Voice Controlled Automation System

Muhammad Salman Haleem
Department of Electronic Engineering
NED University of Engineering and Technology
Karachi, Pakistan
salmanhaleem85@gmail.com

Abstract—**In this era of technology, rapid advancements are being made in the field of automation and signal processing. The developments made in digital signal processing are being applied in the field of automation, communication systems and biomedical engineering. Controlling through human speech is one of the fascinating applications of digital signal processing (DSP) and Automation Systems. This paper discusses the speech recognition and its application in control mechanism. Speech recognition can be used to automate many tasks that usually require hands-on human interaction, such as recognizing simple spoken commands to perform something like turning on lights or shutting a door or driving a motor. After speech recognition, a particular code related to spoken word is transferred through wireless communication system to 8051 microcontroller and it works accordingly. Many techniques have been used to compare the pattern of the speech signals and recent technological advancements have made recognition of more complex speech patterns possible. Despite these breakthroughs, however, current efforts are still far away from 100% recognition of natural human speech. Therefore, the project is considered involving processing of a speech signal in any form as a challenging and rewarding one.**

Keywords-speech processing through MATLAB; frequency warping; 8051 microcontroller application; serial communication; pattern recognition; data acquisition system.

I. INTRODUCTION

The term speech recognition can be defined as a technique of determining what is being spoken by a particular speaker. It is one of the commonplace applications of the speech processing technology. The project taken under consideration involves speech recognition and its application in control mechanism. It establishes a speech recognition system which controls a system at a remote area via transceiver, for example controlling of a wireless car using voice [6]. Similarly, in control and instrumentation industry devices can be controlled from a control center that is responsible for emanated commands.

Recognizing natural speech is a challenging task. Human speech is parameterized over many variables such as amplitude, pitch, and phonetic emphasis that vary from speaker to speaker [1]. The problem becomes easier, however, when we look at certain subsets of human speech.

II. PROJECT OVERVIEW AND FEATURES

A. Project Plan

The system consists of microphone through 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 detects the exact word spoken. The command signal from the speech processor is generated accordingly which is then send to the microcontroller via wireless transceiver and the microcontroller takes necessary action according to the command signal. Suppose we want to drive an electric motor and we want that the motor rotates anticlockwise then by speaking the word anticlockwise we can rotate the motor in the same direction. Similarly by speaking the word clockwise, we can rotate the motor accordingly. If such a system is installed in a motor car, then by using several commands like left, right, start, stop, forward, backward etc; we can drive a car without using even out hands.

B. Dependency on the speaker

The system is speaker independent i.e. it is able to recognize the words spoken by anyone and it is not bound to the particular speaker [1]-[4]. This makes the flexibility in its usage as everyone is able to operate the system quite efficiently.

C. Software based Speech processor

The speech processor is software based and the algorithms are constructed on MATLAB 7. The input applied via microphone is recorded in PC as a '.wav' file. Since the '.wav' is decipherable to MATLAB so it can be processed by MATLAB very easily and necessary signals can be generated.

Figure 1. System Plan

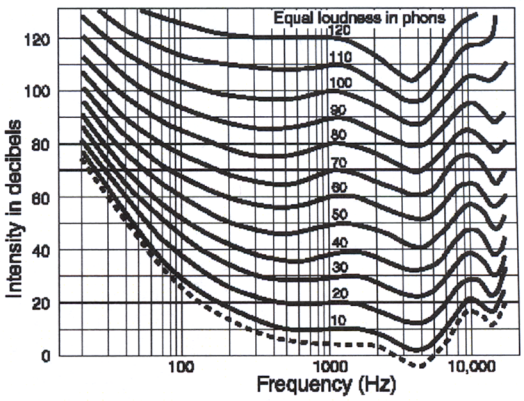

Figure 2. Loudness Curves

978-1-4244-2824-3/08/$25.00 ©2008 IEEE

III. SPEECH PROCESSING

A. Basic attributes to human hearing

In human hearing, pitch is one of the most important attributes of audio signals. According to the Place Theory of hearing [8], the inner ear acts like a frequency analyzer, and the stimulus reaching our ear is decomposed into many sinusoidal components, each of which excites different places along the basilar membrane, where hair cells with distinct characteristic frequencies are linked with neurons. When a particular frequency of vibration hits the ear, the hair cells that are tuned to that frequency will fire nerve impulses.

