

ISOLATED WORD SPEECH RECOGNITION SYSTEM FOR CHILDREN WITH  
DOWN SYNDROME

by

Janah Salama Emeeshat

Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in the

Electrical Engineering

Program

YOUNGSTOWN STATE UNIVERSITY

August, 2017

Isolated Word Speech Recognition System for children with Down syndrome

Janah Salama Emeeshat

I hereby release this thesis to the public. I understand that this thesis will be made available from the OhioLINK ETD Center and the Maag Library Circulation Desk for public access. I also authorize the University or other individuals to make copies of this thesis as needed for scholarly research.

Signature:

---

Janah S. Emeeshat, Student

Date

Approvals:

---

Dr. Frank X. Li, Thesis Advisor

Date

---

Dr. Philip Munro, Committee Member

Date

---

Dr. Faramarz Mossayebi, Committee Member

Date

---

Dr. Salvatore A. Sanders, Dean of Graduate Studies

Date

## **Abstract**

Automatic speech recognition by machine is one of the most effective methods for man-machine communication. Because speech waveform is nonlinear and time-variant, speech recognition requires significant amount of intelligence and fault tolerance in the pattern recognition algorithms. The objective of this work was to develop an isolated word speech recognition system for children with Down syndrome to communicate with others, almost normally. These children are delayed in the use of meaningful speech and slower to obtain a fruitful vocabulary due to their large tongue and other factors. In this thesis, single words were collected from a child with Down syndrome of age 10. Additionally, the same words were collected from a child with typical development in the same age to compare their speech features. Children voices were recorded by a mobile phone first then recorded at a sample rate of 48 kHz using a unidirectional microphone at 24-bits per sample and one recording channel. The model proposed was based on Fast Fourier Transform (FFT) as a feature extraction technique. FFT transforms the sampled data from time domain to frequency domain to investigate features of spoken words and features of different children.

## **Dedication**

*To the one and only  
who was the motivation behind this thesis; my baby brother Ahmad  
May your soul rest in peace.*

## **Acknowledgements**

First and foremost, I would like to thank God for his generous blessings and for giving me the strength and ability to complete this thesis. Also, I would like to express my sincere gratitude to my thesis advisor, Dr. Frank X. Li, for his suggestions, guidance, support, advice, and patience. It has indeed been a privilege to work with him.

I am very thankful for my parents for their love, the chance to pursue my master's degree in The United States of America, and their encouragements and confidence during all my life, as it has made me who I am. I also appreciate the support offered by my brother Ayed.

I would like to express my deepest appreciation to my uncle, Dr. Salem Alemaishat, for his support and valuable advice during my studies, and in every aspect of my daily life. I also would like to express my affection to my closest friends for being always next to me. Thanks to Tahani Sultan and Omneya Ali, and their families, whom I have shared the best moments with and have been supported by during the difficult ones. Last but not least, I am very thankful for all relatives and friends for their love and prayers.

## Table of Contents

Abstract .....	iii
Dedication .....	iv
Acknowledgements .....	v
List of Figures .....	viii
List of Tables .....	viii
Abbreviations .....	ix
Chapter 1 Introduction .....	1
1.1 Speech Recognition .....	1
1.2 Motivation and Basics of the Research .....	2
1.3 Organization of the Thesis .....	3
Chapter 2 Literature Review .....	4
2.1 Background .....	4
2.2 Approaches to Speech Recognition .....	4
2.2.1 Acoustic Phonetic Approach .....	4
2.2.2 Pattern Recognition Approach .....	5
2.2.3 Artificial Intelligence Approach (Knowledge Based Approach) .....	6
2.3 Introduction to Speech Sounds .....	7
2.3.1 Speech Production .....	7
2.3.2 Speech Perception .....	8
2.3.3 Speech Features .....	10
Chapter 3 Speech Recognitions .....	12
3.1.1 Types of Speech Recognition .....	12
3.1.2 Relevant Issues of ASR design .....	13
3.1.3 Developing ASR .....	14
3.2 Speech Recognition of Children with Down Syndrome .....	14
3.2.1 The Phonology of Single Words .....	16
Chapter 4 Voice Recolonization System Modeling and Simulations .....	17
4.1 Data Collection .....	17
4.2 Software .....	17
4.3 System Datasets and Parameters .....	19
4.4 System Block Diagram .....	19
4.5 Voice Recording .....	21
4.6 Pre-processing Stage .....	22

