**2. SPEECH RECOGNITION**

**2.1. Overview**

Speech recognition is a technology that allows converting spoken words to a text or to instructions for security or controlling purposes, by digitizing the sound and matching its pattern against the stored patterns [2].Speech recognition systems used to help people with disabilities to meet their needs; also it is used in smart phone applications. This technology evolved rapidly, where the researchers focus on this technology being very useful and it can be used in many areas and future applications.

The systems that based on speech recognition technology can work instead of human to perform the same purpose that human can do it. Speech recognition system that used by some companies to answer customers questions by phone and give them an answer to solve some of the problems experienced by customers is an example.

**2.2. Applications of speech recognition system**

Speech recognition applications are entered in many areas and fields. Some of these applications are:

* **Modern cars:**

Nowadays some cars companies (like TOYOTA) are developing some features that depend on speech recognition technology; these features (are used by the driver of the car) allow the driver to focus on driving while he can use these features by specking the thing that he want to do instead of using his hand, and these features reduce the probability of accidents [3]. Figure 2.1 shows how the driver can use speech recognition feature in the modern car.



Figure 2.1: Speech recognition technology in modern cars [4].

* **Robotics:**

There is some robotics that does something depending on speech recognition technology (like children toy cars), this robotics moves depending on the words that the user speak them, for example go, back, stop, right, left, up, down, and so on [5]. Figure 2.2 shows two robots controlled by speech.



Figure 2.2: Two robots controlled by speech [6].

* **Speech instead of human hand to entire a security code:**

In some security systems the using of speech is better than the using of hand to entire the code, because there is a probability for the user to do some mistakes like writing a number or letter instead another one or forget one of them to write it, but in systems that use speech, the user speak the code and the system will work. Some of security systems depend on the envelope of the voice signal that mean just special peoples can use this system, that known as Speaker Recognition system. Figure 2.3 shows an example of using speech recognition as a keyboard for entering a password instead of traditional keyboard.



Figure 2.3: Speech password device [7].

* **Input device:**

The latest versions of Windows like Windows7 and Windows8 used speech recognition technology as an input device instead of mouse. Some of the combined features on the latest versions of Windows are go to next page or return to previous one, play a movie or stop it, open an application or close it, and so on. Figure 2.4 shows the feature of speech recognition that used in Windows 8, the user just say ‘next’ and the next page will appear.



Figure 2.4: Speech recognition feature in Windows 8.

As seen above the areas and fields that speech recognition is used in are wide and variety. Now to know how speech recognition system works, firstly we must know types of human speech and some properties of human speech, and how the speech is produced.

**2.3. Types of speech**

There are different types of speech that the speech recognition system applied on them to recognize the word spoken, these types are [8]:

* **Isolated Words:**

One utterance at each processing time, that mean the user speak one word or utterance and must wait the system to recognize it, then he speak another word or utterance and so on, this type of speech is used in some robotics to control them to do something like go, stop, left, right, and back.

* **Connected Words:**

The user in this type of speech, speak more than one word (phrase) but he must pauses after each word or utterance for limited time to give the system a time to recognize the spoken words. This type of speech is like the previous one, but allows separate utterances to be 'run-together' with a minimal pause between them.

* **Continuous Speech:**

This is the best type compared with others, as it gives the user freedom to speak in their own way, but it is difficult to create such these recognizers because they require special techniques or methods to determine utterance boundaries.

**2.4. Speech signal and its basic properties**

Speech is an important communication tool between peoples. The researchers started to develop some systems that depend on speech to be used instead of human like some robotics, or instead of some input devices like mouse and keyboard. The development of speech recognition systems requires a good knowledge of the basic properties of the speech signal, so a brief description on speech signal properties had been explained below.

One of the speech signal properties is slowly varying on time over a short period of time (less than 100 msec), and fairly stationary over a long period of time. The signal characteristics change to reflect the different speech sounds being spoken. Usually the beginning of the speech signal (nearly 1600 samples if the sampling rate is 8000 samples/sec) recording corresponds to silence (or background noise) because the speaker takes some time to speak the word when recording starts. These properties of the speech signal are illustrated in figure 2.5.