The second most important attribute related to human hearing is the loudness and the intensity of the sound signal. According to Psycho acoustical experiments made in the 1930's by Fletcher and Munson [11], the perceived loudness is not simply a function of intensity, but also of the sound's frequency. It can be shown in the form of graph as shown in figure 2.

The above curves show that the ear is not equally sensitive to all frequencies, particularly in the low and high frequency ranges. The curves are lowest in the range from 1 to 5 kHz, with a dip at 4 kHz, indicating that the ear is most sensitive to frequencies in this range. The lowest curve represents the threshold of hearing while the highest curve represents the threshold of pain.

B. Theory of Bark Scale

According to psychoacoustic measurements that seemed to demonstrate human inner ear frequency resolution [11], early speech researchers designed an overlapping bank of bandpass filters which could mimic the frequency response of the human cochlea (inner ear). The bandpass filters were tuned to different frequencies [12], and the passbands were made similar to the observed bandwidths of the human ear. To simulate the frequency resolution of ears, the filter consists of bandpass filters whose center frequencies are arranged nonlinearly on the frequency axis as critical bandwidth of hearing is not constant and the lower central frequencies have higher resolution and lower bandwidth while the higher frequencies have the reverse. The bandpass filters also have different bandwidths, which are thought to account for our ears' limited spectral resolution at high frequencies. On the basis of Place Theory [8], Bark Scale was constructed by early researchers. The equation of the Bark Scale is given as:

$$B = \frac{26.81}{1+\frac{1960}{f}} - 0.53 \qquad (1)$$

C. Difference between Two voice Signals

An uttered voice can differ from a stored template due to interference, noise, and other magnitude distortions which corrupt the input signal and can make it sound different from the reference signal. Also, unexpected pauses, unusually fast or slow speaking styles, and other changes in speed can randomly shift the position of the input relative to the template The same person can utter the same word in slightly different ways each time [1]-[4]. The person can pause, speak faster, speak slower, or emphasize certain syllables. These differences are called intra-speaker differences. The differences between the same words uttered by the different speakers or different words uttered by same speaker or different speakers are called inter-speaker differences [6]. These differences are large as compared to intra speaker differences.

D. Components of Speech Recognition System

The speech capturing device consists of a microphone and an analog to digital converter which digitally encodes the raw speech waveform. The DSP module performs the endpoint (word boundary) detection to separate speech from non-speech and convert the raw waveform into a frequency domain representation. It also performs further windowing, scaling, filtering and data compression. The goal is to enhance and retain only those components of the spectral representation that are useful for recognition purposes, thereby reducing the amount of information that the pattern-matching algorithm must contend with. A set of these speech parameters for one interval of time is known as a speech frame [7].

The pre-processed speech is buffered for the recognition algorithm in preprocessed signal storage. Stored reference patterns can be matched against the user's speech sample once

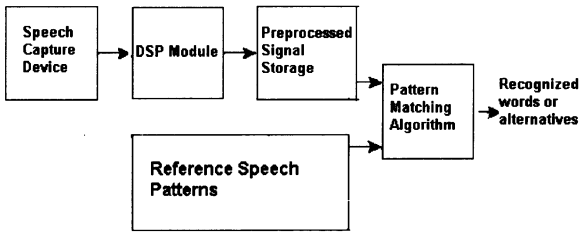

Figure 3. Basic Components of the Speech Recognition System

it has been pre-processed by the DSP module. This information is stored as a set of speech templates or as generative speech models. The algorithm must compute a measure of goodness-of-fit between the pre-processed signal from the user's speech and all the stored templates or speech models. A selection process chooses the template or model with the best match.

E. Pattern matching of speech signals

The comparison of two speech signals is nothing but basically their pattern matching. The speech signal can be represented as the set of numbers representing certain features of the speech that is to be described. For further processing it is useful to construct a vector out of these numbers by assigning each measured value to one component of the vector [1]. As an example, consider an air conditioning system which will measure the temperature and relative humidity in an office. If those parameters are measured every minute or so and the temperature is put into the first component and the

humidity into the second component of a vector, a series of two–dimensional vectors will be obtained describing how the air in the office changes as a function of time. Since these so–called feature vectors have two components, the vectors can be interpreted as points in a two–dimensional vector space. Thus, a two–dimensional map of the measurements can be drawn and it is shown in figure 4. Each point in the map will be a representative of the temperature and humidity in the office at a given time. In the map the comfortable value–pairs are shown as points labelled"+" and the less comfortable ones are shown as"-".

