**3. STRUCTURE OF SPEECH RECOGNITION SYSTEM**

**3.1. Overview**

Speech recognition system allows a computer to recognize the speech and the words that express through the microphone or speaks by phone. The core of these speech recognition systems consists of a set of statistical models representing the various sounds of the language to be recognised. Speech has temporal structure and can be encoded as a sequence of spectral vectors spanning the audio frequency range. In this chapter the structure of speech recognition system, its basic blocks are described. The basic techniques used for speech recognition have been described.

**3.2. The basic structure**

Speech recognition system includes set of blocks and processes. Figure 3.1 shows general structure of speech recognition system [21].



Figure 3.1: General structure of speech recognition system.

As shown above the system is consisted of five blocks, and the process of every block is explained below.

**3.2.1. Speech**

In this block speech raw are entering by a microphone as a continuous signal. Sampling is applied to convert the continuous speech signal to discrete form. This process is done by multiplying the continuous signal by train of impulses that have value of 1 volt, same time or space between the impulses is there. The number of samples that produced in one second called Sample Rate or Sampling Frequency and its unit is Hertz (Hz), commonly sampling rate used in speech signal are (4KHz, 8KHz, 10KHz, 16KHz, 22KHz), an example of sampling process is illustrated in figure 3.2, sampling can be done directly by the microphone. After that A/D converter is used to convert the signal from analog to digital, so the signal has limited values of amplitude that the system can deal with it, there are many types of A/D converters depending on their sizes, and usually 256 bit A/D converter is used. Now the signal is ready to be used in the next block.



Figure 3.2: Sampling process [41].

**3.2.2. Signal preprocessing**

Now the signal was converted to a sequence of samples as shown in figure 3.2 above. In this block preprocessing on the signal will be done. Noise and difference in amplitude of the signal can distort the integrity of a word while timing variations can cause a large spread amongst samples of the same word, so de-noising must be done to get the speech data without noise, band-pass filter is used for de-noising since the frequencies of the speech in range of 300Hz to 3750Hz. Then Endpoint detection algorithm is used to detect the starting and the ending of the word, in this thesis Zero Crossing based Endpoint Detection algorithm was used. It is important to detect the start and the end points of the word, in order to drop the silence regions of the speech signal, so good features can be extracted and also the memory that used to store the patterns can be reduced.

**3.2.3. Features extraction**

The process of feature extraction is done in many steps, the first step is preemphasis, in this stage the speech signal enters to a low order filter to spectrally flatten the signal and to make it less susceptible to finite precision effects, a high-pass FIR filter is used for this purpose and its function is illustrated in equation (3.1).



(3.1)

The second step is to split the signal to many same long blocks, every block has $N$samples of speech signal called Frame, and there is an overlapping of $M$samples between every two adjacent frames to sure that there is no losing samples, framing is done because the speech signal is slowly varying over time (quasi-stationary), the signal is fairly stationary, therefore speech signals are often analysed in short time segments, which are referred to as short-time spectral analysis, figure 3.3 shows framing process. After that windowing is applied on the frames and every frame is multiplied with a window function $w(n)$ with length $N$, where $N$is the number of samples in each frame, Windowing is used to avoid problems due to truncation of the signal, and usually Hamming window is used in speech recognition systems, the function of Hamming window is illustrated in equation (3.2) below.



(3.2)



Figure 3.3: Framing process [29].

The last step is the extraction of features from the windowed frames by one of the techniques that used for this purpose. Features extraction means extracting data from a signal as minimum as possible without losing any important data from the signal. To improve the recognition rate of the system, good features must be extracted from the speech signal that mean wide variation between the classes and a more flexibility for the class itself because the signal of the same word from the same person not exactly appear at every time. The quality of a good feature is that it gives maximum information about the class within a much smaller dimension. Further, these features are important in deciding the overall recognition system.

Several techniques are used for speech features extraction; some of these techniques are Spectrogram, Mel Frequency Cepstral Coefficients (MFCC), and Linear Predictive Coding (LPC). These techniques had been used in this thesis and they had been described in details later in this chapter.

**3.2.4. Speech classification**

In some systems called Pattern Matching. This block is responsible for recognizing the speech patterns and classifies these patterns each alone, many algorithms were developed by the researchers for this purpose, and some of these algorithms are vector quantization, Hidden Markov Model (HMM), Euclidean Squared Distance, Artificial Neural Network (ANN), Etc. In this thesis ANN algorithm was used for pattern matching. Researches had shown that ANN is gave the best recognition rate depending on many results that were gotten by many researchers. More details on types of ANN models that are used in speech recognition systems are described in the next chapter.

**3.2.5. Output**

The last block is output of the system, the output is the word that classified and the decision of matching between the words is made in this block. In some systems like “robotic” if there is matching between spoken word and stored word this mean an instruction is gave to the robot to do something like go left, or go right, or stop, but if there is no matching the system will do nothing. In some systems like “speech based password” the system still asking the user to repeat speaking of the word if there is no matching with the stored word.

