Faculty of Engineering

Systems & Biomedical engineering Department

“Computer Aided Recognition of Voice Disorders”

Al Hassan Yosri Ibrahim
Joseph Girgis Ghebryal
Mazen Atef Abdel Fattah

Supervised by

Dr. Manal Abdel Wahed
First of all, we acknowledge God (Allah) the Almighty for the great help during our whole life.

We would like to express our thanks and appreciation to our supervisor Dr. Manal Abdel Wahed as she encouraged us to implement our idea and believed in our abilities. She supported us in all the work stages with her knowledge and guidance in a friendly manner.

We would like to express our appreciation to Dr. Yasser Tolba, Phonetics Physician in the phonetics department in Kobry El Kobba military hospital for providing us with the datasets used in this project and providing us with the necessary medical background.
<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>2</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>3</td>
</tr>
<tr>
<td>List of Tables</td>
<td>6</td>
</tr>
<tr>
<td>List of Figures</td>
<td>7</td>
</tr>
<tr>
<td>Abstract</td>
<td>10</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
<td>12</td>
</tr>
<tr>
<td>1.1 Motivation for the Study</td>
<td>12</td>
</tr>
<tr>
<td>1.2 Objective of the Study</td>
<td>16</td>
</tr>
<tr>
<td>1.3 Research Outline</td>
<td>17</td>
</tr>
<tr>
<td>Chapter 2: Medical Background</td>
<td>18</td>
</tr>
<tr>
<td>2.1 Anatomy of the Speech Production System</td>
<td>18</td>
</tr>
<tr>
<td>2.2 The Sub-glottal Part</td>
<td>18</td>
</tr>
<tr>
<td>2.2.1 The Vibratory Part</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2 The Cartilages of the Larynx Skeleton</td>
<td>22</td>
</tr>
<tr>
<td>2.2.3 The Laryngeal Cavity</td>
<td>23</td>
</tr>
<tr>
<td>2.2.4 The Supra-glottal Part</td>
<td>24</td>
</tr>
<tr>
<td>2.3 The Process of Speech Production</td>
<td>24</td>
</tr>
<tr>
<td>2.3.1 The Vibration of the Vocal Folds</td>
<td>25</td>
</tr>
<tr>
<td>2.3.2 Articulation</td>
<td>26</td>
</tr>
<tr>
<td>2.4 Pathologies of the Larynx</td>
<td>26</td>
</tr>
<tr>
<td>2.4.1 Organic Dysphonia</td>
<td>27</td>
</tr>
<tr>
<td>2.4.2 Neurological Dysphonia</td>
<td>28</td>
</tr>
<tr>
<td>2.4.3 Dysfunctional Dysphonia</td>
<td>29</td>
</tr>
<tr>
<td>2.5 Clinical Assessment of Voice</td>
<td>29</td>
</tr>
<tr>
<td>2.5.1 Patient Self-Assessment</td>
<td>30</td>
</tr>
<tr>
<td>2.5.2 Perceptual Analysis</td>
<td>30</td>
</tr>
<tr>
<td>2.5.3 Aerodynamic Measurement</td>
<td>31</td>
</tr>
<tr>
<td>2.5.4 Acoustic Analysis</td>
<td>31</td>
</tr>
<tr>
<td>2.5.5 Vocal Fold Motion Measurement</td>
<td>32</td>
</tr>
<tr>
<td>Chapter 3: Literature Survey</td>
<td>35</td>
</tr>
<tr>
<td>3.1 Common Features in Voice Pathology Assessment</td>
<td>36</td>
</tr>
<tr>
<td>3.1.1  Fundamental Frequency</td>
<td>36</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----</td>
</tr>
<tr>
<td>3.1.2  Cepstral Features</td>
<td>36</td>
</tr>
<tr>
<td>3.1.3  Acoustic Features Computed in MDVP Software</td>
<td>39</td>
</tr>
</tbody>
</table>

Chapter 4: The Proposed System

| 4.1  Developed Software       | 46 |
| 4.2  Speech Signal Acquisition | 47 |
| 4.3  Signal Pre-processing    | 48 |
| 4.3.1  Pre-emphasis           | 48 |
| 4.3.2  Frame Blocking         | 49 |
| 4.3.3  Windowing              | 50 |
| 4.3.4  Segmentation           | 50 |
| 4.4  Feature Extraction       | 51 |
| 4.4.1  Zero Crossing Rate (ZCR) | 51 |
| 4.4.2  Short Time Energy      | 55 |
| 4.4.3  Mel Frequency Cepstral Coefficients (MFCC) | 57 |
| 4.4.4  Linear Predictive Coding (LPC) | 61 |

Chapter 5: Experimental work and Results

| 5.1  Datasets                 | 65 |
| 5.2  Experimental work        | 65 |
| 5.2.1  Segmentation process   | 65 |
| 5.2.2  Feature Extraction     | 72 |
| 5.2.2.1  Short Time Energy    | 72 |
| 5.2.2.2  Zero Crossing Rate (ZCR) | 76 |
| 5.2.2.3  Linear Predictive Coding (LPC) | 83 |
| 5.2.2.4  Mel Frequency Cepstral Coefficients (MFCC) | 87 |

| 5.3  Classification          | 91 |

Chapter 6: Conclusions and Future Work

| 6.1  Conclusions             | 97 |

**Table of Contents (Cont’d)**
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2 Challenges and Drawbacks</td>
<td>99</td>
</tr>
<tr>
<td>6.3 Future Work</td>
<td>100</td>
</tr>
<tr>
<td>References</td>
<td>102</td>
</tr>
<tr>
<td>Table No.</td>
<td>List of Tables</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>5.1</td>
<td>Accuracy, sensitivity and specificity for 200 ms segment length</td>
</tr>
<tr>
<td>5.2</td>
<td>Accuracy, sensitivity and specificity for 300 ms segment length</td>
</tr>
<tr>
<td>5.3</td>
<td>Accuracy, sensitivity and specificity for 400 ms segment length</td>
</tr>
<tr>
<td>5.4</td>
<td>Accuracy, sensitivity and specificity for 500 ms segment length</td>
</tr>
<tr>
<td>5.5</td>
<td>Accuracy, Sensitivity and Specificity for different thresholds in the short time energy mean feature</td>
</tr>
<tr>
<td>5.6</td>
<td>Accuracy, Sensitivity and Specificity for different thresholds in the ZCR Max</td>
</tr>
<tr>
<td>5.7</td>
<td>Accuracy, Sensitivity and Specificity for different thresholds in the ZCR Mean</td>
</tr>
<tr>
<td>5.8</td>
<td>Accuracy, Sensitivity and Specificity for different thresholds in the LPC</td>
</tr>
<tr>
<td>5.9</td>
<td>Accuracy, Sensitivity and Specificity for different thresholds in the MFCC</td>
</tr>
<tr>
<td>5.10</td>
<td>Accuracy, Sensitivity and Specificity for the single features and the combined ones</td>
</tr>
<tr>
<td>Figure No.</td>
<td>List of Figures</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1.1</td>
<td>Speech waveform and spectrogram of the word &quot;please&quot;</td>
</tr>
<tr>
<td>2.1</td>
<td>The vocal apparatus (sagittal view)</td>
</tr>
<tr>
<td>2.2</td>
<td>The lungs and the bronchi</td>
</tr>
<tr>
<td>2.3</td>
<td>Different views of the larynx cartilages</td>
</tr>
<tr>
<td>2.4</td>
<td>Laryngoscopic view of the interior of the larynx</td>
</tr>
<tr>
<td>2.5</td>
<td>Movements of the vocal folds during one cycle</td>
</tr>
<tr>
<td>2.6</td>
<td>Imaging of vocal fold motion by digital stroboscopy</td>
</tr>
<tr>
<td>2.7</td>
<td>Principles of electroglottography</td>
</tr>
<tr>
<td>3.1</td>
<td>Common structure of voice pathology recognition systems</td>
</tr>
<tr>
<td>3.2</td>
<td>Weighting functions for Mel-scale filtering</td>
</tr>
<tr>
<td>3.3</td>
<td>Relation between Hz and Mel scales</td>
</tr>
<tr>
<td>4.1</td>
<td>Block diagram of the Proposed System</td>
</tr>
<tr>
<td></td>
<td>List of Figures (Cont’d)</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>4.2</td>
<td>Snapshot from GUI</td>
</tr>
<tr>
<td>4.3</td>
<td>Signal Acquisition in GUI</td>
</tr>
<tr>
<td>4.4</td>
<td>Examples of time representation of sustained vowel /ah/ said by a normal speaker and a Pathologic speaker before and after pre-emphasis</td>
</tr>
<tr>
<td>4.5</td>
<td>framing process on a continuous signal.</td>
</tr>
<tr>
<td>4.6</td>
<td>Segmentation panel</td>
</tr>
<tr>
<td>4.7</td>
<td>Illustration of the computation of the zero crossing rate</td>
</tr>
<tr>
<td>4.8</td>
<td>Feature panel</td>
</tr>
<tr>
<td>4.9</td>
<td>ZCR mean for a 400 ms segment of normal audio signal</td>
</tr>
<tr>
<td>4.10</td>
<td>ZCR Max for a 400 ms segment of normal audio</td>
</tr>
<tr>
<td>4.11</td>
<td>Feature panel</td>
</tr>
<tr>
<td>4.12</td>
<td>Energy mean for a 400 ms segment of normal audio</td>
</tr>
<tr>
<td>4.13</td>
<td>MFCC Calculation</td>
</tr>
<tr>
<td>4.14</td>
<td>Frequency domain features panel</td>
</tr>
<tr>
<td>5.1</td>
<td>Accuracy VS Segment length</td>
</tr>
<tr>
<td>5.2</td>
<td>Sensitivity VS Segment length</td>
</tr>
<tr>
<td>5.3</td>
<td>Specificity VS Segment length</td>
</tr>
<tr>
<td>5.4</td>
<td>Short time energy mean accuracy, sensitivity and specificity curve</td>
</tr>
<tr>
<td>5.5</td>
<td>Short time energy mean values</td>
</tr>
<tr>
<td>5.6</td>
<td>ZCR Max accuracy, sensitivity and specificity curve</td>
</tr>
<tr>
<td>5.7</td>
<td>ZCR Max values</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>5.8</td>
<td>ZCR Mean accuracy, sensitivity and specificity curve</td>
</tr>
<tr>
<td>5.9</td>
<td>ZCR Mean values</td>
</tr>
<tr>
<td>5.10</td>
<td>LPC accuracy, sensitivity and specificity curve</td>
</tr>
<tr>
<td>5.11</td>
<td>LPC values</td>
</tr>
<tr>
<td>5.12</td>
<td>MFCC accuracy, sensitivity and specificity curve</td>
</tr>
<tr>
<td>5.13</td>
<td>MFCC values</td>
</tr>
</tbody>
</table>
Abstract

Laryngeal diseases and vocal fold pathologies have strong impacts in the resulting quality of the voice production. Many approaches have been developed to analyze the acoustic parameters for the objective judgment of the pathological voice.