4.6.1 DC Offset Removal.....	22
4.6.2 Amplitude Normalization .....	23
4.7 Endpoint Detection (Word Boundary Detection) .....	26
4.7.1 Time-domain waveforms .....	33
4.8 Feature Extraction.....	35
4.8.1 Fast Fourier Transform (FFT).....	36
4.8.2 Formants .....	45
Chapter 5 Conclusion and Future Research.....	48
5.1 Summary and Closing Remarks .....	48
5.2 Future Work and Recommendations .....	49
Bibliography .....	50

## List of Figures

Figure 1 Speech apparatus cross-section .....	7
Figure 2 Human ear cross-section.....	9
Figure 3 Voice recognition process .....	20
Figure 4 Voice recognition flow diagram.....	21
Figure 5 Shifted FFT signal of child C saying the word “Apple” .....	24
Figure 6 Single-sided FFT signal of child C saying the word “Apple” .....	25
Figure 7 Double-sided scaled and filtered signal.....	25
Figure 8 Single-sided scaled and filtered signal .....	26
Figure 9 Endpoint detection block diagram.....	27
Figure 10 The waveform of the word “Apple” before applying endpoint detection .....	31
Figure 11 The waveform of the word “Apple” after applying endpoint detection.....	32
Figure 12 Voice recording of “Cat” for child c with Down syndrome.....	34
Figure 13 Time-domain plot of a child with Down syndrome of the word “Apple” .....	34
Figure 14 Extracting a frame from time-domain signal.....	41
Figure 15 Extracting 512 points of a frame .....	41
Figure 16 Formants Identification .....	45
Figure 17 FFT plots of two children pronouncing the same word .....	47

## List of Tables

<b>Table 1 FFT MegaCore function Parameters .....</b>	<b>49</b>
---	-----------



## Abbreviations

ADC	Analog to Digital Converter
FFT	Fast Fourier Transform
DFT	Discrete Fourier Transform
ASR	Automatic Speech Recognition
CTFT	Continuous-Time Fourier Transform
DAC	Digital to Analog Converter
HDL	Hardware Description Language
VHSIC	Very High Speed Integrated Circuit
VHDL	VHSIC Hardware Description Language
FPGA	Field-Programmable Gate Array
MFCC	Mel-Frequency Cepstral Coefficient

# Chapter 1 Introduction

## 1.1 Speech Recognition

Speech is the essential communication intermediate between people. This communication process has a convoluted structure consisting not only of transmission of voice. Body language, language, topic, and the ability of the listener devote to this process as well. As the dictum says, what you can tell is restricted to what the auditor can understand. In this respect, the behavior of a speech recognizer system mainly depends on how and for which task you design it.

In the last six decades, efforts have been made to automate the recognition of human speech. The term “Speech Recognition” is one that comprises many distinct ways to the problem of recognizing human speech. It differs from isolated word recognition to continuous speech recognition, from speaker-dependent recognition to speaker-independent recognition, and from a small word to a large word. The speech recognition issue, as developed in recent years, is a highly reckoning exhaustive problem, as it needs fast processors, and a huge amount of memory.

Automatic Speech Recognition system (ASR), which was considered to be a part of science fiction for many years, is now a crucial branch of information and communication technology. ASR is used to transform spoken words into text. It has very critical applications such as dictation, security control, and foreign language translation. Moreover, ASR can help special needs people to interact more easily with society.

## **1.2 Motivation and Basics of the Research**

This thesis focuses on isolated word speech recognition for children with Down syndrome. The motivation behind this idea is that my brother, who passed away, was a child with Down syndrome. He faced many difficulties in communication with people because his speech was not clear enough to be understood due to his large tongue and smaller dimensions of the speech apparatus than that of a normal child. His situation inspired me to create a system that could help other children with Down syndrome. Moreover, I also wanted to honor my baby brother. With so much usefulness that speech recognition could bring to our lives, there are persuasive reasons for researching and improving speech recognition technology. However, achieving recognition is a completely difficult task. The complexity is created because of the number of speakers, the irregularity of utterances, the complexity of the language, and the environmental conditions. This thesis deals with isolated word recognition through the use of Digital Signal Processing (DSP) and the use of MATLAB. Moreover, VHSIC Hardware Description Language (VHDL) is used to extract the features of the speech signals.

The purpose of this research was to develop a system for recognizing speech of children with Down syndrome. Children's speeches have been recorded via MATLAB at a sampling frequency of 48 kHz for 3 seconds as an audio signal. These signals have been processed by MATLAB to get the formants and their frequencies. In addition, a code was written for features extractions from speech samples to try to prove if these frequencies are in the audible range (i.e. the range of frequencies that can be heard by humans).

