SPEECH CONTROLLED CAR

Session 2006 -2010

Project Advisors
Mr. Farhan Iqbal
Mr. Mubashir Baig
Mr. Asmar Khan Azar

Sumitted By

Basit Saeed	060820-018
Muhammad Ahmad	070920-037
Numan Ahmed	071020-215

Department of Electrical Engineering
University of Management and Technology
Lahore

A report submitted to the

Department of Electrical Engineering

in partial fulfillment of the requirements of

Degree

Bachelor of Science

in

Telecommunication Engineering

by

Basit Saeed

Muhammad Ahmad

Numan Ahmed

University of Management and Technology

March 24, 2011

Acknowledgments

We are graceful to the Allah Almighty who provides all the resources of every kind to us, so we make their proper use for the benefit of mankind. May He keep us with all the resources and the guidance to keep helping the humanity.

We would also like to thank Mr. Asmar Khan Azar, Mr. Farhan Iqbal and Mr. Mubashir Baig for his guidance and encouraging us to work hard and smart. He has been a constant source of guidance throughout the course of this project. His critical comments on our work have certainly made us think of new ideas and techniques. We are also thankful to our friends and families whose silent support led us to complete our project.

Advisor: _____

Co-Advisor: _____

Basit Saeed	060820-018	_____
Muhammad Ahmad	070920-037	_____
Numan Ahmed	071020-215	_____

Date

March 24, 2011

Abstract

An increasingly popular way to interact with machines is to simply talk to them. However, there is often a trade-off between ease of use and system complexity. Thus, the main objective of this project is to design and implement a voice control car using RF technology for wireless data transmission it is capable of accurately identifying a single sound while remaining simple and fast. For this purpose, an algorithm like Linear Predictive Coding is prototyped and tested using MATLAB. On the hardware side we interfaced RF module with Microcontroller, speech recognition is performed using PC, Microcontroller then convert bits from serial port to message and transmit it, receiving side RF respond and order car to move, Speech is a complex and non stationary signal produced as a result of several and complex transformations, sound signal changes a lot due to many factors so it is not possible to have 100% accurate result, Result of our system are more than 70 percent, The system as though looks astonishing for lay man but we are hoping that this could be a small step for helping blind and disable people to solve their everyday problems.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Speech recognition has emerged a lot during past few years and project like voice controlled browser, wheel chair, speech control home appliances, and most recently used by Microsoft for its upcoming gaming machine to control its power and software menu through human voice, and ours project voice control robotic car are astonishing for lay man. We design a car which could be controlled by human voice. Basic steps for the project are: first sound is converted from analog to digital form then voice is converted from time domain to frequency domain using Fast Fourier Transform finally matching is done using LPC and HMM. First we will train speaker voice by taking average samples and then these fingerprints will be compared if it matches. Matlab or Visual Studio will send bit through Serial Port to RF module via Microcontroller, RF will convert bit from Microcontroller to message and send it to the Receiver, and on the other hand Microcontroller will check the received message and move the Car.

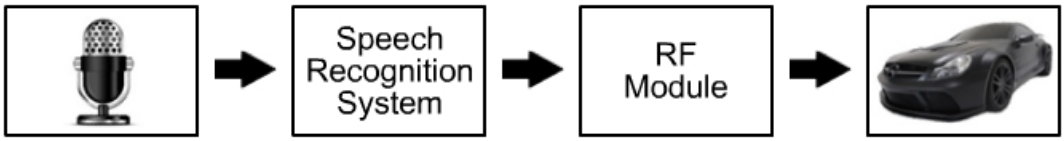

Figure 1.1: General System

Speech Recognition converts spoken words to machine-readable input; the human hearing system is capable of capturing noise over a very wide frequency spectrum, from 20 Hz on the low frequency end to upwards of 20,000 Hz on the high frequency end. The human voice, however, does not have this kind of range. Typical frequencies for the human voice are on the order of 100 Hz to 2,000 Hz. Human voice slightly changes with mood, speaking after sleeping, illness, with environment etc. So it is not an easy task to get 100 percent or even 90 percent perfect results each and every time.

