaking to it, makes it easier to work with that device while, increasing the efficiency and effectiveness. At the most basic level, speech commands allow the user to perform parallel tasks (i.e., hands and eyes are busy elsewhere) while continuing to work with the computer, appliance, instrument or robot.

There are several ways in which humans interact with each other- speech, eye contact, gesture, facial expression and etc. Among them speech is a fundamental communication method in human-human interaction. This is because people can readily exchange information without the need for any other tool through speech. For intelligent robots to interact with people, an efficient human-robot communication interface is important. The typical situation where a speech interface can be successfully employed is captured in one or both of these broad characteristics:

- The user's hands or eyes are occupied
- The use of conventional input devices is inconvenient or undesired.

Along with its application in hazardous environment and as an aid to disabled people, the robot is capable of performing a small set of tasks such as go to places and deliver objects using its transport compartment and its ability to navigate in the environment.

Speech recognition systems have constraints concerning the style of speech they can recognize. There are three styles of speech: *isolated, connected and continuous*.

Isolated speech recognition systems can just handle words that are spoken separately. This is the constraint we have considered. This work presents the development of a voice activated command and control framework specifically for the control of robot.

II. BLOCK DIAGRAM

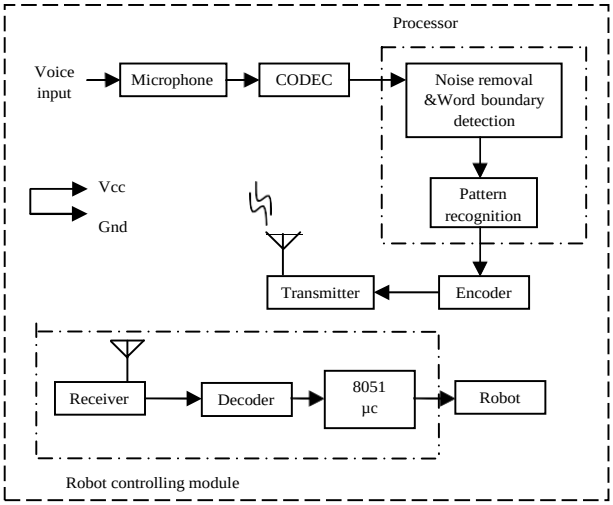

Figure 2.1: Block diagram of speech recognition robot

2.1. Microphone

The acoustic sound pressure wave is transformed into a digital signal suitable for voice processing. Here a microphone is used to acquire the voiced signal, which converts the acoustic signal into an analog signal.

2.2 CODEC:

The CODEC consists of A/D and D/A convertors. An analog signal is applied to the combination of delta-sigma modulator and decimation filter to convert it to the corresponding digital signal.

2.3 Processor:

2.3.1 Noise removal and word boundary detection:

Acquired voice signal is filtered to remove the background noise and processed to extract only the signal of interest in this block.

2.3.2 Pattern Recognition:

The signal obtained from the above explained block is then matched with the stored templates and corresponding signal is sent for transmission. Speech processing algorithms are implemented using LabVIEW embedded.

2.4 Encoder and Transmitter:

On match in the pattern recognition block, a four bit data is sent for transmission. This four bit data is encoded using the HT12E and this encoded data is sent to TLP434A for transmitting it.

2.5 Robot Controlling Module:

2.5.1 Receiver:

The transmitted signal is received by the robot controlling module which has RLP434A. Its received information is sentto the decoder.

2.5.2 Decoder:

The information so received is decoded by HT12D and the decoded data is sent to the microcontroller.

2.5.3 Microcontroller:

Microcontroller 8051 reads the decoded data and based on its value, one of the several functions in it is executed, then sends the proper sequence of pulses to the driving circuit forthe movement of the robot.

2.6 Robot:

It is constructed using two stepper motors supported by twocaster wheels.

III. METHODOLOGY

The work is divided into two phases. In first phase, Speech recognition is done and in second phase Robot's movementis controlled.

Speech recognition refers to the ability to listen (input in audio format) spoken words and recognise them as words of some known language. The steps required to make computers perform speech recognition are: word boundary detection, feature extraction and recognition.