The aim of this research is to propose a user friendly system for the discrimination between normal and diseased voice. It is to help physicians in this field as an automatic tool to support their diagnosis and decision for further investigation. The feature extraction technique has been applied on the voice signal in the time domain and in the frequency domain. Time domain features used in this research are: Zero Crossing Rate (ZCR) and Short time Energy. Frequency domain features used are: Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). Classification was based on threshold detection of each feature or group of features.

The database used in this work was developed and collected in a sound proof room of the Phonetics department of Kobri Elkobba military hospital. The acoustic samples correspond to sustained phonations (1-3 sec long) of vowel /ah/ from patients (males and females) with wide variety of vocal folds disorders. After analyzing 50 different voice signals, 30 diseased signals and 20 normal, our decision was to choose 400 ms segment in the middle of vowel “ah” as this segment gives the best accuracy and it gives relatively high sensitivity and specificity. The analysis resulted in the following conditions for normal voice signal: Energy Mean > 0.07, ZCR Max < 0.23, ZCR Mean [0.09:0.13], LPC [110:130] or [167:220], and finally MFCC [130:150].
The proposed system yielded the highest Accuracy of 90% with combining both ZCR Mean AND ZCR Max, highest sensitivity of 100% with ZCR Mean, and highest specificity of 97% with combining both ZCR Max AND MFCC, also with combining both ZCR Mean AND ZCR Max.

This method is quantitative and non-invasive nature, allowing the identification and monitoring of vocal system diseases, achieving early detection of laryngeal pathologies, and reducing the cost and time required for basic analysis.
CHAPTER 1

INTRODUCTION

1.1 Motivation for the Study

The laryngeal pathology has received much attention nowadays due to the modern way of life which led to an increased number of professionals whose working activity greatly depends on the use of their voice such as teachers, TV presenters, and singers; also unhealthy social habits such as smoking and too much alcohol drink may cause voice disorder [1]. People are subjected to the risk of voice problems due to errors after surgical operations such as laser cordectomy, or Para thyroidectomy, etc.

Acoustic analysis has proved to be an excellent tool for voice disorder detection and assessment [2]. Voice assessment techniques may be categorized into two categories: subjective and objective techniques. Ear, Nose and Throat doctors use a subjective technique, which relies on the doctor's hearing to the patient's voice which may cause errors. The objective technique is based on physical measurements obtained during phonation. It includes measures of vocal fold vibratory movement, such as laryngoscopy, glottography, and digital stroboscopy.

These techniques are more accurate in diagnosing various laryngeal diseases due to their ability to capture the vocal folds movements. However, they are invasive, require costly resources and require experienced professionals. Also, it may cause much discomfort and sometimes generating resistance by the patients during examination, which may cause distortions in the data and thus produce false diagnoses [3].
The vocal and voice diseases have to be diagnosed in the early stage. It is well known that most of these diseases cause changes in the acoustic voice signal. Therefore, the voice signal can be a useful tool to diagnose them. Researchers have been interested for many years to find reliable ways in analyzing the content of the speech signal.

The analysis of voice signal is usually performed by the extraction of acoustic parameters using digital signal processing techniques. After that, these parameters are analyzed to determine the characteristic of the voice: non-pathologic, pathologic and the type of pathology.

The diagnosis of pathologies from the voice signal presents several advantages with regard to other methods:

- It is a non-invasive tool and easy to use, the patient only has to speak.
- It makes it possible to build an automatic computer based diagnosis system.
- The diagnostic is objective because it is based on the value of acoustic parameters.
- It can be used as a complementary tool to those methods based on the direct observation of the vocal folds using laryngoscopy.
- It can be used for the evaluation of surgical and pharmacological treatments and rehabilitation processes, by measuring the deviation degree between the pathological and the normal voice patterns.
- It can also be used for the early detection of vocal fold pathology.
There are many algorithms to calculate acoustic parameters and it was shown that there is a great correlation between these parameters and the pathologies. Furthermore, there are a great number of acoustic parameters that can be extracted and analyzed, however it is not absolutely clear their usefulness for solving the problem. The selection of the most appropriate parameters is still an open problem.

The speech signal is a slowly time varying (quasi-periodic) signal which depends on known physical movements. When examined over a short period of time, its characteristics are fairly stationary; however over long periods of time the signal characteristics change to reflect the different speech sounds being spoken.

Speech signals are represented in two domains i.e. time and frequency domain. Figure 1.1 shows both the time domain (waveform) and frequency domain (spectrogram) analysis of speech signal "please" uttered by both male and female speakers.
Human speech is a product of several physiological parts working jointly to convey an acoustic message from the speaker to the listener. The main components of the speech production system are the lungs and glottis comprising the source function and the vocal tract comprising the filter function.

The voice utterance results from three components of voice production: voiced sound, resonance, and articulation. Voiced sound produced by vocal fold vibration is amplified and modified by the vocal tract resonators (the throat, mouth cavity, and nasal passages), and finally vocal tract articulators (the tongue, soft palate, and lips) modify the voiced sound; therefore, recognizable words are produced.
Vocal folds are in the form of two elastic hands of muscle tissue located in the larynx (voice box) directly above the trachea, and most of the disorders result from their malformation. The health condition and functionality of vocal folds have some effects on the quality of voice. If the vocal folds become inflamed, some growths may develop on them or they become paralyzed and, as a result, the speech production process may fail. In cases with disordered voice, speech samples carry symptoms of disorder from their origin. Therefore, any abnormality in larynx often affects the quality of voice signal. The common disorders are vocal fold paralysis, vocal fold edema, vocal fold nodules, polyps, carcinoma, adductor spasmodic dysphonia, anterior and posterior (A-P) squeezing, and others [2]. Disorders usually show up in speech signal in the form of acoustic perceptual measures, such as hoarseness, breathiness, harshness, and inability to project the voice loudly.

1.2 Objectives of the Study

The aim of this research is to propose a user friendly system for the discrimination between normal and diseased voice. It is to help physicians in this field as an automatic tool to support their diagnosis and decision for further investigation. The feature extraction technique has been applied on the voice signal in the time domain and in the frequency domain. Time domain features used in this research are: Zero Crossing Rate (ZCR) and Short time Energy. Frequency domain features used are: Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). Classification was based on threshold detection of each feature or group of features.
1.3 Research Outline

This chapter introduces the problem statement and the solutions to develop in this project.

Chapter two presents the basic medical background of this research.

Chapter three presents a literature review done in the voice pathology recognition area.

Chapter four illustrates in details the proposed system. The chapter explains features used in the feature extraction stage.

Chapter five discusses the experimental work and results.

Finally, Chapter six concludes the contributions, list the major challenges and drawbacks in our work and provides possible future research directions of the current work.
CHAPTER 2

MEDICAL BACKGROUND

Speech production is a very important process. It involves many different but coordinated phases. The movement of the lips, tongue, and other organs is among the subtlest and most adept of any actions performed by human beings. Here, the mechanism of speech production will be discussed which includes the human vocal organs and the discrete-time speech production model.

2.1 Anatomy of the speech production system

Speech is produced by the vocal apparatus as shown in figure 2.1, which can be divided into three parts:

- The sub-glottal part (lungs and trachea).
- The vibratory part (larynx and vocal folds).
- The supra-glottal part (pharynx, oral and nasal cavities).

The following description is inspired from M.Behlau et al. [4].

2.2 The sub-glottal part

The first component of this apparatus is the lungs as shown in Figure 2.2, that provide the necessary air and that can thus be described as the “generator”. They play a major role in the exchanges of oxygen and carbon dioxide (CO₂) between the air and the blood during the respiration. The lungs can inflate or deflate by increasing or decreasing the volume of the thoracic cavity, due to the action of the diaphragm and the thoracic
muscles. The trachea is a tube linking the lungs and the larynx whose walls are reinforced by about 20 C-shaped rings.

During inspiration, the volume of the lungs increases and their inner pressure decreases, inducing an aspiration of the air from the outside to the lungs by passing through the trachea. The increasing volume of the lungs and the thoracic cavity induces a stretching of both the elastic tissue in the lungs and the muscles of the rib cage.

Figure 2.1: The vocal apparatus (sagittal view), [4].
Figure 2.2: The lungs and the bronchi, [4].

During expiration, the volume of the thoracic cavity and the lungs decreases, this corresponds to an increase of the inner pressure of the lungs, which forces air to come out. During quiet respiration, the time taken for the inspiration and expiration are approximately equal and both phases form a rhythmic process comprising from 12 to 15 cycles per minute.

In case of speech production, the situation is quite different. The inspiration phase is more rapid than in quiet respiration and deeper in terms of volume of inhaled air. The expiration phase (during which speech is produced) is on the other hand longer. During this phase, the airstream coming from the lungs is transformed into a series of pulses by constriction at the vocal folds in the larynx or any other place in the vocal tract.
2.2.1 The vibratory part

The larynx as shown in figure 2.3, commonly called the "voice box," is a tube shaped structure comprised of a complex system of muscle, cartilage, and connective tissue. The Larynx serves a number of purposes, ones that are essential to life. These purposes are called "biological", while speaking and singing are called "non-biological".

A. Biological Function:

- To act as a valve to prevent air from escaping the lungs, e.g. weightlifting.
- To prevent foreign substances from entering the lungs, trachea and glottis, e.g. while swallowing, the epiglottis covers the opening to the larynx.
- To forcefully expell foreign substances which threaten the trachea, e.g. coughing.