1.1 Objectives:

- Voice Acquisition
- A/D Conversion
- Fourier Transform
- Linear Predictive Coding
- Hidden Markov Models
- Serial Port Interfacing
- Interfaced Microcontroller with RF Module on both transmitting and receiver side
- Car Controlling

1.2 How Voice is Generated?

Speech organs are located in the mouth and throat when we speak air push out from lungs through the larynx and epiglottis by breaking the vocal cords producing a continuous tone whose pitch can be changed by shape of larynx.

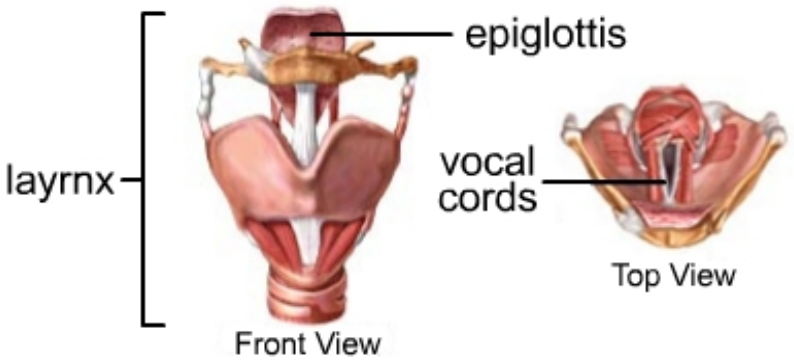

Figure 1.2: Human Neck

Voice production occurs in larynx during breathing vocal cords are separated but during speech the cartilages of larynx comes closer by the action of muscles as shown in the figure above. The tension of vibrating cord changed when cartilages comes closed to each other and this alters the pitch of the spoken sound this way high nodes are produce by the vibration of tight vocal cords and low nodes are produced by vibrating loose cords. [1]

1.3 Sound:

Sounds are pressure waves of air. If there wasn't any air, we wouldn't be able to hear sounds. When you clap your hands, the air that was between your hands is pushed aside. This increases the air pressure in the space near your hands, because more air molecules are temporarily compressed into less space. The high pressure pushes the air molecules outwards in all directions at the speed of sound, which is about 340 meters per second. When the pressure wave reaches your ear, it pushes on your eardrum slightly, causing you to hear the clap. [2]

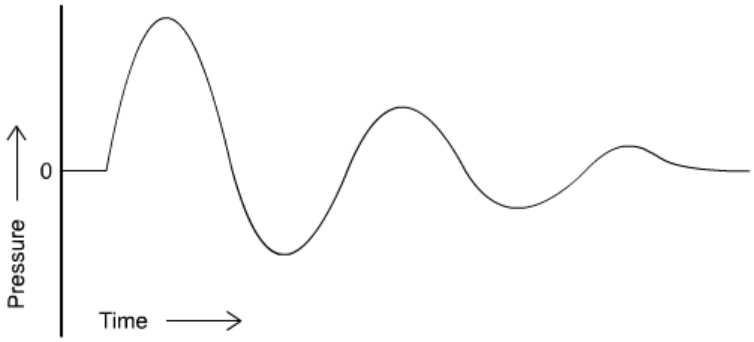

Figure 1.3: Sound Wave

1.3.1 Analog Signal:

Analog signals were first used in the 1800's; it is continuous electrical signals that vary in time. Analog signals are continuous where digital signals are discrete. [2]