F. Techniques for Pattern Comparison

There are several techniques for comparing the pattern of two speech signals. Dynamic time warping is an algorithm that is used for measuring the similarity between two sequences which may vary in time or speed. Similarly another form of pattern matching that is normally used with the

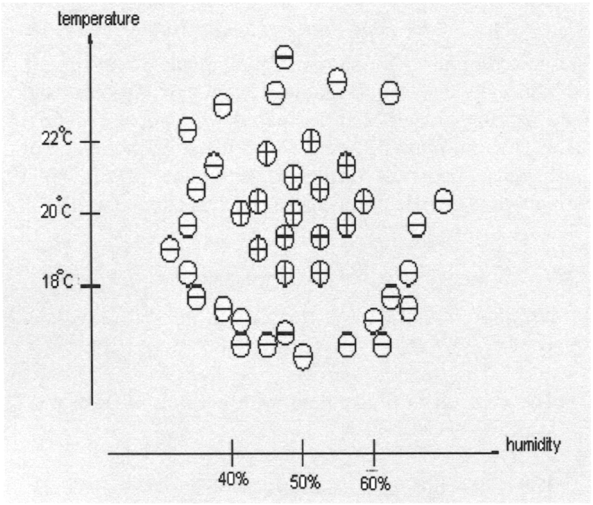

Figure 4. A map of feature vectors.

speaker verification systems utilizes multiple templates to represent frames of speech, is referred to as vector quantization (VQ) [1].

The method of Frequency Warping on which this project is based is the most accurate among the three techniques. The method of frequency warping for speech recognition is based on the fact that the voice signal acts as a stimulus to human hearing decomposes into many sinusoidal components with each component having a distinct frequency. The components are divided into critical bandwidths of hearing with each other and having a particular center frequency [12].

IV. Frequency Warping as a Speech Comparison Technique

A. Feature Extraction

Before speech pattern comparison, the features of the spoken words will be extracted. For this, a function is constructed using Data Acquisition Toolbox [9]. During voice recording, some noise or no spoken signal is also stored before and after the actual voice sample. So it is removed for real analysis. As discussed above that the critical bandwidth of human hearing is not constant, so computer can not analyze the whole speech signal instantaneously but the algorithm is set in such a way that it will select the particular number of samples from the entire set of samples of the voice signals. The feature interval of the voice signal which the computer processor will select at a time is 32ms. Succeeding feature intervals will be taken by the combination of half of the interval of previous break apart and the new samples for rest of the interval [6]. . Commonly used measuring intervals are from 20-40 ms. In the frequency domain, shorter intervals give you good time resolution but poorer frequency resolution, and longer intervals give you poorer time resolution but better frequency resolution. 32 ms was selected as a compromise because it has a feature-length short enough to resolve individual sound details, but long enough to process the signal quickly. For the particular feature interval, two time domain and 24 frequency domain features will be extracted. The feature extraction is done with all the

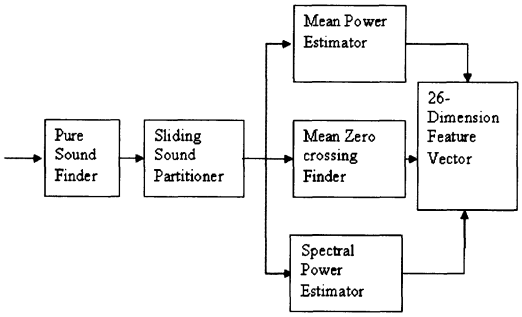

Figure 5. Block Diagram of Feature Extraction

speech templates to be stored for reference and the real time command signal. Noting that for each command, at least five examples are stored and their mean is taken as the reference template for the particular command [6].