B. Non-Biological Function:

- The production of sound.
2.2.2 The cartilages of the larynx skeleton

- The thyroid cartilage as in figure 2.3.a - the "adam's apple" on men, this V shaped cartilage features a notch in the front which can be felt with the edge of your thumb.
- The cricoid cartilage as in figures 2.3.a and 2.3.c is a ring shaped cartilage connected to the trachea.
- The Arytenoid cartilages as in figure 2.3.b are pyramid shaped and sit behind the thyroid cartilage. The vocal folds are attached to these cartilages and it is their movement that opens and closes the glottis (the space between the vocal folds).
- The epiglottis as in figures 2.3.a, 2.3.b and 2.3.c is a leaf-shaped elastic cartilage attached to the internal surface of the thyroid cartilage. This cartilage closes the opening of the larynx when we swallow food or drink.
The cartilages and membranes forming the skeleton framework of the larynx surround the laryngeal cavity which is continuous with the trachea and the laryngopharynx as shown in figure 2.4. This cavity is narrowed at two points by two pairs of mucous folds: the vestibular or false vocal folds and the true vocal folds.

The true vocal folds are concerned with the production of sound. They contain two vocal ligaments which are made of elastic tissue, pearly white in colour, 20 to 25 millimetres long, and coated with mucus. Each ligament is attached in front to the thyroid cartilage and behind to the arytenoid cartilage. The space between the two true vocal folds is known as the glottis.
2.2.4 The supra-glottal part

The supra-glottal part (also called vocal tract) comprises the oral and nasal cavities and the pharynx as shown in figure 2.1. The oral cavity is demarcated by the lips, the palate and the tongue. The nasal cavities are the incoming point of air in the respiratory system and are in charge of the warming and humidification of the air. The pharynx is a muscular tube attached to the base of the skull and continuous with the oesophagus. It can be divided in three parts: the nasopharynx (behind the nasal cavities), the oropharynx (behind the mouth) and the laryngopharynx (behind and around the laryngeal inlet).

2.3 The process of speech production

Speech production involves the coordinated movement of a number of structures within the head, neck and thorax. Three distinct stages are necessary for the production of speech: the generation of an air stream (expiration); the conversion of this air stream into a series of vibrations (the phonation) and the modification of these vibrations (the articulation).
2.3.1 The vibration of the vocal folds

During expiration, the air is expelled from the lungs through the trachea. If the vocal folds are maintained closed by the action of the intrinsic laryngeal muscles, the subglottal pressure rises. When this pressure is sufficient enough to overcome the muscular force, the vocal folds are forced to open a little and a small amount of air is released into the vocal tract, leading to a decrease of the pressure below the folds. As the muscular action exceeds the force produced by the subglottal force, the vocal folds close, aided by the Bernoulli effect and by their natural elastic recoil. This cycle repeats at a certain frequency (the fundamental frequency) until the vocal folds are relaxed or when there is no longer airflow [5].

Figure 2.5 shows that vocal fold vibration repeats four phases within a cycle: the closed phase, opening phase, open phase, and closing phase. Voiced speech sounds like vowels, caused by the vibrating vocal cords, resulting in quasi-periodic speech waveforms. The fundamental frequency ($F_0$) of vibration depends on the mass and tension. It is about 80–400Hz in males, and about 120–800Hz in females [5].

For most of the speech sounds, we can assume that the general properties of excitation and vocal tract are fixed over a period of 10 – 20 msec. Vowel sounds are usually used for laryngeal function assessment because they reflect the physical condition of vocal folds where it vibrates at a sustained frequency.
2.3.2 Articulation

The spectrum of the source signal has to be modified to produce intelligible sounds. This is achieved by the articulation, the pharynx and the other cavities with which it communicates (the nasal cavities, mouth, and larynx) act as a “resonator” that alters the sounds issuing from your vocal folds, amplifying some frequencies while attenuating others.

The transformation of the sounds from the larynx is then completed by the position of the soft palate, tongue, teeth, lips, and other parts of the mouth, which act as “modulators” aiming at modifying the size and the shape of the different parts of the vocal tract. The resulting wave is emitted to the outside by the lips.

2.4 Pathologies of the larynx

Dysphonia is the medical term for disorders of the voice, is defined as impairment in the ability to produce voice sounds using the vocal organs. The dysphonic voice can be hoarse or excessively breathy, harsh,
or rough. Different types of dysphonia exist. First, organic dysphonia are caused by pathological changes to the vocal folds. Second, neurological dysphonia are caused by some problem in the nervous system as it interacts with the larynx. Third, dysfunctional dysphonia are characterized by difficulties in phonation without obvious organic alteration of the vocal folds.

2.4.1 Organic Dysphonia

This kind of voice disorder is essentially caused by morphological changes in the anatomy of the larynx, especially at the glottal level. A very common organic dysphonia is the laryngitis. Caused by infectious agents (viruses/bacteria) or prolonged speaking sessions, this consists in an inflammation of the larynx. It provokes an asymmetrical vibration of the vocal folds and an increasing of their mass and stiffness. Typical symptoms are hoarseness and soreness.

When focusing on the vocal folds, another common pathology is the vocal polyps and nodules, often resulting from the abuse or overuse of the voice like teachers and singers. Both polyps and nodules arise at the portion of the vocal folds that is the most sought, namely the place of glottal closure. Appearing on one of the vocal folds, polyps take the form of red or whitish fleshy masses, fixed to the vocal folds or linked to them by a stalk. Contrary to the polyps, the nodules may appear symmetrically on the two folds as firm whitish lumps.

Both polyps and nodules have the same effect, in the sense that the closure cannot be complete due to the presence of the lumps, they prevent the vocal folds from meeting in the midline. This results in a breathy voice
due to an air leakage during phonation (leading to the presence of noise in the produced speech) and a modification of the mass and the elasticity of the vocal folds (leading to hoarseness and raspy voice).

Another pathology that prevents normal vibration of vocal folds is a Cyst, it growth beneath the surface layer of the vocal fold mucosa, it causes a gap between the two vocal folds. This results in a breathy, rough and hoarse voice.

The vibration of the vocal folds can also be altered by the presence of Edema. For instance, the Reinke’s Edema is located at one or two vocal folds and consists in a swelling of the entire layer of the superficial lamina propria (or Reinke’s layer). It occurs exclusively in smokers, and some have proposed that it is a reaction to repeated exposure of the vocal folds to the heat of inhaled cigarette smoke or drugs. Although it does not prevent the glottal closure, the Edema influences the vibration of the folds by increasing the mass of the vocal fold on which it is located, inducing low-pitched and a hoarseness in the produced voice.

In the most severe cases, Laryngeal cancer is a change of normal tissue to tissue that grows uncontrollably at the vocal folds, in the form of a squamous cell carcinoma that begins at the surface of the larynx. If unchecked, it spreads into deeper tissues. The cancer can be surgically removed and the vocal preserved. Smoking is responsible for the majority of laryngeal cancers. Cancer causes hoarseness, even when it is quite small.

2.4.2 Neurological Dysphonia

On the other hand vocal fold paralysis or paresis, one or both vocal folds do not move, results from a lesion of the neural or muscular
mechanism, often causing a gap between the two vocal folds which allows air to leak through and disrupts vibration. The sound of the voice may be weak, breathy, rough, diplophonic (two pitches occurring at the same time), or just a whisper.

Vocal fold paralysis can result from damage to one of the two nerves that go from the brain to the larynx. Also, it may come from an error after surgical operations such as after laser cordectomy, after total thyroidectomy and after total Para thyroidectomy.

2.4.3 Dysfunctional Dysphonia

Dysfunctional dysphonia are caused by poor muscle functioning. Hyperadduction and hyperabduction are two examples of dysfunctional dysphonia. In hyperadduction where the vocal folds adduct (come together) very tightly, producing a valve that restricts airflow inducing a poor voice quality. In hyperabduction where the vocal folds abduct (do not come together) to produce voice, it causes a weak and a breathy voice.

2.5 Clinical assessment of voice

As voice is a complex acoustic phenomenon, and voice production involves the interplay of different anatomic structures and physiologic systems. Therefore, characterization and quantification of the voice is a challenging and multidimensional undertaking. When feeling trouble with his voice, the patient has to undergo a clinical assessment in order to qualify and quantify the magnitude of the problem. Vocal function testing is used to assess the severity degree of the dysphonia or the efficiency of a treatment.
Voice assessment techniques may be categorized into five groups. Two involve the subjective judgments of humans: patient self-assessment and perceptual analysis. The other three categories consist of objective physical measurements obtained during phonation, including acoustic analysis, aerodynamic measurement and vocal fold motion measurement.

2.5.1 Patient self-assessment

This evaluation is performed by the patient himself. An example of evaluation is the Voice Handicap Index (VHI) as in Jacobson et al. [7], consisting in asking 30 questions to the patient, for which the answer can range from 0 (normal) to 3 (highly disturbed). These questions include 10 of physical, functional, and emotional occurrences. VHI is a statically robust instrument and it may prove to be useful clinically and in research.

2.5.2 Perceptual analysis

In this task, the dysphonia is rated in terms of perception. For this, the speech therapists examine different perceptual scales of the voice. The most used scales in practice are the Hirano’s GRBAS scales which proposed in Hirano [8]; each scale is rated from 0 (normal) to 3 (severe):

- Grade (G): the overall degree of voice abnormality.
- Roughness (R): audible impression of glottal cycles irregularities and abnormal fluctuations of fundamental frequency.
- Breathiness (B): audible impression of air leakage during phonation.
- Asthenia (A): voice weakness (hypotonic voice).
- Strain (S): forced voice (hypertonic voice).
2.5.3 Aerodynamic measurement

These measures form a quantitative evaluation of the vocal apparatus behavior during phonation. They are directly performed by the clinician or measured by dedicated systems, such as Multi Dimensional Voice Program (MDVP) system:

- Maximum Phonation Time: time during which the patient is able to maintain a vowel after a maximal inspiration.
- Phonatory Volume: volume of air really expired during the measure of the maximum phonation time.
- Mean Flow Rate: ratio between the phonatory volume and the maximum phonation time.
- Subglottal Pressure: estimated by a measuring the inner-mouth pressure during the production of a particular syllable.
- Quotient of Phonation: ratio between the vital capacity (maximum amount of air that can be inspired) and the maximum phonation time.