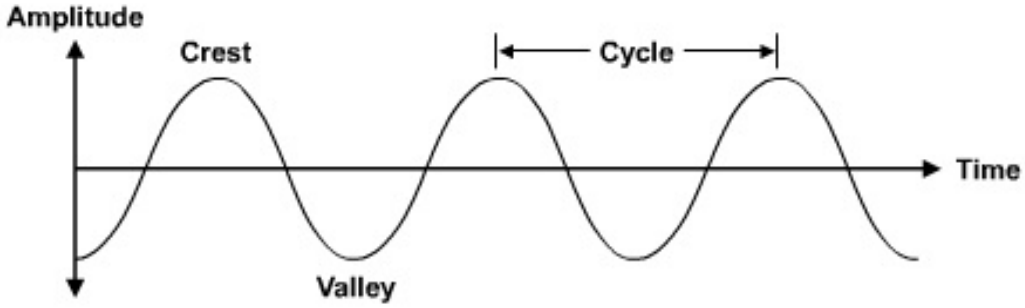

Figure 1.4: Analog Signal

Its two important characteristics are Frequency and Amplitude. Frequency is measured in Hz or cycles per second. Humans can hear frequencies between 20 Hz and 20,000 Hz (20 KHz). Amplitude is measured in decibels; Sound is simply pressure waves in air converted to electrical signals by a microphone. Disadvantage of analog signal are signals may fade with time and distance, signals may get combined with interference from other sources (static) etc.

1.3.2 Digital Signal:

With a digital signal, we are using an analog signal to transmit numbers, which we convert into bits and then transmit the bits. Suppose we want to transmit the number 6. In binary, that number is 110. We first decide that, "high" mean a 1 and "low" mean a 0. Thus, 6 might look like:

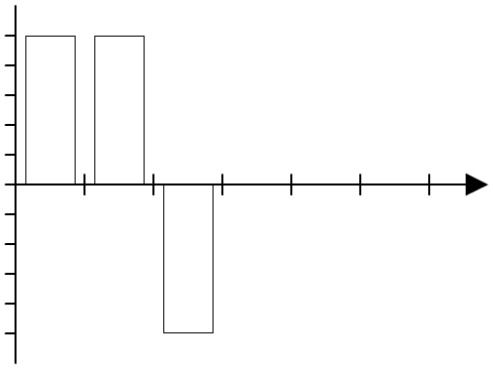

Figure 1.5: Digital Signal

The black line is the signal, which rises to the maximum to indicate a 1 and falls to the minimum to indicate a 0. [2]

1.3.3 Degradation and Restoration of Digital Signal:

One the basic advantage of digital signal over analog is that it can be restored because we know that each bit is either 0 or 1. Thus, the previous signal might be degraded to the following:

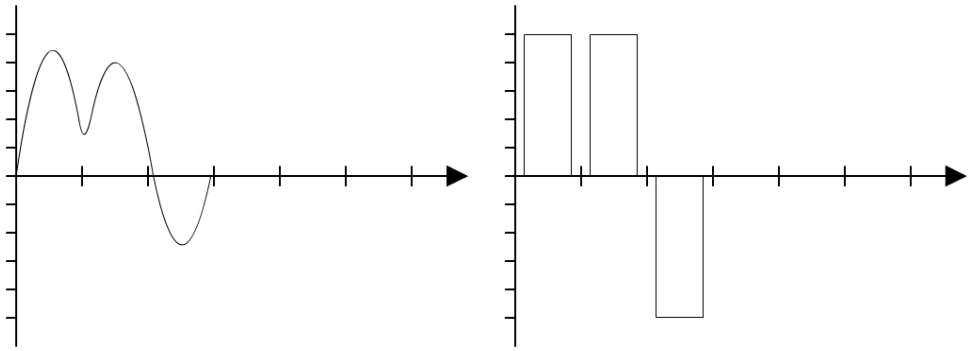

Figure 1.6: Analog Degraded and Digital Signal

Despite the general noise of the signal, we can still figure out which are the 0s and which are the 1s, and restore it. This restoration isn't possible with analog signals, because with analog there aren't just two possibilities. Analog signals are continuous where digital signals are discrete. Analog signals are continuously varying where digital signals are based on 0's and 1's (ON's and OFF's). As an example, consider a light switch that is either ON or OFF (digital) and a dimmer switch (analog) that allows you to vary the light in different degrees of brightness. [2]