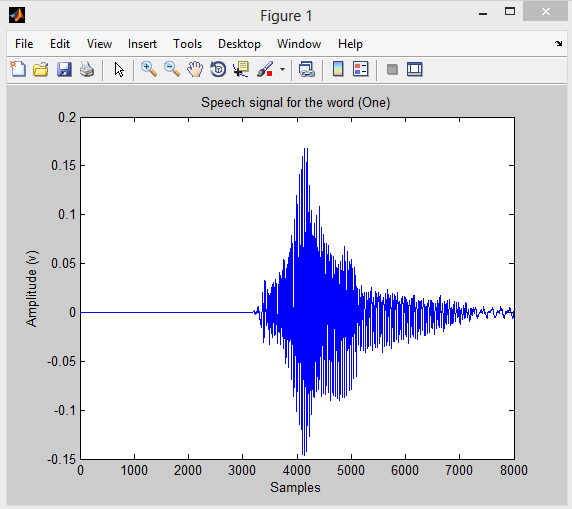


Figure 2.5: Speech signal for the word “one”.

Speech can be represented phonetically by a finite set of symbols called the phonemes of the language, the number of which depends upon the language and the refinement of the analysis. For most languages the number of phonemes is between 32 and 64. Sounds like /SH/ and /S/ look like (spectrally shaped) random noise, while the vowel sounds /UH/, /IY/, and /EY/ are highly structured and quasi-periodic. These differences result from the distinctively different ways that these sounds are produced [9].

**2.5. Speech production**

The knowledge of how Humans generate the speech and how the speech signal is perception by them is an important step to build a good speech recognition system. In this section a brief explaining of speech production was described.

Huang mentioned that speech is produced by air-pressure waves emanating from the mouth and the nostrils of a speaker (Huang et al, 2001). Figure 2.6 shows Human speech production apparatus.

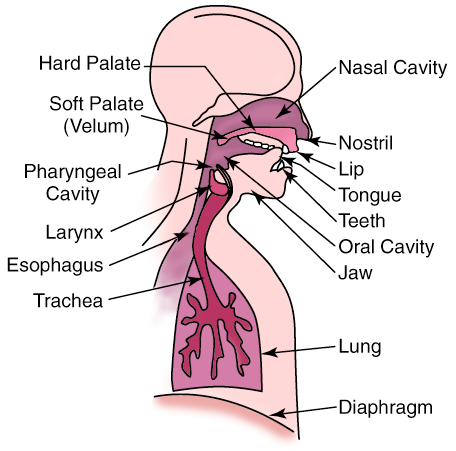
**

Figure 2.6: Human speech production apparatus [40].

The process of speech production is done after using the organs that involved producing the speech. These organs are lungs, larynx, nose and mouth. The air expelled from the lungs is modulated in different ways to produce the acoustic power in the audio frequency range. After, the rest of the vocal organs, such as vocal cords, vocal tract, nasal cavity, tongue, and lips, modify the properties of the resulting sound to produce the speech waveform signal. These properties can be principally determinate thanks to the acoustical resonance process performed into the vocal tract. The main resonant modes are known as formants, being the two lowest frequency formants the most important ones in determining the phonetics properties of speech sounds. This resonant system can be viewed as a filter that shapes the spectrum of the source sound to produce speech (Holmes & Holmes, 2001) [10].

**2.6. Review of speech recognition**

Many papers and thesis were made on speech recognition, and different techniques and algorithms were used for this purpose.

In reference [11], Voice Recognition system for controlling robotic applications was developed. LPC & HMM were used, where LPC is a method to analyse voice signals by giving characteristics into LPC coefficients, and HMM is a form of signal modelling where voice signals are analysed to fined maximum probability and recognize words given by a new input based from the defined codebook.

Five basic movement of the robot “forward”, “left”, “right” and “stop” were tested, and the system could to recognize them. 5 male and 5 female samples were chosen as input with the same keyword given. The performance test involving 5 samples with 25 voice data files given 25 files correctly recognized or resulting the 100% accuracy, where the same test involving 5 samples with 25 voice data files given 17 files correctly recognized or resulting the 68% accuracy.