**3.3. Features extraction techniques**

Several methods and techniques are used for speech feature extraction, some of these methods depend on spectral analysis and some of them depend on prediction. In this thesis, three methods had been used (LPC, MFCC, and Spectrogram) and the description of these methods was explained here.

**3.3.1. Linear Predictive Coding (LPC)**

Linear predictive coding is a technique used mostly in speech processing to estimate basic speech parameters like pitch formants and spectral envelope of the speech signal in compressed form.

LPC is one of the most useful methods for encoding good quality speech at a low bit rate. The coefficients are generated by the linear combination of the past speech samples using the autocorrelation or auto variance method. Equation (3.3) shows how to compute the coefficients of LPC.

$$\tilde{x}\_{(n)}=a\_{1}x\_{(n-1)}+a\_{2}x\_{(n-2)}+. . . . . .+a\_{p}x\_{(n-p)} (3.3)$$

Where $\tilde{x}\_{(n)}$ is the predicted of $x\_{(n)}$ based on the summation of past samples, $a\_{i}$ is the linear prediction coefficients, $p$ is the number of coefficients, and $n$ is the samples.

The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech and estimated speech signal over a finite duration. This could be used to give a unique set of predictor coefficients. The transfer function of the time varying digital filter is given in equation (3.4) below.



 (3.4)

LPC uses the Levinson-Durbin recursion to solve the numerical equations that arise from the least squares formulation. This computation of the linear prediction coefficients is often referred to as the autocorrelation method [22].

**3.3.1.1. LPC estimation**

LPC coefficients can be constructed by using LPC estimation processing. As mentioned previously, speech signal is a not stationary signal and the used methods for LPC estimation assume that the signal is stationary. So, framing the speech signal is used to make the signal quasi stationary. A more stationery signal result in a better LPC estimation because the signal is better described by the LPC coefficients and therefore minimize the residual signal. The residual signal also called the error signal [23]. Figure 3.4 shows a block diagram of LPC estimation.



Figure 3.4: Block diagram of LPC estimation [23].

Where $S$ is the input signal, $g$ is the gain of the residual signal (prediction error signal) and $a$ is a vector containing the LPC coefficients to a specific order. The order of LPC is an important factor for LPC estimation because of the bigger number of the LPC coefficients means better estimation of the vocal tract.

**3.3.1.2. LPC analysis**

The useful from LPC analysis is to calculate the residual signal or error signal. The inverse transfer function that was gotten from LPC estimation is filtered with the original signal to determine the residual signal. In some cases error signal is equal to the noise of speech production, these cases are appear when the inverse transfer function from LPC analysis is equal to the vocal tract transfer function [24].



Figure 3.5: Block diagram of LPC analysis [23].

Where *S* is the input signal, $g$ and $a$ is calculated from LPC estimation and $e$ is the residual signal for LPC analysis.

**3.3.1.3. LPC synthesis**

To construct the original signal, LPC synthesis is used. This process can be done by filtering the residual signal that was gotten from LPC analysis with the transfer function of the vocal tract. So, LPC synthesis can be done depending on LPC estimation and LPC analysis.



Figure 3.6: Block diagram of LPC synthesis [23].

Where *e* is the error signal found from LPC analysis, *g* and $a$ are found from LPC estimation.

**3.3.1.4. LPC applications**

The main use of LPC is to analyse and resynthesize the speech. It is used as a method for features extraction in some systems that use speech recognition. LPC is used also in some phone companies as a compression method. US government used LPC in wireless communication, where LPC is a secure technique. Paul Lansky made the well-known computer music piece as an idle chatter using LPC. In some educational toys, LPC was used as Speak and Spell. LPC is one of the many methods that were used in lossless audio codec, so, it is used in MPEG-4 ALS (an extension to the MPEG-4 part3 audio standard to allow lossless audio compression) and FLAC (Free Lossless Audio Codec) [25].

LPC can be applied in many areas of speech processing (Makhoul, 1975). For example, linear prediction can be used in clinical practise for analysing voice disorders (Baken and Orlikoff, 1999). In the area of engineering LP is a basic method for speech analysis, for compression and coding (Atal and Hanauer, 1971) or for speech and speaker recognition (Choi et al., 2000; Bimbot et al., 2000). In speech analysis LP has been widely used in formant extraction in voice-source analysis (Alku, 1992) [26].

**3.3.2. Mel Frequency Cepstral Coefficients (MFCC)**

It is a method used to extract the features from the speech signal. MFCC is based on human hearing perceptions because it is observed that human ear acts as a filter, these filters are non-uniformly spaced on the frequency axis (More filters in the low frequency regions and Less no. of filters in high frequency regions) as shown in figure 3.7. So MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz [27].