1.4 Sampling:

Analog signals are continuous in time. In digital signal processing, we do not use the whole analog signal but replace it by its amplitudes taken at regular intervals. This is sampling. The problem is we must sample the signal so that from the samples we can reconstruct the original analog signal perfectly. In short sampling is taking discrete samples after regular time interval. [2]

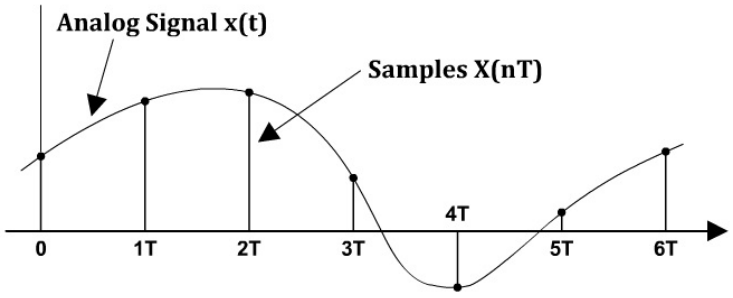

Figure 1.7: Sampling

1.4.1 Quantization:

Quantization is defined as the process of converting an analog signal to digital representation. It is performed by an analog to digital converter. If we convert analog signals to digital data advantage is that we can manipulate or calculate through powerful

computer and software's. To do this we must sample our analog waveform at well defined discrete times so we can maintain a close relationship between time in analog domain and time in digital domain. If we do this we can construct the signal in the digital domain, do our processing on it and then later reconstruct it back to analog if needed. [2]

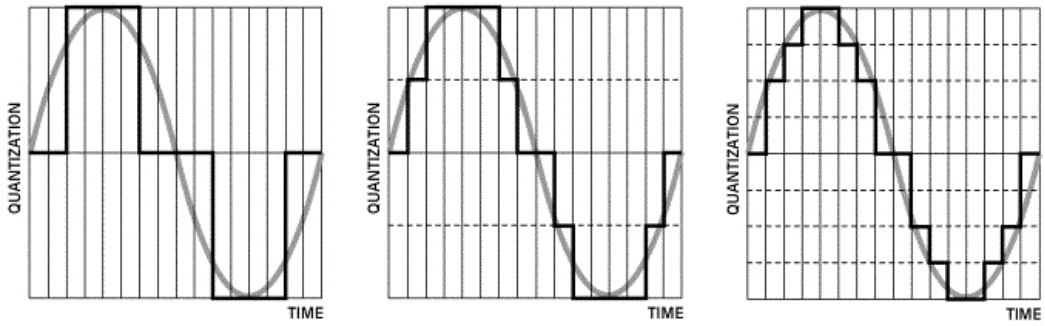

Figure 1.8: Quantization

1.4.2 Nyquist Theorem:

Any analog signal consists of components at various frequencies; the highest frequency component in an analog signal determines the bandwidth of that signal. Higher the frequency, greater the bandwidth. Suppose the highest frequency component or a given analog signal is F (MAX). According to the Nyquist Theorem, the sampling rate must be at least 2F (MAX), or twice the highest analog frequency component. If the sampling rate is less than 2F (MAX), some of the highest frequency components in the analog input signal will not be correctly represented in the digitized output. When such a digital signal is converted back to analog form by a digital-to-analog converter, false frequency components appear that were not in the original analog signal. This undesirable condition is called aliasing. [2]

1.4.3 Sampling Rate:

Sampling rate determines the sound frequency range; the range of frequencies represented in a waveform is called its bandwidth. Waveforms sampled at a high sampling rate can represent a broad range of frequencies and hence have broad bandwidth. In fact, the maximum bandwidth of a sampled waveform is determined exactly by its sampling rate; the maximum frequency represent able in a sampled waveform is called its Nyquist Frequency, and is equal to one half the sampling rates. [2]