### **1.3 Organization of the Thesis**

This thesis gives a comprehensive description about the developed system, and is organized as follows:

- Chapter 1 starts with an introduction to the research and the idea behind it.
- The second chapter provides insight about related works in the field of Automatic Speech Recognition and the main approaches used.
- Chapter 3 gives details about speech recognition systems and children with Down syndrome.
- Chapter 4 is mostly dedicated to system development, and explains in detail how it operates.
- In the last chapter of this thesis, concluding remarks are stated. Future work, which may follow this study, is also presented.

# Chapter 2 Literature Review

## 2.1 Background

Speech Recognition is also addressed as Automatic Speech Recognition (ASR), or computer speech recognition, which is the procedure of adapting a speech signal to a flow of single words through methods of an algorithm actualized as a computer program. This chapter discusses the researchers' efforts in this exciting and challenging field, the dominant topics researched today, and advances accomplished in the past 60 years of research.

## 2.2 Approaches to Speech Recognition

Essentially, three approaches exist to speech recognition. They are [1]:

- Acoustic Phonetic Approach
- Pattern Recognition Approach
- Artificial Intelligence Approach

### 2.2.1 Acoustic Phonetic Approach

The earliest approaches to speech recognition took the basis of finding speech sounds and providing suitable labels to them. Hemdal and Hughes took that basis and hypothesized that there exist finite unique phonetic units (phonemes) in spoken language, and that these units are widely characterized by a set of acoustic properties that are alternating with respect to time of a speech signal [2]. According to this approach, the message carrying components of speech are to be derived explicitly with the perseverance of appropriate binary acoustic properties such as voiced-unvoiced

classification, nasality, frication, and continuous features such as formant locations, and ratio of high and low frequencies. These properties will be discussed in chapter 3.

The acoustic phonetic approach has not been broadly used in most commercial applications because it is difficult to get a reliable phoneme lattice [3]. This approach is materialized in a specific series: spectral analysis, features detection, segmentation and labeling, and recognizing valid words. In the validation process, linguistic constraints on a specific task (i.e., the vocabulary, the syntax, and other semantic rules) are implemented to access the lexicon for word decoding.

### **2.2.2 Pattern Recognition Approach**

The pattern-matching approach was first asserted by Itakurain 1975, which got enormous support from other researchers [1]. This approach has become the ruling method for speech recognition in the last six decades [3]. It includes two essential steps: pattern training and pattern comparison. The fundamental feature of this approach is that it utilizes a well formulated mathematical framework and there-after authenticates coherent speech pattern representations, for decent pattern comparison, from a set of labeled training samples by a formal training algorithm.

A speech pattern representation can be in the shape of a speech template or a statistical model [4]. Moreover, it can be applied to a sound (smaller than a word), a word, or a phrase. The second step of the approach is the pattern comparison, a comparison made between the speeches that are recognizable with each possible template learned in the training stage to decide the identity of the unknown by using the matching algorithm.

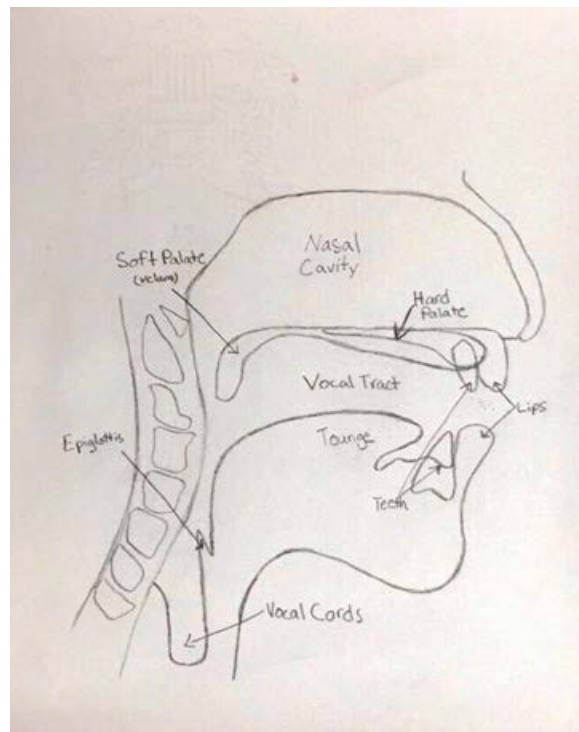
### **2.2.3 Artificial Intelligence Approach (Knowledge Based Approach)**

This approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. It sheds light on how to automate the speech recognition process according to the way a person implements intelligence in visualizing, analyzing, and characterizing speech, depending on a group of measured acoustic features. Both acoustic phonetic and template based approach failed on their own to examine appreciable insight into human speech processing [4]. In the classical knowledge based approach, the production rules are constructed curiously from empirical linguistic knowledge or from the conclusions from the speech spectrogram. Knowledge aids the algorithm to execute better, and plays an important role in choosing a suitable input representation, the definition of units of speech, or the design of the algorithm itself. Samouelian hypothesized a data driven methodology for CSR in which the knowledge about the structure and characteristics of the speech wave are collected notably from the database by using inductive inference [5]. This approach was found to have benefits of solving the problem of inter and intra speaker speech variability, and the proficiency to create decision trees. Tripathy states a knowledge based approach using fuzzy inference for the categorization of spoken English vowels that gives better results over the standard Mel-Frequency Cepstral Coefficient (MFCC) feature extraction [6].

## 2.3 Introduction to Speech Sounds

### 2.3.1 Speech Production

Speech sound is assembled by a group of well-controlled movements of various speech apparatus. Figure 1 shows a schematic cross-section through the vocal tract of the apparatus.



**Figure 1 Speech apparatus cross-section**

The vocal tract is an essential acoustic tube, which is the part of the mouth cavity enclosed by the vocal cords and the lips. As air is expelled from the lungs, the vocal cords are anxious and then quivered by the airflow. The fundamental frequency is called the frequency of oscillation, and it depends on many properties of the vocal cords, such as length, mass and tension.



It can vary:

From 50 to 200 Hz for a male voice,

From 150 to 450 Hz for a female voice,

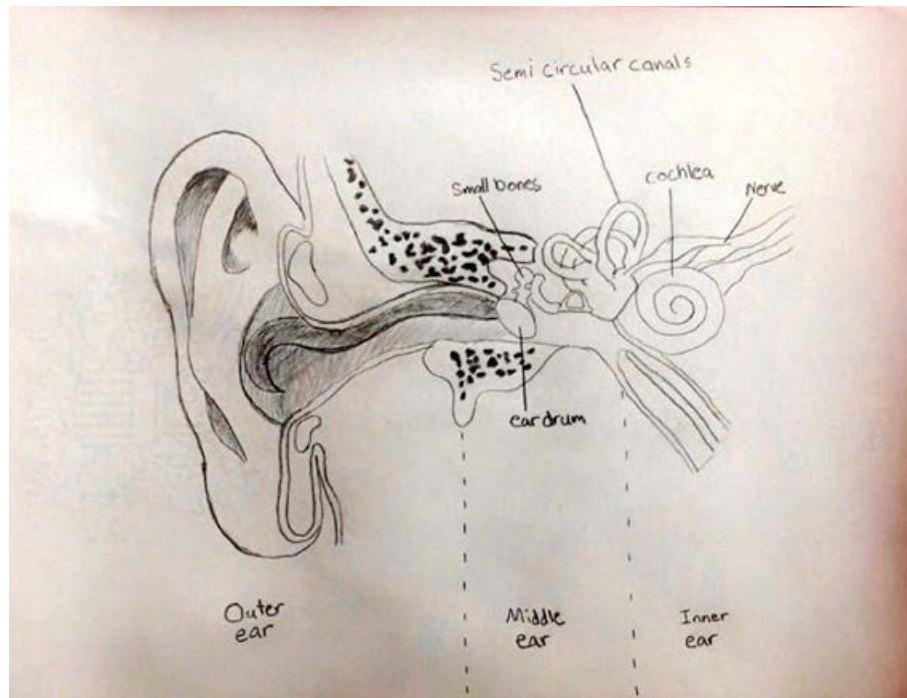
From 200 to 600 Hz for a child's voice

Concurrently with this process, the appearance of the vocal tube is changed by distinct positions of the velum, tongue, jaw, and lips [7]. The vocal tract, as an acoustic resonator, decides fickle resonant frequencies by modifying the shape and the size of the vocal tract. The resonant frequency is called the formant frequency, or simply formant. The nasal tract is a supplementary acoustic tube that can be acoustically collaborated with the vocal tract to form nasal sounds.

Miscellaneous speech sounds are produced not only by modifying the size and the shape of the vocal tract, but also by the type of excitation. Alongside the airflow from the lung, the excitation could occur by some other sources such as fricative excitation, plosive excitation and whispered excitation, the excitation sources are explained in chapter 3 [8].

### **2.3.2 Speech Perception**

As the vocal system can produce speech sounds, the auditory system can detect the change in air pressure of audible sounds [1]. Figure 2 illustrates a cross-section diagram of the human ear, the ear involves three parts: the outer ear, the middle ear, and the inner ear.