2.5.4 Acoustic analysis

Acoustic analysis is widely used in testing vocal function because it is objective, non-invasive, and may be easily performed using many available commercial systems. Acoustic features are computed in order to attempt to quantify the voice disorder. Using acoustic algorithms analyze the complex sound waves of the voice in the domains of frequency, intensity, and time. These features are computed on a fragment of sustained vowel /a/ using dedicated systems such as MDVP.
They mainly consist in:

- Fundamental frequency as well as its extreme variations and irregularity (jitter).
- Extreme variations and irregularity of intensity (shimmer).
- Harmonic to noise ratio (HNR): ratio between energy of the harmonic and non harmonic part of the signal in order to reflect the presence of noise during phonation.

2.5.5 Vocal fold motion measurement

Speech production mechanisms arise from the functions of the internal organs of the human body that are mostly invisible. Therefore, better understanding of speech production processes relies on the development of observation techniques [5].

Differences often occur between what is heard by the clinician and what can be viewed by looking at the vocal folds movement. This phenomenon is however too fast to be correctly observed by human eye. Imaging of the vocal folds during speech has been conducted with a combination of an endoscope and video camera as shown in figure 2.7. A solid-type endoscope is capable of observing vocal fold vibration with stroboscopic or real-time digital imaging techniques during sustained phonation [5].
An electroglottograph (EGG) as shown in figure 2.8 is a non-invasive device that indexes the contact area between the two vocal folds. A small, high-frequency current is passed between two electrodes that are secured around the neck at the level of the larynx. The opening and closing of the vocal folds causes variation in the electrical resistance of the current. These changes in resistance are then displayed on screen [5].
Although videostroboscopy greatly expands the diagnostic sensitivity of some aspects of office-based laryngoscopy, its interpretation depends on the skill and experience of the clinician performing the study, and specifically the skill and experience of the diagnostic interpreter. The quality of the images collected is directly related to the skill of the operator performing the procedure.

Several technologies have been developed to improve objective measurements. Software was developed to measure the glottic-area waveform (GAW), a plot of the glottic area against the time of opening and closing of the glottis during a representative vibratory cycle (taken from the stroboscopic image). From this information, glottal opening and closing rates are calculated. These measurements are purported to be correlates of vocal-fold pliability and differ statistically in preoperative and postoperative states for benign vocal-fold lesions.
CHAPTER 3

LITERATURE REVIEW

The field of voice pathologies recognition using information from the speech signal has been investigated over the past 20 years, either for classifying normal or pathological voice or for separating different pathologies.

In voice pathology recognition systems as shown in figure 3.1; the speech waveform needs to be converted into digital format before it is suitable for processing in the speech recognition system. The features are extracted from the speech signal and used as inputs of a classifier in order to provide a decision about the presence of pathology in the vocal folds.

Figure 3.1: Common structure of voice pathology recognition systems.
As the features and the classifiers are fundamental in such a system, section 3.1 is devoted to the description of the features used to provide a decision about the pathological nature of the voice, and are often cited in literature. These features are described and their relation to voice pathology is discussed.

3.1 Classical features in voice pathology assessment

This section gives an overview of the most frequently used features in voice pathology assessment.

3.1.1 Fundamental frequency

The fundamental frequency ($F_0$) is surely the most obvious feature for research on pathologies, mainly because it is assumed to reflect the contribution of the vocal folds behavior in the speech signal. It’s a function of the mass, elasticity and length of vocal folds as in Davis [9]. The fundamental frequency is the basis of many features in voice pathology assessment such as the jitter or some formulations of the harmonic-to-noise ratio. Estimating the fundamental period is a basic problem in speech processing for which many algorithms have been proposed, with various assumptions and robustness.

3.1.2 Cepstral features

The cepstrum is a transformation of a signal $x(t)$ from the temporal domain to another domain, similar to the temporal domain. It is defined as the inverse Fourier Transform of the log spectrum of $x(t)$:
\[ C(x(t)) = F^{-1}(\log(F(x(t)))) \] 

(3.1)

Where \( F(x(t)) \) is the fast fourier transform of the signal \( x(t) \).

One of the most widely used cepstral representations for speech is obtained by computing the Mel-Frequency Cepstral Coefficients (MFCCs). They are used for instance in speech recognition as in [11]. These coefficients are computed for each speech frame by weighting the magnitude spectrum by a mel-filterbank as depicted in figure 3.2.

The term mel refers to a kind of measurement related to perceived frequency. The mapping between the real frequency scale (Hz) and the perceived frequency scales (mels) is approximately linear below 1 kHz and logarithmic at higher frequencies as shown in figure 3.3. The suggested formula that models this relationship is described in Deller et al. [10] as follows:

\[ F_{mel} = 2595 \cdot \log_{10} \left( 1 + \frac{f}{700} \right) \text{ where } f \text{ is the real frequency (Hz)} \]

Figure 3.2: Weighting functions for Mel-scale filtering.
Then compute the log of each filter output and finally compute the Discrete Cosine Transform (DCT) of the log-mel-spectrum. The MFCCs are the resulting coefficients of this DCT operation.

Considering the MFCCs provides several advantages: human perception is taken into account by considering a perceptive scale of frequencies; the spectral envelope of the speech frame is summarized into a limited number of coefficients.

As MFCCs are widely used in various speech processing fields, it is normal that a certain amount of studies includes these coefficients in the input features set of classifiers. For instance, Godino-Llorente et al. [11] aim at classifying between normal and pathological samples. MFCCs used as the inputs of respectively a Multi Layer Perception (MLP) or a Gaussian Mixture Model (GMM) while MFCCs and other acoustic features are used as inputs of a Support Vector Machines classifier (SVM) in Godino-Llorente et al. [12].

Dibazar et al. [13], [15], and Dibazar and Narayanan [14], used 12 MFCCs, the fundamental period and their respective first and second
derivatives in order to capture source information by means of the fundamental period and the spectral envelope information by means of the MFCCs. The obtained features were used to train HMMs in order to discriminate between normal and pathological samples, as well as some particular pathologies.

Fredouille et al. [16], aimed to determine the grade corresponding to a particular voice sample. For this 16 MFCCs and their first derivatives were computed using a 24 mel -filterbank and are taken as input of GMM designed for modeling each grade. The study presented by Fredouille et al. [17], also aimed at training GMMs but this time for both the normal/pathological classification and the classification of the grades and by using 24 MFCCs as well as their first and second derivatives for the particular case of the classification of different types of dysarthria. Finally, 11 MFCCs as well as their first derivatives were extracted from voice samples by Bocklet et al. [18] in order to train a GMM aiming at determining an intelligibility score for each pathological speaker.

### 3.1.3 Acoustic features computed in MDVP software

The Multi Dimensional Voice Program (MDVP) is a software produced by KayPentax Corp for clinical purposes. This software is used to assess the production of a patient by computing acoustic descriptors related to the fundamental frequency and the amplitude of the signal. It calculates over 30 parameters, a brief description about these parameters can be found in [19]:

- $F_0$: Average fundamental frequency for the vocalization.
- **$T_0$:** Average fundamental period for the vocalization.
- **$F_{hi}$:** Highest fundamental frequency for the vocalization.
- **$F_{lo}$:** Lowest fundamental frequency for the vocalization.
- **$STD$:** Standard deviation of the fundamental frequency for the vocalization.
- **$PFR$:** Phonatory fundamental frequency range in semi-tones.
- **$F_{ftr}$:** Frequency of the most intensive low-frequency $F_0$ modulating component.
- **$F_{atr}$:** Frequency of the most intensive low-frequency amplitude modulating component.
- **$Jit_a$:** Absolute jitter, computed as the period-to-period variability of the fundamental period (s).
- **$Jitt$:** Jitter, computed as the percentage of period-to-period variability of the fundamental period (%).
- **$RAP$:** Relative average perturbation computed as the period-to-period variability of the fundamental period with a smoothing factor of 3 periods (%).
- **$PPQ$:** Pitch period perturbation quotient computed as the period-to-period variability of the fundamental period with a smoothing factor of 5 periods (%).
- **$sPPQ$:** Pitch period perturbation quotient computed as the period-to-period variability of the fundamental period with a smoothing factor of 55 periods (%).
- **$vF_0$:** Relative standard deviation of the period-to-period computed fundamental period.
- **$Shim_a$:** Absolute shimmer computed as the period-to-period variability of the peak-to-peak amplitude.
- **Shim**: Shimmer, computed as the percentage of variability of the peak-to-peak amplitude from a period to another (%).
- **Shim\(_{dB}\)**: Absolute shimmer expressed in dB.
- **APQ**: Amplitude perturbation quotient, computed as the percentage of variability of the peak-to-peak amplitude with a smoothing factor of 11 periods (%).
- **sAPQ**: Smoothed amplitude perturbation quotient, computed as the percentage of variability of the peak-to-peak amplitude with a smoothing factor of 55 periods (%).
- **NHR**: Noise-to-Harmonic Ratio, computed as the ratio of the energy of the non-harmonic part of speech in the [1500-4500] Hz band to the energy of the harmonic part in the [70-4500] Hz band.
- **\(vA_m\)**: Standard deviation of the period-to-period computed peak-to-peak amplitude.
- **VTI**: Voice Turbulence Index, computed in stable phonation areas as the ratio of the energy of the non harmonic part in high frequencies to the energy of the harmonic part.
- **SPI**: Soft Phonation Index, ratio of the harmonic energy in low to high frequencies.
- **FTRI**: \(F_0\) Tremor Intensity Index, computed as the ratio of the frequency magnitude of the most intense low-frequency modulating component of \(F_0\) to the total frequency magnitude of the speech signal.
- **ATRI**: Amplitude Tremor Intensity, computed as the ratio of the amplitude of the most intensive low-frequency amplitude-modulating component of \(F_0\) to the total amplitude of the speech signal.
- **DVB**: Degree of Voice Breaks, computed as the ratio of the total length of the areas representing the voice breaks to the total duration of the speech signal.
- **DSH**: Degree of Sub Harmonics, computed as an estimation of the percentage of occurrence of sub-harmonics of $F_0$ in the speech signal.
- **DUV**: Degree of Unvoiced, computed as an estimation of the percentage of occurrence of unvoiced segments.
- **NVB**: Number of Voice Breaks, computed as how many times the estimated $F_0$ contour is interrupted from the beginning of the first to the end of the last voiced area.
- **NSH**: Number of Sub-Harmonic segments found during the analysis of the speech signals.
- **NUV**: Number of Unvoiced segments detected in the speech signal.
- **SEG**: Total number of segments considered by MDVP software during the autocorrelation analysis.
- **PER**: Pitch Periods detected during the period-to-period pitch extraction in MDVP.