1.4.4 Aliasing:

If a signal is sampled at sampling rate smaller than twice the Nyquist frequency false lower frequency components appear in the sampled data this is called aliasing. For example, the human ear can hear sounds from 20 - 20,000Hz. This means that if we want

to recreate an audio signal with the complete range of recorded frequencies we must sample the signal at a minimum of 40,000Hz. [2]

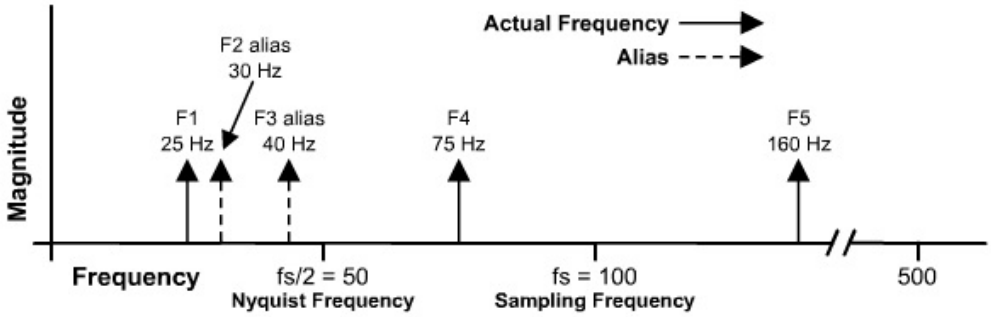

Figure 1.9: Aliasing

The alias frequency is the absolute value of the difference between the frequency of the input signal and the closest integer multiple of the sampling rate as an example, consider an input signal containing several frequencies. Assuming Fs the sampling frequency is 100 Hz therefore the Nyquist frequency = fs/2 = 50Hz. Using the Nyquist theorem frequencies below the Nyquist frequency are sampled correctly and the frequencies above the Nyquist frequency appear as aliases.

1.5 Time Domain:

Here is a signal in domain one way to represent this signal is the sum of numbers that occurred each point in time take a look at the figure the value of function f(t) at time t_1 here is the number $f(t_1)$ at this particular time $d(t - t_1)$. [3]

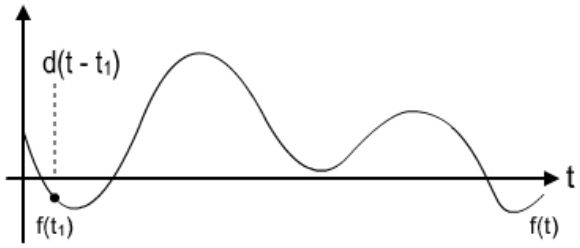

Figure 1.10: Time Domain

In order to create entire signal rather than just one point representation time is varied to have access to all point. So another way to represent this signal is by the sum of the numbers that occurs each point in time multiplied by the delta function of each point.

$$\sum f\ (t_n)\ \delta(t - t_n) \qquad (1.1)$$

As this equation show signal in time domain is sum of amplitudes at discrete points in time and we sum up the entire signal in time to get our function.

1.6 Frequency Domain:

It is a representation in which signal strength can be represented graphically as a function of frequency, instead of a function of time.

1.6.1 Fourier Transform:

Fourier transform convert signals from time domain to frequency domain, Voice is consist of various frequencies of low and high magnitude so in order to extract these frequencies Fourier Transform is used. Fourier transform convert signal from time domain into frequency domain.

Mathematical Representation:

$$\int_{-\infty}^{\infty} f(t)e^{j\omega t}\delta t \tag{1.2}$$

Where: $$e^{\pm jx} = cos(x) \pm jsin(x) \tag{1.3}$$

This equation tells us that we have to multiply our function f(t) with $e^{j\omega t}$, so to make it easier that function is represented in sine and cosine component. [3]

Discrete Representation:

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{-jk2\pi n/N} \qquad \text{for: n = 0 --- N-1} \tag{1.4}$$

N represents number of DFT points X(k) is our function and k varies from zero to N-1.