1) Time Domain Features: The two time-domain features which were extracted are mean-power and the mean zero crossing of the feature chunk [7]. The mean power for a signal **g(t)** over an interval **N** is simply given by:

$$p(t) = \frac{1}{N}\sum_{t=0}^{t=N} g^2(t) \qquad (2)$$

Mean zero-crossings is the average number of times a signal crosses the zero-axis over an interval.

2) Frequency Domain Features: For frequency-domain features, it is better to weight the samples with a Hamming window. Hamming window is used for amplitude weighting of

the time signal used with gated continuous signals to give them a slow onset and cut-off in order to reduce the generation of side lobes in their frequency spectrum [6]. Then absolute Fast Fourier Transform of the weighted sample is taken to form frequency scale vector.

The Bark scale goes up to a maximum value of index of 24, so a filter bank based on the Bark scale utilizes 24 bandpass filters which are centered on the Bark center frequencies and whose bandwidths are equal to Bark scale critical bandwidths. Because these bandpass filters generally overlap, a triangular weighting scheme is applied to each filter in order to give the center frequency the greatest weight [12]. The plot of the filter bank that we have used in this project is shown in figure 7.

The frequency scale vector is then divided according to the Bark bandpass filters. Then each set of frequencies is triangularly weighted, and the base-10 log power of the spectrum is calculated over each filter interval. Then finally individual power values are concatenated together to form a single 24-element feature vector.

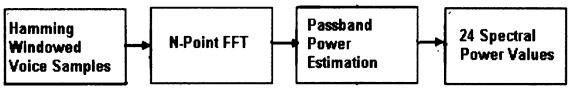

Figure 6. Block Diagram of the Frequency Domain Feature Extractor

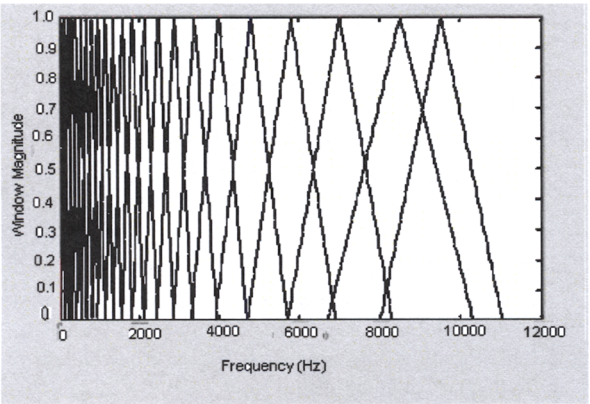

Figure 7. Bark Scale Filter Bank

B. *Final Template Making*

Before the recognition engine can recognize new voice inputs, it must first be able to recognize the original data provided for its training. This process is called "template creation" and needs to be performed only once for a particular training set. For comparing two speech signals spoken by same person, the formula used is given as;

$$\text{Error} = \frac{\text{CurrentValue} - \text{ActualValue}}{\text{ActualValue}} \quad (3)$$

The error template is formed by including the minimum error values resulting by the comparison of each value of every two example.

For calculating inter speaker differences, the procedure is same except the formula used which is known as Euclidean distance formula [1] and is given as:

$$D = \sqrt{sum(V_1 - V_2)^2} \quad (4)$$

The values coming from both types of calculation combine to form the final template for each word.

C. *Real Time Pattern Comparison*

The voice signal given as command signal in real time is passed through the same feature extraction process and template making process and compared with the stored final template. The minimum difference with the stored template of the particular word will result in generation of the code indicating that the particular word has been spoken.

V. WIRELESS COMMUNICATION VIA SERIAL PORT

By definition, serial data is transmitted one bit at a time. The order in which the bits are transmitted is given below:

- The start bit is transmitted with a value of 0, that is, logic 0.
- The data bits are transmitted. The first data bit corresponds to the least significant bit (LSB), while the last data bit corresponds to the most significant bit (MSB).
- The parity bit (if defined) is transmitted.
- One or two stop bits are transmitted, each with a value of 1, which is logic 1.

The number of bits transferred per second is given by the *baud rate*. The transferred bits include the start bit, the data bits, the parity bit (if defined), and the stop bits [5].