As these features are considered as “classic” in the domain of voice pathology assessment, some authors directly used the values of features provided with the recordings of Kay Database in their method, as Godino-Llorente et al. [1], Dibazar et al. [14] and Godino-Llorente et al. [20]. Other authors used the MDVP software to compute the acoustic features from their own recordings, such as Godino-Llorente et al. [21] in which cepstral measures were added to the MDVP features set. Finally, some authors were inspired by the MDVP software and proposed their own implementation of
some MDVP features. For instance, the studies of Moran et al. [22] proposed a classification system between normal and pathological samples after transmission through a telephone channel. The involved features were smoothed with different factors as in the MDVP software.
CHAPTER 4

**The Proposed System**

The aim of this research is to propose a user friendly system for the discrimination between normal and diseased voice. It is to help physicians in this field as an automatic tool to support their diagnosis and decision for further investigation. The feature extraction technique has been applied on the voice signal in the time domain and in the frequency domain. Time domain features used in this research are: Zero Crossing Rate (ZCR) and Short time Energy. Frequency domain features used are: Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). Classification was based on threshold detection of each feature or group of features.

Figure 4.1 represents a block diagram of the different steps carried out in the process set up for the recognition of voice alterations. A description of each step is presented in the following sections.
Figure 4.1: Block diagram of the proposed system
4.1 Developed Software

We used Matlab to implement the software required for all stages. We built a Graphical User Interface (GUI) (Figure 4.2) which enables physicians and researchers to record voice signal of the patient and framing the signal in order to process this signal and get features.

![Snapshot from GUI](image)

Figure 4.2: Snapshot from GUI

Our software enables the user to record, play the signal, control its volume, regard the shape of the wave and its analysis, play any frame of his selection and change its size easily and choose any
feature from a given menu of features available to see its wave shape. It can get the sampling frequency, duration and no of channels of the given signal. Our software is able to recognize the diseased signal. This work can be applied in the field of preventive medicine in order to achieve early detection of laryngeal pathologies.

4.2 Speech Signal Acquisition

The voice Signal is recorded using a button in the GUI called “Record”. When clicking on this button an audio signal of 2 seconds is recorded from the microphone connected to the computer and saved on the hard disk. The default sampling rate in our software is 22050 HZ. Moreover, our software allows the user to load any existing audio signal with (*.wav) extension and process it using button called “Open Wav file”.

After the signal is recorded / loaded, the GUI shows its sampling frequency, duration and number of channels.

![Figure 4.3: Signal Acquisition in GUI](image)

Figure 4.3: Signal Acquisition in GUI
4.3 Signal Pre-processing

4.3.1 Pre-emphasis

There is a need to spectrally flatten the signal. A process called pre-emphasis is applied to emphasize the high-frequency portion of the spectrum because the boosting of high-frequency energy gives more information to the acoustic model. Pre-emphasis is accomplished by passing the signal through high-pass filter whose transfer function $H(z)$ is given by Rabiner and Huang [42]:

$$H(z) = 1 - az^{-1} \text{ where } 0.9 \leq a \leq 1$$

The value for the pre-emphasis parameter ‘a’ determined adaptively from literature review and researches to be 0.97 [24]

When the user clicks on the “Record” button in the GUI, the audio signal is being filtered automatically by passing the signal through high-pass filter with the transfer function: $H(Z) = 1-0.97 Z^{-1}$

Figure 4.4 illustrates the time representation of normal and pathological speech signal before and after pre-emphasis step.
4.3.2 Frame Blocking

The human speech signal is a slowly time varying signal and can be treated as a stationary random process when considered under a short time frame [24]. Therefore, the speech signal is usually separated into small duration blocks, called frames, and the spectral analysis is performed on these frames. The commonly used frame length is 10-30 milliseconds for speech recognition tasks because the positions of the articulators do not change much in the period of frame length [24]. Figure 4.5 shows the framing process on a continuous signal.
4.3.3 Windowing

After the speech signal being partitioned into frames, each frame is multiplied by a window function prior to the spectral analysis to reduce the effect of discontinuity introduced by the framing process.

4.3.4 Segmentation

After testing different segment lengths we selected a 400 ms segment in the middle of the vowel “ah”. User select the appropriate segment length by defining Time 1 and Time 2 so that the total time length is 400 ms in the middle of the vowel “ah”. Choosing 400 ms segment length will be illustrated in chapter 5 “Experimental work and Results” with comparison between different segment lengths and accuracy, sensitivity and specificity of each segment length.
Figure 4.6 shows the segmentation panel in our GUI.

![Segmentation Panel Image]

Figure 4.6: Segmentation panel

4.4 Feature Extraction

Feature extraction aims at giving a useful representation of the speech signal by capturing the important information from it. As stated before, we used two time domain features: Zero Crossing Rate (ZCR) and Short time Energy and two Frequency domain features: Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) in the feature extraction process.

4.4.1 Zero Crossing Rate (ZCR)

The zero crossing rate (expressed in Hz) is a time domain feature that defines the frequency at which the speech signal, represented in the time domain, crosses the zero-axis. ZCR is commonly used in endpoint detection, especially in detecting the start and end of unvoiced sound.

Noise, unvoiced and disordered voices have high ZCR. If a sample with zero value is considered a case of ZCR, then the value of ZCR is higher. Otherwise, it is lower.
The zero-crossing rate is computed here by counting the number of sign alterations between consecutive samples (from positive to negative and the opposite) in the frame and dividing this number by the duration of the frame.

**Algorithm used in calculating ZCR**

- Divide the speech signal into windows
- Compare successive samples in the window to find a transition from positive to negative
- Mark a transition as zero crossing
- Total number of zero crossings form a ZCR of the window
- Calculate normalized ZCR

- Repeat until all the windows are finished

We build a time domain feature extraction panel in the GUI, which help in selecting the time domain feature and calculating statistics for the selected segment. User selects from 2 popup menus the time domain feature and the statistic. Features menu includes ZCR and short time energy, Statistics menu includes mean, median, standard deviation (std), max, standard deviation by mean (stdbymean) and min. Figure 4.8 shows the feature extraction panel in which a user selects the ZCR from features and mean from statistics.

![Feature Extraction Panel](image)

**Figure 4.8: Feature panel**

Figure 4.9 shows calculated ZCR mean for a 400 ms segment of normal audio signal.
Figure 4.9: ZCR mean for a 400 ms segment of normal audio signal

Figure 4.10 shows calculated ZCR max for a 400 ms segment of normal audio signal
One of the simplest representations of a signal is its energy. In the case of a real discrete time signal $x(n)$ the energy is defined in general as:

$$E = \sum_{n=-\infty}^{\infty} x^2(n)$$

For no stationary signals such as speech, it is often more appropriate to consider a time varying energy calculation such as the following:

Figure 4.10: ZCR Max for a 400 ms segment of normal audio signal
\[ E(n) = \sum_{m=0}^{N-1} [w(m)x(n-m)]^2 \]

where \( w(m) \) is a weighting sequence or window which selects segment of \( x(n) \), and \( N \) is the number of samples in the window. For the simple case of \( w(m) = 1 \), \( E(n) \) is the sum of the squares of the \( N \) most recent values of \( x(n) \).

One difficulty with energy measurements is that they are very sensitive to large signal levels (because they enter the computation as a square), thereby emphasizing large sample to simple variations in \( E(n) \).

Figure 4.11 shows the short time energy feature from the feature menu with the mean from the statistics menu.

![Feature panel](image)

**Figure 4.11:** Feature panel
4.4.3 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is a well known and popular technique in speech recognition [24]. MFCC’s are based on the known variation of the human ear’s critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [24]. The process of computing MFCCs is described in more detail next.

Figure 4.12: Energy mean for a 400 ms segment of normal audio
Let $x(n)$ denotes the input speech signal. The complete calculation process of the coefficients can be described in the next three steps as follows [42]:

Step 1) Transform the input speech signal from time to frequency domain by short-time FFT

$$X(k) = \sum_{n=0}^{N-1} x(n).w(n).e^{-j2\pi nk/N}$$

Where $n = 0, 1, 2... N-1$; $N$ is the frame size; $w(n)$ is the Hamming window function.

Step 2) Find the energy spectrum, $E(k)$, of each frame, and calculate the energy $E_m$ in each mel window

$$E(k) = |x(k)|^2, \quad E_m = W_m(K).E(K)$$

Where $1 \leq m \leq M$; $M$ is the number of the mel windows in mel scale, which generally ranges from 20 to 24. $W_m(k)$ is the triangular weighted function associated with the $m$th mel window in mel scale.

Step 3) Taking logarithms and applying Discrete Cosine Transform (DCT) to the results, the mel frequency cepstral coefficients $C_j$ [11] are:
\[ C_j = \sum_{i=1}^{M} \log(E_m) \cdot \cos(j(i - 0.5)) \]

Where \(1 \leq j \leq L\); \(L\) is the desired order of the MFCC.

Figure 4.13 shows the block diagram of MFCC calculation steps.
Figure 4.13: MFCC Calculation

- Speech signal
- Pre-emphasis
- Frame blocking
- Windowing
- FFT
- Mel Filter Bank
- Natural log
- DCT
- Calculate Energy
- Energy
- MFCC (C1-C12)
- Form the 13-element feature vector
- Calculate the first derivative (26-element)
- Calculate the second derivative (39-element)

MFCC
We built a frequency domain feature extraction panel contains 2 buttons return the value of LPC and MFCC. Figure 4.14 shows the frequency domain features panel in the GUI.

![Frequency Domain Features Panel](image)

Figure 4.14: Frequency domain features panel

### 4.4.4 Linear Predictive Coding (LPC)

Linear prediction is a good tool for analysis of speech signals. Linear predication models the human vocal tract as an infinite impulse response (IIR) system that produces the speech signal [24]. In speech coding, the success of LPC have been explained by the fact that an all pole model is a reasonable approximation for the transfer function of the vocal tract. All pole model is also appropriate in terms of human hearing, because the ear is more sensitive to spectral peaks than spectral valleys. Hence an all pole model is useful not only because it may be a physical model for a signal, but because it is a perceptually meaningful parametric representation for a signal.[24]

The LPCs methods are based on the fact that the signal can be approximated from a weighted sum of precedent samples. This approximation is given by:
where \( a_k \) (\( 1 < k < p \)) is a set of real constants known as predictor coefficients, that must be calculated, and \( p \) is the predictor order. The problem of linear prediction resides on finding the predictor coefficients \( a_k \) that minimize the error between the real value of the function and the approximated function.