In reference [12], isolated words automatic Arabic speech recognition system was developed in this reference. LPC and HMM methods were used for feature extraction, and pattern classification respectively. A new pattern classification algorithm was developed; where neural network trained using Al-Alaui Algorithm was used. This new method gave comparable results to the already implemented HMM method for the recognition of words, and it has overcome HMM in the recognition of sentences.

Reference [13] is talking about animal identification system based on animal voice pattern recognition; many algorithms for different purposes were used to develop the system, Zero-Cross-Rate (ZCR) algorithm for endpoints detection of the voice signal, Mel Frequency Cepstral Coefficients (MFCC) algorithm for feature extraction, and Dynamic Time Warping (DTW) algorithm for pattern matching.

Security system based on voice identification as the access control key was developed in reference [14], the system was developed by building a microcontroller circuit to controlling access to a door by using MATLAB (Simulink) function blocks to test the reliability of the system. When there is a matching the system produces “1” logic, otherwise the system produces “0” logic. The system succeeded to activate the door opening mechanism using a voice command that only works for the authenticated individual. The system is proven to be able to provide medium-security access control and also has an adjustable security level setting to account for the variations in one's voice each time a voice identification occurs.

In reference [15], emotion recognition results from speech signal system was developed. The system particularly focuses on extracting emotion features from the short utterances typical of Interactive Voice Response (IVR) applications. Sadness, boredom, happy, and cold anger emotions were chosen for classification using neural networks, Support Vector Machines (SVM), K-Nearest Neighbours, and decision trees. The database that used in this reference was taken from the Linguistic Data Consortium at University of Pennsylvania, which is recorded by 8 actors expressing 15 emotions. Recognition Rate obtained in this reference is over 90% accuracy.

In reference [16], Speech Recognition system was developed. Two methods LPC and DWT were used for feature extraction, and ANN was used for pattern classification to recognize speaker independent spoken isolated words. Words from Malayam, one of the four major Dravidian languages of southern India were used for recognition. 50 speakers uttered 20 isolated words for each of them. Both the methods produce good recognition accuracy, but discrete wavelet transforms are found to be more suitable for recognizing speech because of their multi-resolution characteristics and efficient time frequency localizations.

In reference [17], Neural Network Hidden Markov Model (N.N.H.M.M) hybrids was used as a technique for speaker independent continuous speech recognition system, N.N was used to perform acoustic modelling, and H.M.M was used to perform temporal modelling, the best word accuracy achieved by this thesis was 90.5% when N.N.H.M.M hybrids was used and 86.0% when pure H.M.M was used under similar conditions.

Reference [5] is talking about implementation of speech recognition system on a mobile robot for controlling movement of the robot. LPC algorithm was used to extract features from the speech signal, and backpropagation neural network was used to make the matching between the patterns that produced by LPC processing. Data that obtained from LPC process become the input to the neural network, 576 data were the output of LPC for each utterance. 30 persons were chosen to speak 7 utterances for each one of them. Indonesian words were used as a language for controlling, and the highest recognition rate that obtained in this system was 91.4%.

In reference [18], speech recognition system was developed to control a mobile robot, two algorithms were used one for extract the features from the speech signal, it is LPC, and the other algorithm used for pattern matching this algorithm is H.M.M. This system was implemented on personal computer, and the system take the speech signal by a microphone in 1 second and the signal processed with LPC. The highest recognition rate that achieved was 100 %.

In reference [19], the author was presented a security system based on speaker identification. Mel Frequency Cepstral Coefficients (MFCC) method was used for feature extraction, and vector quantization technique is used to minimize the amount of data to be handled.21 speakers, which include 13male and 8 female, were chosen to make the database of the system. The system succeeded to identify the speakers and the highest rate of recognition that the system achieved was 100%.

The isolated word speech recognition system was developed in reference [20], the algorithms that were used in this reference are LPC for feature extraction, Vector Quantization (VQ) is used to create reference templates for speaker recognition, and Dynamic Time Warping (DTW) was used for feature matching, 12 words of Lithuanian language pronounced ten times by ten speakers were the database that the system was worked with, and the recognition error rate obtained in this reference was 2.5% when VQ was used and 1.94% without using VQ.