The communication via RS-232 line is asynchronous. This means that the transmitted byte must be identified by start and stop bits. The start bit indicates when the data byte is about to begin and the stop bit(s) indicates when the data byte has been transferred. The process of identifying bytes with the serial data format follows these steps:

- When a serial port pin is idle (not transmitting data), then it is in an "on" state (logic1).
- When data is about to be transmitted, the serial port pin switches to an "off" state (logic 0) due to the start bit.
- The serial port pin switches back to an "on" state (logic 1) due to the stop bit(s). This indicates the end of the byte.

The data bits transferred through a serial port might represent device commands, sensor readings, error messages, and so on. The data can be transferred as either binary data or ASCII data. ASCII equivalent will be transferred if the data is in the form of alphabetic characters.

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

VI. Controlling Through Microcontroller

ATMEL 89C51 microcontroller is used to control the system according to the command provided in the form of speech signal. The serial output coming from the receiver is

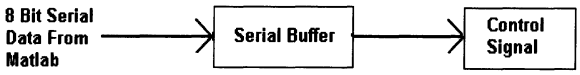

Figure 8. Serial Data Execution

sent to 89C51 (8051) microcontroller. The microcontroller accepts the serial data, processes it and provides the output on one of its ports accordingly. The baud rate of microcontroller is set according to the baud rate of the serial data send by computer.

One of the 89C51's many powerful features is its integrated UART, also known as a serial port [10]. The fact that the 8051 has an integrated serial port means that one may be able to read from and write values to the serial port very easily. Here it is needed to configure the serial port's operation mode and baud rate. For serial port configuration, what has to be done is to write the data from the serial port to Serial Buffer (SBUF), an Special Function Register (SFR) dedicated to the serial port. The interrupt service routine of the 89C51 will automatically let the controller know about the reception of the serial data so that it can control the system according to the command send in the form of the speech signal. For configuring the baud rate compatible to the serial port, the timer registers of the microcontrollers are set according to the particular baud rate of the serial port of the computer.

VII. Limitations

The present design of the system has few limitations. Firstly the design is computer software based and it will be unable to be implemented without the computer. This makes it unacceptable for the portable devices. Secondly, for more accurate recognition and to make it speaker independent, the more speech samples of the particular command spoken by the different persons have to be stored. This makes not only higher occupation of the Hard Disk of the computer but also makes the processing time of the algorithm high which increases the system's delay and makes it slow. Also the higher number of commands will result in the same problem.

VIII. Future Enhancements

Voice controlled automation system can be used to operate the Distributed Control Systems without the interaction of the handwork and interfacing with the spoken words only. Its wireless feature is able to make the control system less complicated. Also with the modification in the algorithm of the speech recognition, the system can be made speaker dependent which can be used in automatic speaker verification system. The implementation of the similar type of algorithm on the 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 the clutch and gear system.

Acknowledgment

The author would like to acknowledge the NED University of Engineering and Technology for this project.

References

[1] B. Plannerer, "An Introduction to Speech Recognition," March 2005.
[2] Richard D. Peacocke and Daryl H. Graf, "An Introduction to Speech and Speaker Recognition," Computer archive, Volume 23, August 1990, pp.26-33.
[3] Joseph P. Campbell, "Speaker Recognition: A Tutorial,"Proceedings of the IEEE, Volume 85, Issue 9, Sep 1997, pp.1437 – 1462.
[4] D.A. Reynolds, "An overview of automatic speaker recognition technology," Speech and Signal Processing, Volume 4, 2002, pp.4072 - 4075.
[5] Craig Peacock, "Interfacing the Serial/RS-232 Port," 15th June 2005.
[6] Adrian Abordo and Jon Liao, "Voice Command Recognition," June 2003.
[7] Joseph Picone, "Signal Modeling Techniques in Speech Recognition," Proceedings of IEEE, June 1993.
[8] Kyogu Lee, "Pitch Perception: Place Theory, Temporal Theory, and Beyond," 2004
[9] Brian D. Storey, "Using the MATLAB Data Acquisition Toolbox."
[10] Scott Mackenzie, "The 8051 Microcontroller," Prentice Hall, 1999
[11] H.Fletcher and W.A. Munson, "Loudness, its definition, measurement and calculation," J. Acoust. Soc. Amer., Volume 5, October 1933, pp.82-108.
[12] Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go, and Chungyong Lee, "Optimizing Feature Extraction for Speech Recognition," IEEE transactions on speech and audio processing, Volume 11, no. 1, January 2003