To minimize the total quadratic error is necessary to calculate the autocorrelation coefficients. This is a matrix equation with different recursive solutions, the commonly used is the Levinson recursion.

The number of predictor coefficients is obtained by substituting the sampling frequency value (\( f_s \)):

\[
p = 4 + \frac{f_s}{1000} = 4 + \frac{8000}{1000} = 12
\]

The sequence of the minimal error could be interpreted as the output of the \( H(z) \) filter when it is excited by the \( S_n \) signal. \( H(z) \) is usually known as an inverted filter. The approximated transfer function could be obtained if it is assumed that the transfer function \( S(z) \) of the signal is modeled as an only pole filter with the form:
\[
\hat{S}(z) = \frac{A}{H(z)} = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}
\]

**LPC advantages:**

- LPC provides good model of speech signal
- Provides linear characteristics
- LPC leads to a reasonable source-vocal tract separation
- LPC is analytically tractable model
- The method of LPC is mathematically precise and straightforward to implement in either software or hardware

**LPC disadvantages:**

- A serious problem with the LPC is that they are highly correlated but it is desirable to obtain less correlated features for acoustic modeling.
- An inherent drawback of conventional LPC is its inability to include speech-specific a priori information in the modeling process.
4.5 Classification

Classification was based on threshold detection of each feature and group of features.

Classification process is to be discussed in chapter 5 “Experimental work and Results”
Chapter 5

EXPERIMENTAL WORK and RESULTS

5.1 Datasets

The database used in this work was developed and collected in a sound proof room of the Phoniatics department of Kobri Elkobba military hospital. The acoustic samples correspond to sustained phonations (1-3 sec long) of vowel /ah/ from patients (males and females) with wide variety of vocal folds disorders.

The files were obtained with low noise level, constant microphone distance around 15 cm from the talker's lips, and 22050 Hz sampling rate. The experiments were done on 50 voice signals, 45 voice signals recorded in the hospital (15 Normal and 30 Diseased) in addition to 5 signals recorded with the developed software.

5.2 Experimental Work

After finishing signal acquisition and signal preprocessing processes we started the feature extraction process. We applied feature extraction process on 4 different features, two time domain features: Zero Crossing Rate (ZCR) and Short time Energy., two frequency domain features: Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC).

5.2.1 Segmentation process

In our developed software we selected 400 ms segment length after testing 4 different segment lengths 200, 300, 400 and 500 ms.
We started with 200 ms segment length and checked the accuracy, sensitivity and specificity for the selected features, after that we tried another frame length of 300 ms and the accuracy, sensitivity and specificity increased for some features, then we increased the segment length to 400 ms and we observed that the accuracy, sensitivity and specificity increased again but when choosing 500 ms segment length the accuracy, sensitivity and specificity dropped. The following sentences show the definitions for sensitivity and specificity:

1. Sensitivity (SE): Likelihood that an event will be detected given that it is present which is in our method the number of errors in the normal voice signals.

2. Specificity (SP): Likelihood that the absence of an event will be detected given that it is absent which is in our method the number of errors in the disordered voice signals.

The following tables show the results for calculations of accuracy, sensitivity and specificity for different segment lengths.

Table 5.1 accuracy, sensitivity and specificity for 200 ms segment length

<table>
<thead>
<tr>
<th>Feature</th>
<th>sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy mean</td>
<td>95%</td>
<td>60%</td>
<td>74%</td>
</tr>
<tr>
<td>ZCR Max.</td>
<td>70%</td>
<td>83%</td>
<td>78%</td>
</tr>
<tr>
<td>ZCR Mean</td>
<td>85%</td>
<td>77%</td>
<td>80%</td>
</tr>
<tr>
<td>LPC</td>
<td>20%</td>
<td>80%</td>
<td>56%</td>
</tr>
<tr>
<td>MFCC</td>
<td>0%</td>
<td>90%</td>
<td>54%</td>
</tr>
<tr>
<td>Average</td>
<td>54%</td>
<td>78%</td>
<td>68%</td>
</tr>
</tbody>
</table>
Table 5.2 accuracy, sensitivity and specificity for 300 ms segment length

<table>
<thead>
<tr>
<th>Feature</th>
<th>sensitivity</th>
<th>Specificity</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy mean</td>
<td>75%</td>
<td>67%</td>
<td>70%</td>
</tr>
<tr>
<td>ZCR Max.</td>
<td>75%</td>
<td>87%</td>
<td>82%</td>
</tr>
<tr>
<td>ZCR Mean</td>
<td>80%</td>
<td>73%</td>
<td>76%</td>
</tr>
<tr>
<td>LPC</td>
<td>55%</td>
<td>60%</td>
<td>58%</td>
</tr>
<tr>
<td>MFCC</td>
<td>45%</td>
<td>63%</td>
<td>56%</td>
</tr>
<tr>
<td>Average</td>
<td>66%</td>
<td>70%</td>
<td>68%</td>
</tr>
</tbody>
</table>

Table 5.3 accuracy, sensitivity and specificity for 400 ms segment length

<table>
<thead>
<tr>
<th>Feature</th>
<th>sensitivity</th>
<th>specificity</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy mean</td>
<td>85%</td>
<td>70%</td>
<td>76%</td>
</tr>
<tr>
<td>ZCR Max.</td>
<td>80%</td>
<td>90%</td>
<td>86%</td>
</tr>
<tr>
<td>ZCR Mean</td>
<td>100%</td>
<td>67%</td>
<td>80%</td>
</tr>
<tr>
<td>LPC</td>
<td>75%</td>
<td>87%</td>
<td>82%</td>
</tr>
<tr>
<td>MFCC</td>
<td>60%</td>
<td>57%</td>
<td>58%</td>
</tr>
<tr>
<td>Average</td>
<td>80%</td>
<td>74%</td>
<td>76%</td>
</tr>
</tbody>
</table>

Table 5.4 accuracy, sensitivity and specificity for 500 ms segment length

<table>
<thead>
<tr>
<th>Feature</th>
<th>sensitivity</th>
<th>specificity</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy mean</td>
<td>60%</td>
<td>77%</td>
<td>70%</td>
</tr>
<tr>
<td>ZCR Max.</td>
<td>70%</td>
<td>83%</td>
<td>78%</td>
</tr>
<tr>
<td>ZCR Mean</td>
<td>90%</td>
<td>67%</td>
<td>76%</td>
</tr>
<tr>
<td>LPC</td>
<td>30%</td>
<td>87%</td>
<td>64%</td>
</tr>
<tr>
<td>MFCC</td>
<td>0%</td>
<td>83%</td>
<td>50%</td>
</tr>
<tr>
<td>Average</td>
<td>50%</td>
<td>79%</td>
<td>68%</td>
</tr>
</tbody>
</table>
The following figures show the results of calculations of accuracy, sensitivity and specificity for the features (Energy Mean, ZCR Max, ZCR Mean, LPC, MFCC) calculated in different segment lengths (200, 300, 400 and 500 ms).
Figure 5.1 Accuracy VS Segment length

(Bars at each segment length represent Energy Mean, ZCR Max, ZCR Mean, LPC and MFCC from left to right Respectively)
Figure 5.2 Sensitivity VS Segment length

(Bars at each segment length represent Energy Mean, ZCR Max, ZCR Mean, LPC and MFCC from left to right Respectively)
Figure 5.3 Specificity VS Segment length

(Bars at each segment length represent Energy Mean, ZCR Max, ZCR Mean, LPC and MFCC from left to right Respectively)
After analyzing those data, our decision was to choose 400 ms segment in the middle of vowel “ah” as this segment gives the best accuracy and it gives relatively high sensitivity and specificity.

5.2.2 Feature Extraction

After choosing the appropriate segment length, we started to analyze the features on the different audio signals in the dataset.

5.2.2.1 Short Time Energy

We started our feature extraction process by analyzing short time energy mean for all the voice signals in the dataset. Firstly, analyzing the normal (healthy) signals and then analyzing the disordered ones.

We applied a systematic technique on all signals which is choosing 400 ms segment length in the middle of the vowel “ah” and from the time domain feature extraction panel in the developed software we chose “short time energy” and from the statistics “mean”. According to the literature, the mean of short time energy gives a significance between normal and disordered voice signals. [24]

After analyzing all the signals -Normal and disordered- we tested different thresholds to find the best threshold to classify the signals into normal and disordered ones. In this process we are also observing the accuracy, sensitivity and specificity for each threshold to get the threshold which achieves the best accuracy and in the same time preserves a high sensitivity and specificity. Table 5.5 shows accuracy, sensitivity and specificity for different thresholds in the short time energy mean feature.
Table 5.5 Accuracy, Sensitivity and Specificity for different thresholds in the short time energy mean feature

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100%</td>
<td>0%</td>
<td>40%</td>
</tr>
<tr>
<td>0.02</td>
<td>100%</td>
<td>0%</td>
<td>40%</td>
</tr>
<tr>
<td>0.04</td>
<td>95%</td>
<td>30%</td>
<td>46%</td>
</tr>
<tr>
<td>0.06</td>
<td>90%</td>
<td>60%</td>
<td>72%</td>
</tr>
<tr>
<td>0.07</td>
<td>85%</td>
<td>70%</td>
<td>76%</td>
</tr>
<tr>
<td>0.08</td>
<td>55%</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>0.1</td>
<td>20%</td>
<td>87%</td>
<td>60%</td>
</tr>
<tr>
<td>0.12</td>
<td>10%</td>
<td>97%</td>
<td>62%</td>
</tr>
<tr>
<td>0.14</td>
<td>5%</td>
<td>100%</td>
<td>62%</td>
</tr>
<tr>
<td>0.16</td>
<td>5%</td>
<td>100%</td>
<td>62%</td>
</tr>
<tr>
<td>0.18</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
</tbody>
</table>
Figure 5.4 shows the Short time energy mean accuracy, sensitivity and specificity curve for different thresholds.
The following figure shows the different energy mean values for all signals in the datasets.