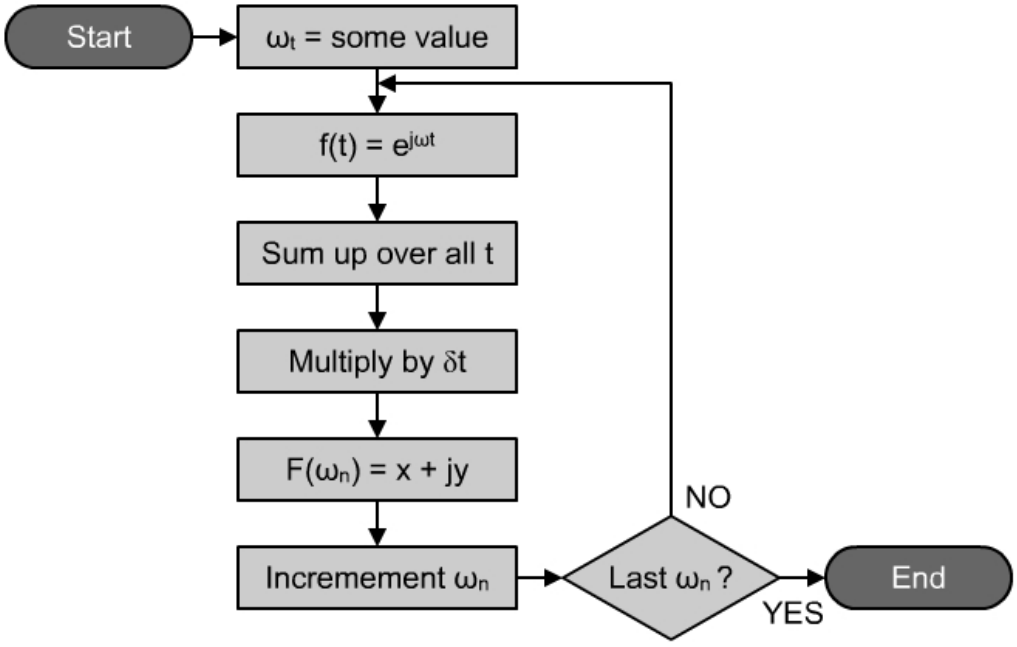

Figure 1.11: Fourier Transform Block Diagram

1.6.2 Fast Fourier Transform:

If we look at the equation of discrete Fourier transform you will see that it is complicated to work out with many addition and multiplications involving complex numbers as well even a single N=8 point DFT requires $(N-1)^2$ complex multiplications and N(N-1) complex addition while FFT approach is to break down one large problem into smaller chunks and thus it will reduce a lot of calculation resulting fast performance, we can say that FFT (Fast Fourier Transform) is the less computation approach for solving DFT (Discrete Fourier Transform). [3]

1.7 Linear Predictive Coding:

Linear Predictive Coding (or LPC) is a method of predicting a sample of a speech signal based on several previous samples. We can use the LPC coefficients to separate a speech signal into two parts: the transfer function (which contains the vocal quality) and the excitation (which contains the pitch). We can predict that the n^{th} sample in a sequence of speech samples is represented by the weighted sum of the "p" previous samples:

$$s = \sum_{k=1}^{P} a_k x[n-k] + e[n] \tag{1.5}$$

We take the Z-transform of above equation:

$$E(Z) = S(Z) - \sum_{k=1}^{P} a_k S[Z] Z^{-k} \tag{1.6}$$

$$E(Z) = S(Z)\,[1 - \sum_{k=1}^{P} a_k Z^{-k}] = S(Z) * A(Z) \tag{1.7}$$

In speech processing, computing the LPC coefficients of a signal gives us its a_k values. We can get the filter A(z). A(z) is the transfer function between the original signal s[n] and the excitation component e[n]. The transfer function of a speech signal is the part dealing with the voice quality: what distinguishes one person's voice from another. The excitation component of a speech signal is the part dealing with the particular sounds and words that are produced. In the time domain, the excitation and transfer function are convolved to create the output voice signal. As shown in the figure below, we can put the original signal through the filter to get the excitation component. Putting the excitation component through the inverse filter (1 / A(z)) gives us the original signal back.