Figure 3.7: Mel scale filter bank [28].

The formula for the Mel scale is illustrated in equation (3.5), and the plot of Mel scale is shown in figure 3.8:

𝑀 = 2595\* log10 (𝑓/700 + 1) (3.5)



Figure 3.8: Mel scale plot [22].

The Mel scale, named by Stevens, Volkmann and Newman in 1937, and the name Mel comes from the word “Melody” to indicate that the scale is based on pitch comparisons. There are several methods that compute MFCCs, The steps that are applied in the traditional method to get the coefficients are started with reemphasizing the sampled signal and then applying the framing and windowing on it, then taking the Fast Fourier Transform (FFT) for each windowed frame, the signal now is a power spectrum, this signal enters to a Mel filter bank and the length of the output is equal to the number of filters created, after taking a discrete cosine transform to the log of the filter bank`s output, an array of features that describe the spectral shape of the signal [28].

Figure 3.9 below shows the steps of MFCC:



Figure 3.9: Block diagram of MFCC [22].

As seen above the processing is started from converting the speech signal from analog to digital, then making a preemphasis, framing, and windowing on the digital signal, these all process are explained in the beginning of this chapter. After that Fast Fourier Transform (FFT) is applied on each windowed frame, the equation (3.6) below shows FFT function:



(3.6)

After that overlapped triangular windows are used as a filters for the power spectrum signal that is got from Fast Fourier Transform above, the filters must be as a Mel frequency to simulate the auditory process of Human, so a logarithm function is applied on the filters (see equation 3.5) as explained previously. And the last step is to convert the signal to a time domain again by applying Discrete Cosine Transform (DCT) and the coefficients are ready to be used in the pattern matching stage. The MFCC`s can be calculated by this equation (3.7) [29]:



(3.7)

Where, *P* is the order, *k* is the number of DFT Magnitude coefficients, $X\_{k}$ is the *k`*th order log-energy output from the Mel filter bank.

**3.3.3. Spectrogram**

A Spectrogram, or Sonogram, is a visual representation of the spectrum of frequencies in a sound [30].Spectrogram is used in many fields like speech processing, sonar and music. It is also used to show the difference between spoken words phonetically. Three dimensions are used to represent the spectrogram of any signal. Time is represented in the horizontal dimension and is read from left to right, frequency is represented in the vertical dimension and amplitude is represented as an intensity of the colour. Higher amplitude is darker colour. Voice analysis is performed by spectrogram in two main kinds of analysis; these kinds are wideband and narrowband [31].

**3.3.3.1. Wideband spectrogram**

It can be said that the spectrogram is a wideband when the spectral snapshots are between 100-200 Hz. By counting the number of vertical lines per unit in the graphic of the spectrogram the fundamental frequency can be found. Also, the frequencies and relative strengths of the first two formants (F1 and F2) are visible as dark, rather blurry concentrations of energy [32].Figure 3.10 shows the wideband spectrogram of signal for vowels utterances i, e, a, o and u.



Figure 3.10: Spectrogram of wideband speech signal [32].

A good resolution is allowed in the wideband spectrogram analysis, and as seen above the energy peaks from each individual vibration of the vocal folds in the graph. Individual harmonics cannot be singled out by this type of spectrogram, this means frequency components within 300-500 Hz bandwidth are not easily distinguished.

**3.3.3.2. Narrowband spectrogram**

Unlike wideband spectrogram, narrowband spectrogram has a not good enough resolution to isolate each individual cycle of vibration, but each individual harmonic in it can be singled out. Also, the formant structure of the sound is not rendered as clearly as with a wideband spectrogram. See the dark horizontal stripes; they represent each harmonic, in the graphic below. Also see that the large clusters of formant energy which were seen in the wideband spectrogram are not present in narrowband spectrogram [32].Figure 3.11 shows the narrowband spectrogram of speech signal for vowels utterances i, e, a, o and u.

.



Figure 3.11: Spectrogram of narrowband speech signal [32].

Spectrogram is one of the many tools that are used as a helpful tool in the voice clinic to provide feedback as part of a voice therapy or training program.

**3.3.3.3. How to get spectrogram of a speech signal**

Spectrogram of a speech signal can be gotten by take a Fast Fourier Transform (FFT) for each frame of the speech signal to convert from time domain to frequency domain for the frames and this mean the spectrum of the frames, and then make the horizontal axis for frequency and the vertical axis for amplitude as in figure 3.12.



Figure 3.12: Exchange the axis of the frame.

And then represent the amplitude as a grey colour from 0 to 255, 0 represents black colour and 255 represents white colour, Higher the amplitude, darker the corresponding region. Now representing the vertical axis for the frequency, the horizontal axis for time (number of frames for the speech signal), and the amplitude represent the intensity of the colour, figure 3.13 illustrate the getting of the Spectrogram from the speech signal.



Figure 3.13: Steps to get spectrogram [42].