![Energy Mean](image)

Figure 5.5 Energy mean values

After analyzing these data, we chose a condition which is:

**For the normal voices:**

Energy mean $> 0.07$ \hspace{1cm} (5.1)

This condition gives the following percentages:

- **Accuracy**: 76 %
- **Sensitivity**: 85 %
- **Specificity**: 70 %
Those percentages based on testing 50 different voice signals, 30 diseased signals and 20 normal ones.

5.2.2.2 Zero Crossing Rate (ZCR)

We repeated the same systematic technique on all signals which is choosing 400 ms segment length in the middle of the vowel “ah” and from the time domain feature extraction panel in the developed software we chose “ZCR” and from the statistics “mean” and “max”. According to the literature, the mean and the max of ZCR give a significance between normal and disordered voice signals.

After analyzing all the signals -Normal and disordered- we tested different thresholds to find the best threshold to classify the signals into normal and disordered ones. In this process we are also observing the accuracy, sensitivity and specificity for each threshold to get the threshold which achieves the best accuracy and in the same time preserves a high sensitivity and specificity.

Table 5.6 shows accuracy, sensitivity and specificity for different thresholds in the ZCR, Max:
Table 5.6 Accuracy, Sensitivity and Specificity for different thresholds in the ZCR Max

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>0.1</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>0.2</td>
<td>55%</td>
<td>94%</td>
<td>78%</td>
</tr>
<tr>
<td>0.23</td>
<td>80%</td>
<td>90%</td>
<td>86%</td>
</tr>
<tr>
<td>0.3</td>
<td>95%</td>
<td>47%</td>
<td>66%</td>
</tr>
<tr>
<td>0.4</td>
<td>100%</td>
<td>34%</td>
<td>60%</td>
</tr>
<tr>
<td>0.5</td>
<td>100%</td>
<td>14%</td>
<td>48%</td>
</tr>
<tr>
<td>0.6</td>
<td>100%</td>
<td>7%</td>
<td>44%</td>
</tr>
<tr>
<td>0.7</td>
<td>100%</td>
<td>0%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Figure 5.6 shows the ZCR Max accuracy, sensitivity and specificity curve for different thresholds
Figure 5.6 ZCR Max accuracy, sensitivity and specificity curve

The following figure shows the ZCR Max values for all signals in the datasets.
After analyzing these data, we chose a condition which is:

For the normal voices:

\[ \text{ZCR Max} < 0.23 \] (5.2)

This condition gives the following percentages:

Accuracy: 86%

Sensitivity: 80%

Specificity: 90%
Those percentages based on testing 50 different voice signals, 30 diseased signals and 20 normal ones.

Table 5.7 shows accuracy, sensitivity and specificity for different thresholds in the ZCR, Mean:

Table 5.7 Accuracy, Sensitivity and Specificity for different thresholds in the ZCR Mean

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>0.03</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>0.06</td>
<td>0%</td>
<td>97%</td>
<td>58%</td>
</tr>
<tr>
<td>0.09</td>
<td>0%</td>
<td>94%</td>
<td>56%</td>
</tr>
<tr>
<td>0.1</td>
<td>25%</td>
<td>90%</td>
<td>64%</td>
</tr>
<tr>
<td>0.11</td>
<td>80%</td>
<td>83%</td>
<td>82%</td>
</tr>
<tr>
<td>0.12</td>
<td>80%</td>
<td>66%</td>
<td>72%</td>
</tr>
<tr>
<td>0.13</td>
<td>100%</td>
<td>60%</td>
<td>76%</td>
</tr>
<tr>
<td>0.18</td>
<td>100%</td>
<td>40%</td>
<td>64%</td>
</tr>
<tr>
<td>0.23</td>
<td>100%</td>
<td>30%</td>
<td>58%</td>
</tr>
<tr>
<td>0.33</td>
<td>100%</td>
<td>7%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Figure 5.8 shows the ZCR Mean accuracy, sensitivity and specificity curve for different thresholds.
The following figure shows the ZCR Mean values for all signals in the datasets.

Figure 5.8 ZCR Mean accuracy, sensitivity and specificity curve
After analyzing these data, we found that the most of normal voice signals ranges from 0.09 to 0.13 so choosing a single threshold will affect the Accuracy, Sensitivity and Specificity percentages so we chose a condition which is:

For the normal voices:

\[0.09 < \text{ZCR Mean} < 0.13\]  

(5.3)

This condition gives the following percentages:

- **Accuracy**: 80%
- **Sensitivity**: 100%
- **Specificity**: 67%
Those percentages based on testing 50 different voice signals, 30 diseased signals and 20 normal ones.

### 5.2.2.3 Linear Predictive Coding (LPC)

We repeated the same systematic technique on all signals which is choosing 400 ms segment length in the middle of the vowel “ah” and from the frequency domain feature extraction panel in the developed software we chose “LPC”.

After analyzing all the signals -Normal and disordered- we tested different thresholds to find the best threshold to classify the signals into normal and disordered ones. In this process we are also observing the accuracy, sensitivity and specificity for each threshold to get the threshold which achieves the best accuracy and in the same time preserves a high sensitivity and specificity.

Table 5.8 shows accuracy, sensitivity and specificity for different thresholds in the LPC:
Table 5.8 Accuracy, Sensitivity and Specificity for different thresholds in the LPC

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>50</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>100</td>
<td>0%</td>
<td>90%</td>
<td>45%</td>
</tr>
<tr>
<td>110</td>
<td>0%</td>
<td>90%</td>
<td>45%</td>
</tr>
<tr>
<td>120</td>
<td>25%</td>
<td>90%</td>
<td>64%</td>
</tr>
<tr>
<td>130</td>
<td>45%</td>
<td>84%</td>
<td>68%</td>
</tr>
<tr>
<td>150</td>
<td>60%</td>
<td>56%</td>
<td>58%</td>
</tr>
<tr>
<td>170</td>
<td>80%</td>
<td>37%</td>
<td>54%</td>
</tr>
<tr>
<td>190</td>
<td>90%</td>
<td>30%</td>
<td>54%</td>
</tr>
<tr>
<td>220</td>
<td>100%</td>
<td>30%</td>
<td>58%</td>
</tr>
<tr>
<td>250</td>
<td>100%</td>
<td>27%</td>
<td>56%</td>
</tr>
<tr>
<td>350</td>
<td>100%</td>
<td>20%</td>
<td>52%</td>
</tr>
<tr>
<td>450</td>
<td>100%</td>
<td>10%</td>
<td>46%</td>
</tr>
<tr>
<td>550</td>
<td>100%</td>
<td>3%</td>
<td>42%</td>
</tr>
<tr>
<td>600</td>
<td>100%</td>
<td>0%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Figure 5.10 shows the LPC accuracy, sensitivity and specificity curve for different thresholds
The following figure shows the LPC values for all signals in the datasets.

Figure 5.10 LPC accuracy, sensitivity and specificity curve
After analyzing these data, we found that the most of normal voice signals ranges from 110 to 130 and from 167 to 220 so choosing a single threshold will affect the Accuracy, Sensitivity and Specificity percentages so we chose a condition which is:

**For the normal voices:**

\[ 110 < \text{LPC} < 130 \]
\[ 167 < \text{LPC} < 220 \]  \hspace{1cm} (5.4)
This condition gives the following percentages:

**Accuracy: 82 %**

**Sensitivity: 75 %**

**Specificity: 87 %**

Those percentages based on testing 50 different voice signals, 30 diseased signals and 20 normal ones.

**5.2.2.4 Mel Frequency Cepstral Coefficients (MFCC)**

We repeated the same systematic technique on all signals which is choosing 400 ms segment length in the middle of the vowel “ah” and from the frequency domain feature extraction panel in the developed software we chose “LPC”.

After analyzing all the signals -Normal and disordered- we tested different thresholds to find the best threshold to classify the signals into normal and disordered ones. In this process we are also observing the accuracy, sensitivity and specificity for each threshold to get the threshold which achieves the best accuracy and in the same time preserves a high sensitivity and specificity.
Table 5.9 shows accuracy, sensitivity and specificity for different thresholds in the MFCC:

Table 5.9 Accuracy, Sensitivity and Specificity for different thresholds in the MFCC

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>120</td>
<td>0%</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>130</td>
<td>0%</td>
<td>94%</td>
<td>56%</td>
</tr>
<tr>
<td>140</td>
<td>15%</td>
<td>64%</td>
<td>44%</td>
</tr>
<tr>
<td>150</td>
<td>60%</td>
<td>50%</td>
<td>54%</td>
</tr>
<tr>
<td>170</td>
<td>100%</td>
<td>24%</td>
<td>54%</td>
</tr>
<tr>
<td>190</td>
<td>100%</td>
<td>7%</td>
<td>44%</td>
</tr>
<tr>
<td>210</td>
<td>100%</td>
<td>3%</td>
<td>42%</td>
</tr>
</tbody>
</table>

Figure 5.12 shows the MFCC accuracy, sensitivity and specificity curve for different thresholds
The following figure shows the MFCC values for all signals in the datasets.
After analyzing these data, we found that the most of normal voice signals ranges from 130 to 150 so choosing a single threshold will affect the Accuracy, Sensitivity and Specificity percentages so we chose a condition which is:

**For the normal voices:**

\[130 < \text{MFCC} < 150\] (5.5)

This condition gives the following percentages:

- **Accuracy:** 58%
- **Sensitivity:** 60%
- **Specificity:** 57%
Those percentages based on testing 50 different voice signals, 30 diseased signals and 20 normal ones.

5.3 Classification

After analyzing the previous data, we thought about a method which combines all features and increase the Accuracy, Sensitivity and Specificity percentages. We used the rule “A simple classification method with strong features is better than complicated classifier with weak features”. The method based on applying simple logic conditions “OR” and “AND” on the selected thresholds & ranges for the tested features.

We tested many combinations between the selected features and with applying the previous conditions (From 5.1 to 5.5).