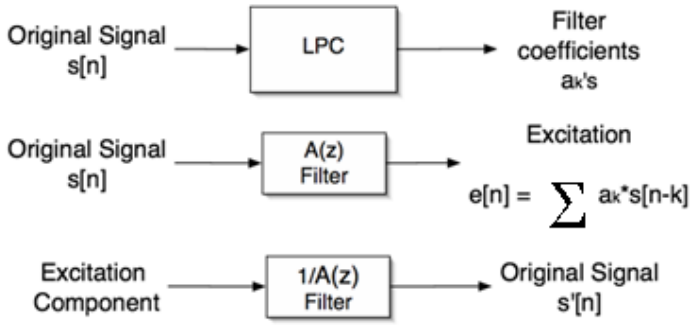

Figure 1.12: Using Linear Predictive Coding to separate the two parts of a speech signal

1.8 Mel Frequency Cepstral Coefficients:

Mel-frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of Cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCCs are commonly derived as follows:

1. Take the Fourier Transform of a signal.
2. Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the Mel frequencies.
4. Take the discrete cosine transform of the list of Mel log powers.
5. The MFCCs are the amplitudes of the resulting spectrum.

1.8.1 Mel scale:

This scale and normal frequency measurement is defined by assigning a perceptual pitch of 1000 Mels to a 1000 Hz tone, 40 dB above the listener's threshold. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments. As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the Mel scale. The name **Mel** comes from the word **melody** to indicate that the scale is based on pitch comparisons.

A popular formula to convert f hertz into m Mel is:

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right) = 1127 \log_{e}\left(\frac{f}{700} + 1\right) \tag{1.8}$$

And the inverse:

$$f = 700(10^{m/2595} - 1) = 700(e^{m/1127} - 1) \tag{1.9}$$

1.9 Hidden Markov Model:

It was developed by Andrei Andreyevich Markov. Markov is particularly remembered for his study of Markov chains, sequences of random variables in which the future variable is determined by the present variable but is independent of the way in which the present state arose from its predecessors. This work launched the theory of stochastic processes. A Markov model is a probabilistic model of symbol sequences in which the probability of the current event is conditioned only by the previous event. An HMM is a non-

deterministic Markov model that is, one where knowledge of the emitted symbol does not determine the state-transition. This means that in order to determine the probability of a given string, we must take more than one path through the states into account. [9]

1.9.1 Markov Random Processes:

A random sequence has the Markov property if its distribution is determined solely by its current state. Any random process having this property is called Markov Random Process. For observable state sequences (state is known from data), this leads to a Markov Chain Model. For non-observable states, this leads to a Hidden Markov Model (HMM).

1.9.2 Time-Based Models:

The models typically examined by statistics:

• Simple Parametric Distributions
• Discrete Distribution Estimates

These are typically based on what is called the "independence assumption". Each data point is independent of the others, and there is no time-sequencing or ordering.

In Hidden Markov Model:

• States are not observable.
• Discrete observations $\{v_1, v_2 ..., v_M\}$ are recorded; a probabilistic function of the state.
• Emission probabilities $B_j(m) \equiv P(O_t=V_m \mid Q_t=S_j)$
• Example: In each turn, there are balls of different colors, but with different probabilities.
• For each observation sequence, there are multiple state sequences [9]

1.9.3 From Markov To Hidden Markov:

The previous model assumes that each state can be uniquely associated with an observable event. Once an observation is made, the state of the system is then trivially retrieved. His model, however, is too restrictive to be of practical use for most realistic problems to make the model more flexible, we will assume that the outcomes or observations of the model are a probabilistic function of each state. Each state can produce a number of outputs according to a unique probability distribution, and each distinct output can potentially be generated at any state. These are known a Hidden Markov Models (HMM), because the state sequence is not directly observable, it can only be approximated from the sequence of observations produced by the system. [9]