Word boundary detection is the process of identifying the start and end of the spoken word in the given speech signal. Feature extraction includes extracting the parameters such as amplitude of the signal, energy of frequencies, pitch etc... Recognition involves the mapping of the given input (in form of various features) with one of the known signal.

In the Speech recognition phase, the voice command is picked by the microphone. This signal is processed to remove the background noise and only the signal of interestis extracted.

After the signal of interest has been derived from the input speech signal, it is now matched with the stored templates. Templates are the list of command signals which are formed taking into consideration the above mentioned parameters. On match, the corresponding code is sent to the receiving part.

Algorithms Used For Word Extraction:

An important problem in speech processing is to detect the speech in the presence of background noise. The accurate detection of a word's start and end points means thatsubsequent processing of the data can be kept to a minimum. For the efficient performance of the algorithm, a number ofspecial situations have to be taken into account such as:

- Words which begin or end with low-energy phonemes (weak fricatives).
- Words which end with an unvoiced plosive.
- Words which end with a nasal.
- Speakers ending words with a trailing off in intensityor a short breath.

In this work two algorithms for word boundary detection areused. Namely,

1. End point detection algorithm.
2. Silence removal algorithm.

3.1.1. End Point Detection Algorithm:

The algorithm uses the difference in the Zero Crossing Rate (ZCR) of the voice sample and the background noise to differentiate between them. If the ZCR of a portion of the speech signal exceeds 50, then this portion will be labelled as unvoiced or background noise whereas any segment showing ZCR around 12 is considered to be the voiced one.

The endpoint detection algorithm functions as follows:

1. The algorithm removes any DC offset in the signal. This isa very important step because the zero-crossing rate of the signal is calculated and plays a role in determining where unvoiced sections of speech exist. If the DC offset is notremoved, we will be unable to find the zero-crossing rate of noise in order to eliminate it from our signal.

2. Compute the average magnitude and zero-crossing rate of the signal as well as the average magnitude and zero-crossing rate of background noise. The average magnitude and zero- crossing rate of the noise is taken from the first hundred milliseconds of the signal. The means and standard deviations of both the average magnitude and zero-crossing rate of noise are calculated, enabling us to determinethresholds for each to separate the actual speech signal from the background noise.

3. At the beginning of the signal, we search for the first point where the signal magnitude exceeds the previously set threshold for the average magnitude. This location marks the beginning of the voiced section of the speech.

4. From this point, search backwards until the

magnitude drops below a lower magnitude threshold.

5. From here, we search the previous twenty-five frames of the signal to locate if and when a point exists where the zero- crossing rate drops below the previously set threshold. This point, if it is found, demonstrates that the speech begins with an unvoiced sound and allows the algorithm to return a starting point for the speech, which includes any unvoiced section at the start of the phrase.

6. The above process will be repeated for the end of the speech signal to locate an endpoint for the speech.

The following figures show the command signals given at the input, and the words extracted from them using the end point detection technique.

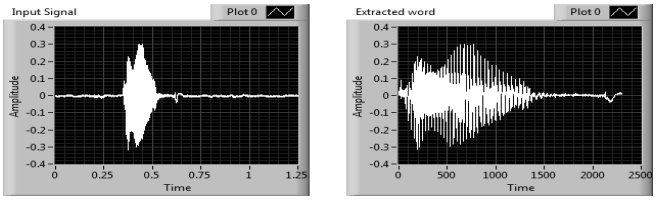

Figure 3.1.: Input and extracted signal for the command 'STOP'

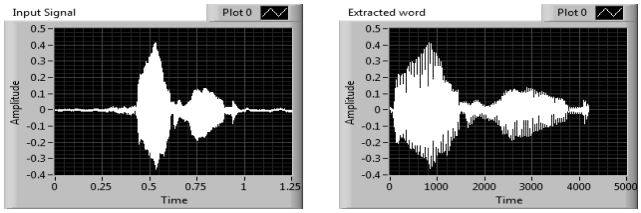

Figure 3.2.: Input and extracted signal for the command 'FORWARD'