We selected the following combination as they give the best results in the Accuracy, Sensitivity and Specificity:

1- ZCRMAXRANGE AND ZCRMEANRANGE

2- ZCRMEANRANGE AND EMEANRANGE

3- ZCRMAXRANGE AND LPCRANGE

4- ZCRMAXRANGE AND MFCCRANGE

5- EMEANRANGE AND MFCCRANGE

The following table shows Accuracy, Sensitivity and Specificity for the single features and the combined ones with the selected conditions:
Table 5.10 Accuracy, Sensitivity and Specificity for the single features and the combined ones

<table>
<thead>
<tr>
<th>Feature</th>
<th>Condition</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy Mean</td>
<td>Normal &gt; 0.07</td>
<td>85%</td>
<td>70%</td>
<td>76%</td>
</tr>
<tr>
<td>ZCR Max.</td>
<td>Normal &lt; 0.23</td>
<td>80%</td>
<td>90%</td>
<td>86%</td>
</tr>
<tr>
<td>ZCR Mean</td>
<td>0.09 &lt; normal &lt; 0.3</td>
<td>100%</td>
<td>67%</td>
<td>80%</td>
</tr>
<tr>
<td>LPC</td>
<td>110 &lt; Normal &lt; 130</td>
<td>75%</td>
<td>87%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>167 &lt; Normal &lt; 220</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>130 &lt; Normal &lt; 150</td>
<td>60%</td>
<td>57%</td>
<td>58%</td>
</tr>
<tr>
<td>ZCR Mean AND ZCR Max</td>
<td></td>
<td>80%</td>
<td>97%</td>
<td>90%</td>
</tr>
<tr>
<td>ZCR Max AND LPC</td>
<td></td>
<td>65%</td>
<td>93%</td>
<td>82%</td>
</tr>
<tr>
<td>ZCR Max AND MFCC</td>
<td></td>
<td>50%</td>
<td>97%</td>
<td>78%</td>
</tr>
<tr>
<td>ZCR Mean AND Energy Mean</td>
<td></td>
<td>85%</td>
<td>87%</td>
<td>86%</td>
</tr>
<tr>
<td>Energy Mean AND MFCC</td>
<td></td>
<td>55%</td>
<td>90%</td>
<td>76%</td>
</tr>
</tbody>
</table>
Then we developed the following algorithm:

Algorithm used in classification

- For the selected segment, get the following values:
  ZCR MAX, ZCR MEAN, Short Time Energy MEAN, LPC and MFCC

- Compare the previous values with the following values assigned to the classifier

  \[ MFCCRANGE = (130 < MFCC < 150) \]

  \[ LPCRANGE = (110 < LPC < 130) \text{ OR } (167 < LPC < 220) \]

  \[ ZCRMAXRANGE = (ZCRMAX < 0.23) \]

  \[ ZCRMEANRANGE = ((ZCRMEAN > 0.09) \text{ AND } (ZCRMEAN < 0.13)) \]

  \[ EMEANRANGE = (EMEAN > 0.07) \]

- IF (ZCRMAXRANGE AND ZCRMEANRANGE) OR
  (ZCRMEANRANGE AND EMEANRANGE) OR (ZCRMAXRANGE AND LPCRANGE) OR (ZCRMAXRANGE AND MFCCRANGE) OR
  (EMEANRANGE AND MFCCRANGE)

  Normal

  else

  Diseased

  In this algorithm the program checks all conditions given and be sure that at least two features are true / satisfied to decide that the voice signal is
normal / healthy. But the output will be a general one (Normal or diseased). As the system has percentages of error (In terms of accuracy, sensitivity and specificity). We modify the previous algorithm to a more accurate one as the following:

\[
\text{if (ZCRMAXRANGE\&ZCRMEANRANGE)}
\]

\[
\text{Then Normal by 90 %}
\]

\[
\text{elseif (ZCRMEANRANGE\&EMEANRANGE)}
\]

\[
\text{Then Normal by 86 %}
\]

\[
\text{elseif(ZCRMAXRANGE\&LPCRANGE)}
\]

\[
\text{Then Normal by 82 %}
\]

\[
\text{elseif(ZCRMAXRANGE\&MFCCRANGE)}
\]

\[
\text{Then Normal by 78 %}
\]

\[
\text{elseif(EMEANRANGE\&MFCCRANGE)}
\]

\[
\text{Then Normal by 76 %}
\]

\[
\text{else}
\]

\[
\text{Diseased}
\]

In this algorithm the program checks the first condition which has the highest accuracy 90 % (ZCRMAXRANGE\&ZCRMEANRANGE), if it is true then the output will indicate that the voice signal is normal by 90 % regardless the output of other conditions whether if they are true or false.
If the first condition is false then the program checks the second condition which has the second highest accuracy 86% (ZCRMEANRANGE&EMEANRANGE) if it is true then the output will indicate that the voice signal is normal by 86% regardless the output of other conditions (Excluding the first one which is false already as indicated) whether if they are true or false.

Then the process continues in the same manner until checking the last condition, if it is false then the output will be a diseased voice signal. This algorithm employs the percentages of accuracies calculated for each condition in the classification process.

The experimental results show that classification based on Maximum Zero Crossing Rate in the signal attained 86% Accuracy, compared to 82%, 80%, 76%, 58% for classification based on LPC, Mean of Zero Crossings Rate, Short time Energy Mean and MFCC respectively.

However that classification based on Maximum Zero Crossing Rate in the signal gives the highest accuracy all over the other features; we cannot depend on this feature only to decide if the voice is normal or disordered. The same concept is also applied on all other features also which means that we cannot depend on one feature alone.

Moreover, combining more than one feature gives better results in Accuracy, Sensitivity and Specificity as we have stated before in results section. We can see that combining the maximum zero crossing rate in the signal with the mean of zero crossing rate in logic “AND” condition gives 90% accuracy. Also combining: Mean of zero crossing rate with the mean of energy in a logic “AND” condition, Maximum zero crossing rate with LPC in a logic “AND” condition, Maximum Zero crossing rate with MFCC
in a logic “AND” condition and Mean of energy and MFCC in a logic “AND” condition gives accuracies 86 %, 82 %, 78% and 76 % respectively.

According to Shama et al. [25], when using 21 energy bands only with K-NN classifier the system’s accuracy was 82%.

When using MFCC with GMM classifier the system accuracy was 83%, Bocklet et al. [26].

6.1 Conclusions

This research proposed a computer based system for pathological/normal voice classification of vocal folds disorders. It is an approach to recognize the presence of pathology from voice records by means of two time domain features: Zero Crossing Rate (ZCR) and Short time Energy and two Frequency domain features: Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC).

It is important to take care also of Sensitivity and Specificity percentages. In some conditions taken, the accuracy results for other thresholds and ranges were higher than this accuracy but affect the sensitivity or specificity percentages.

After testing many conditions and combinations we chose those ones which preserve a minimum 50% for sensitivity and Specificity along with the highest possible accuracy.

For the proposed system we cared about the specificity more than sensitivity. This is because the fact that detecting a diseased signal is the most important which means that it is more tolerable to false indicate the presence of a disease for a normal / healthy voice than indicating that a diseased voice is normal / healthy one. So we can say that the proposed system has a high accuracy, high specificity and a satisfactory sensitivity.
One advantage of the proposed system is the combination between time domain features and frequency domain features. To the best of our knowledge, most of previous work in this field use only one domain features in classification process. Also the time domain feature extraction panel and the frequency domain one in the developed software allow researchers to apply the feature extraction process on the proposed features (ZCR, short time energy, LPC, MFCC) on larger datasets which may lead to a higher accuracy results.

With respect to the developed algorithm, it gives very good results with a simple classification technique which based on choosing threshold and ranges to differentiate between the normal and the pathological ones.

We observed that in case of the used frequency domain features (LPC and MFCC), the normal voice signals have more than one range. This may be used in classification of normal voice signals according to age and gender.

Finally this work can be used as the following:

- A simple and efficient tool which help Doctors and Physicians in hospitals and clinics in diagnosis of voice disorders
- It can be applied in the field of preventive medicine in order to achieve early detection of laryngeal pathologies.
- It is a good first step in the diagnosis of complicated cases or the serious diseases which need a lot of further analysis and larynx’s imaging as an example.
- A simple home application to be used by users at home to check their voice health status.
6.2 Challenges and Drawbacks

One of the greatest challenges in this work was to collect a database with a reliable number of audio signals which have been recorded in a calm environment and with fixed conditions so that it can be used in studying the different features and in the classification process.

We spent a lot of time the process of collecting the signals as a lot of physicians and hospitals’ stuffs are not willing to provide us with audio signals for their patients.

The lack of time and experience was another great challenge in our work. Audio signal processing is a new field for us to deal with, We spent also a period of time before starting the practical work in studying an online course “ Audio signal processing in Matlab” to be able to begin the process of experimental work.

Preserving high percentages for accuracy, sensitivity and specificity was also a challenge for us in the feature extraction process. On the one hand, a high accuracy percentage may be obtained but on the other one, one of the sensitivity and specificity is dropped. So it takes a lot of effort and time to reach results with high accuracy, high specificity and satisfactory sensitivity.

One of the drawbacks of the proposed system is that the system has not a relatively high sensitivity. As discussed before, we cared more about the accuracy and specificity over the sensitivity.

Another drawback is that the segmentation process in the software is done manually. It is clear that automatic segmentation which means that
the software will be able to capture automatically a segment of 400 ms in
the middle of the vowel “Ah” is more efficient.

The fact that the developed software is a graphic user interface built
in Matlab leads to a difficulty for the users to be able to use the software
easily.

### 6.3 Future Work

The future work will be to identify the type of pathologies. For this
purpose, the system should pass through two main steps: the first one deal
with the detection of voice disorder; once the presence is confirmed, the
second step will be voice disorder type identification and recognition.

Since the research work presented combines 2 time domain features
and 2 frequency domain features, it opens up the way to extend this work
for classification tasks between different disorders, perceptual vocal
qualities or the categorization of the speech registers into different degrees
of impairment.

The future works may include the application and analysis of this
approach in larger databases.

Also it may be possible to try to build a complete multi class
classification system so that detection of different type of pathological
speech will be possible. For this propose, we suppose that further research
for more sophisticated feature extraction phase. In such approach our work
will be to increase the percentage of accuracy, specificity and sensitivity of
the system. Also making the segmentation process to be done automatically by the software is a task to be done in the next stage.

Due to the fact that Matlab is not common software for ordinary users, a hardware device or an android application will be our work in the next stage in order to be able to make a tool which can be used for non experienced users.
References