3.1.2. Silence Removal Algorithm:

The proposed method uses Probability Density Function (PDF) of the background noise and a Linear Pattern Classifier for classification of Voiced part of a speech from silence/unvoiced part. We detect silence/unvoiced part from the speech sample using uni-dimensional Mahalanobis Distance function which itself is a Linear Pattern Classifier. The algorithm uses statistical properties of background noise as well as physiological aspect of speech production and does not assume any adhoc threshold. The Silence removal algorithm functions as follows:

1. Calculate the mean and standard deviation of the first 1600 samples of the given utterance. If μ and σ are the mean and the standard deviation respectively then analytically we can write,

$$\mu = \frac{1}{1600} \sum_{i=1}^{1600} x(i)$$

$$\sigma = \sqrt{\frac{1}{1600} \sum_{i=1}^{1600} (x(i) - \mu)^2}$$

The background noise is characterised by this μ and σ. From 1st sample to the last sample of the speech recording, in each sample check whether one-dimensional Mahalanobis distance functions i.e. $|x-\mu|/\sigma$ greater than 3 or not. Analytically, if

$$\frac{|x - \mu|}{\sigma} > 3$$

The sample is to be treated as voiced sample otherwise it is silence/unvoiced.

2. Mark the voiced sample as 1 and unvoiced sample as 0. Divide the whole speech signal into 10 ms non-overlapping windows. Consider there are M no. of zeros and N no. of ones in a window. If M ≥ N then convert each of ones to zeros and vice versa.

3. Collect the voiced part only according to the labelled '1' samples from the windowed array and dump it in a new array. Retrieve the voiced part of the original speech signal from labelled '1' samples.

The following figures show the command signals given at the input, and the words extracted from them using the Silence removal algorithm.

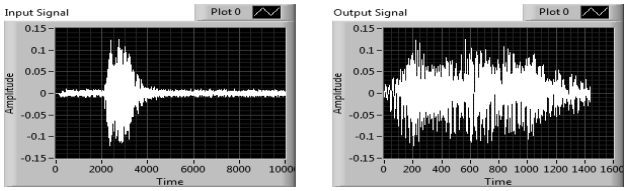

Figure 3.3.: Input and extracted signal for the command 'STOP'

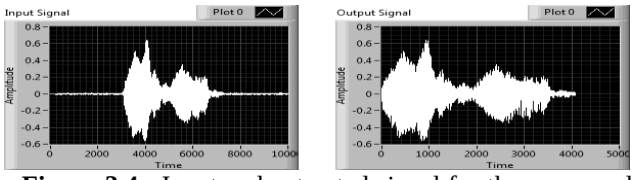

Figure 3.4.: Input and extracted signal for the command 'FORWARD'

The command is extracted from the input speech signal (time domain) following the above explained steps using one of the algorithms.

3.2. Pattern Recognition:

Pattern recognition is the process by which, given an unknown pattern, we decide which word this pattern represents.

A basic process of pattern recognition is template matching. The template method can be dependent or independent of time. The method we have implemented is independent of time. Basic disadvantage of the time domain signal is its variability between different pronunciations of the same word. People do not repeat words exactly the same, down to the minute details of amplitude. On the other hand, there is a tremendous amount of variation in the time domain signal even for the same person within several consecutive pronunciations of the same word. If we try to reduce the data rate of the spoken word, there is a risk of throwing away important information. So, the signal is converted from time domain to frequency domain to get specific number of points. The unknown

pattern is compared to a number of reference patterns stored in memory and the differences between the unknown and the references are noted. The smallest difference indicates a best match and the unknown word is the word that corresponds to the reference pattern that gavethe smallest difference.

The template generation method involves two steps. The first step is the calculation of the ambient noise threshold to eliminate noise to a large extent. The second step is to compute the voiceprint of the speech (the word spoken by the user for identification).

A Fast Fourier Transform of the incoming signal might be an easy method to compare the voice with the template. In order to achieve this, the signal is decimated by 2 and later 4000 point FFT is performed. Later, this signal is passed through a Chebyshev band pass filter (IIR) whose band width is in the normal human voice frequency range. Chebyshev filter used is of fourth order, with 0.5 dB ripple.

There exist many ways to compare two patterns. The most straightforward way is to compute the numerical difference between corresponding samples and then sum all those differences. To make sure that all differences are positive we can take the absolute value of the differences before we sum them. This is done by making any negative distances positive and leaving the positive ones alone. The square of a numberis always a positive number and this overcomes the problem of adding differences with different signs. Hence the filter output is squared and accumulated to get 4000 point dataarray.

The comparison between the 4000 point template and the 4000 point input voice command is carried out by correlation.

Correlation refers to the departure of two random variables from independence. The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. The correlation is 1 in the case of an increasing linear relationship, − 1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The degree of match is obtained using correlation index. +1 indicates maximum match and a negative value or a value close to zero indicates no match. A threshold of 0.6 is set for minimum match to take place. If, the given command has the match value greater than the threshold with any of the template, then that particular command is encoded fortransmission.

IV PERFORMANCE

The performance of a recognizer is evaluated on the basis of the percentage of the words it recognizes correctly out of the total pronounced. For a given recognizer the recognition score will vary from speaker to speaker, with the speaker's emotional condition and with the level of background noise.

The performance of both the algorithms is shown in the Table 4.1 and 4.2. We see that the performance is good when the template is stored in user 1's voice and the same user speaks. For the same templates and different user the performance is not so good.

The below table shows the performance for End Point Algorithm:

USERS	Match (%)
User 1	90
User 2	30
User 3	35
User 4	20

Table 4.1: Result obtained using End Point Algorithm.

The performance for the Silence Removal Algorithm is shown in the below table. The templates are stored in user 1's voice.

USERS	Match (%)
User 1	85
User 2	25
User 3	30
User 4	25

Table 4.2: Result obtained using Silence Removal Algorithm.

V. APPLICATIONS

5.1 Health care:

In the health care domain, even in the wake of improving speech recognition technologies, medical transcriptionists (MTs) have not yet become obsolete. Many *Electronic Medical Records (EMR)* applications can be more effective and may be performed more easily when deployed in conjunction with a speech-recognition engine. Searches, queries, and form filling may all be faster to perform by voice than by using a keyboard.

5.2 People with Disabilities:

People with disabilities are another part of the population that benefit from using speech recognition programs. The technology has improved and there are wheel chairs that canbe controlled by the person's hand through a 'joystick'- like control on the chair. It would be advantageous if the person could control the wheel chair through his voice commands.

5.3 Door locking system:

Voice controlled door locking system takes in voice input as password to open or close the door. Since speech recognition depends on pitch and duration of spoken word, the recognition rate varies for each individual. Therefore the door responds only when the person whose voice print is already stored in memory, gives the command.

5.4 Robot in Hazardous Environment:

Human presence is not safe in every circumstance or environment. For example, during gas leakage it is very obvious that a human cannot enter that

place without harming himself. In such situations a robot can enter that place and give information that is required. For the above case of gas leakage, a camera or an appropriate sensor attached to the robot can be used to collect the information.

VI. CONCLUSION

In this work two algorithms- End point Detection and Silence removal are successfully implemented and the rate of match using both the algorithms is almost the same. The efficiency of match of spoken word with the stored template is around 80% which is moderately good. Graphical programming in LabVIEW is of great help as using it is simple and easy to program. The RF communication used in this work covers the range of 10mts. Based on the command received the robot is precisely controlled by microcontroller to which it is interfaced. The overall performance of the work is very good and there's scope for further enhancements as well.

REFERENCES

[1] G. Saha, Sandipan Chakroborty, Suman Senapati. *A new silence removal and endpoint detection algorithm for speech and speaker recognition applications.*
[2] Joseph P. Campbell. *Speaker recognition: A tutorial*.
[3] *An algorithm for determining the endpoints for isolated utterances*.
[4] L.R. Rabiner & M.R. Sambur. (1975 Feb). *The Bell System Technical Journal*, *54*(2), 297-315.
[5] Dynamic programming algorithms in speech recognition. *Titus Felix Furtună, Academy of Economic Studies, Bucharest*.
[6] Jean-Claude Junqua, Brian Mak & Ben Reaves. *A robust algorithm for word boundary detection in the presence ofnoise*.
[7] Bishnu S. Atal & Llawrence R. Miner. *A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition.*
[8] Muhammad Ali Mazidi, Janice Gillispe Mazidi & Rolin D. McKinlay. *The 8051 microcontroller and embedded systems using assembly andC.*