**Figure 2 Human ear cross-section**

The outer, external ear gathers sound waves and focuses them on the eardrum. The middle ear serves as a mechanical amplifier that converts the vibrations of the eardrum to oscillations of the fluid filled inner ear. The inner ear then transforms the mechanical vibrations into electrical signals that travel to the auditory nerve. The auditory nerve carries impulses from the cochlea (the hearing part of the inner ear) to a relay station in the mid-brain, the cochlear nucleus. These nerve impulses are then carried on to other brain pathways that end in the auditory cortex (the hearing part) of the brain [9].

The human ear can interact with insignificant pressure variations in the air if they are within audible frequency range, roughly 15 Hz to 20 kHz. The ear also acts like a filter and will prefer some frequencies over others. Human ears are most sensitive to frequencies of the range from 1000 Hz to 5000 Hz. This range is related to the resonance of the auditory canal (ear canal).

### 2.3.3 Speech Features

Generally, the automatic speech recognition system can be branched into two main phases: feature extraction phase, and pattern recognition phase. Feature extraction phase aims to search for the speech characteristics that are compact and effective to represent a speech signal, and save them for the second phase, pattern recognition. Any natural language, including English, is based on a group of noticeable and mutually exclusive fundamental units, which are called phonemes. All the phonemes are related to different articulator gestures of a language. There are several ways to classify speech sounds [7, 10]. Based on the type of excitation source of phonemes, speech sounds can be categorized into the following categories:

- **Voiced sounds** that happen when air pressure forces the vocal cords open and makes them vibrate. The vibrating cords harmonize the air stream from the lungs at a rate that could be from 60 times per second for some males to 500 times per second for children. The maximum amplitude of a voiced sound is way higher than that of the unvoiced sound.
- **Nasal sounds** are also voiced. Nevertheless, the nasal cavity is included with the vocal cavity among the utterance. A portion of the airflow is deflected into the nasal tract by opening the velum.
- **Fricatives** are created by exciting the vocal tract with a violent flow generated by airflow through a tight constriction.
- **Voiced fricatives** arise when the vocal tract is excited synchronously by both rough flow and vocal vibration.
- **Plosives** are produced by exciting the vocal tract with a speedy release of pressure by the constrictions of teeth or lips.

- **Affricative sounds** are generated by increasingly releasing an entirely closed and pressurized vocal tract.
- **Whispered sounds** are excited by airflow hurrying through a small triangular opening between the arytenoid cartilages at the end of the nearly closed folds.

## Chapter 3 Speech Recognitions

### 3.1 Speech Recognition by Machine

Speech recognition systems consider that the speech signal is a realization of some message encoded as a sequence of one or more symbols. The essential goal is to “decode” this message and then convert it either into writing or into commands to be processed.

#### 3.1.1 Types of Speech Recognition

Speech recognition systems can be divided into distinct classes by describing what types of utterances they can recognize. These classes are identified as the following:

- **Connected Words:** Connected word systems (or more precisely ‘connected utterances’), allow apportioned utterances to be ‘run-together’ with a minimal pause between them.
- **Continuous Speech:** Continuous speech recognizers let users speak almost naturally, while the computer detects the content. Basically, it is a computer dictation.
- **Spontaneous Speech:** This can be thought of as speech that is natural sounding and not rehearsed. An ASR system for such a speech deals with a variety of natural speech features i.e. words being run together,” ums”,” ahs”, and even slight stutters.
- **Isolated Words:** Isolated word recognizers usually need each utterance to have silence on both sides of the sample window. The user speaks individual words or phrases and the recognizer accepts single words or a single utterance at a time. These systems have “Listen/Not-Listen” states, where they require the speaker to wait (pause) between utterances. The discrete utterance is dealt with in two implicit

assumptions. The first assumption is that the speech consists of a signal that is going to be recognized as a complete entity with no explicit knowledge for the phonetic content of the word/phrase. The second assumption is that each spoken word/phrase has an apparently defined beginning and ending point. In this thesis, an isolated word speech recognition system was developed to recognize two spoken words (Apple and Cat) by a child with Down syndrome.

### 3.1.2 Relevant Issues of ASR design

Since speech waveform is nonlinear and dynamic, speech recognition is an intrinsically complex task. There are different main variabilities of speech signals including within-speaker variability, across-speaker variability, transmission variability, and the environmental conditions under which the speaker is talking.

- **Within-speaker** variability is caused by incompatible pronunciation, speaking speed, and distinct emotions when the words are spoken by same speaker.
- **Across-speaker** variability is due to regional accents and foreign languages.
- **Transducer and transmission** variability is because the words are spoken over several microphone/handsets and the speech signal can be transmitted by all the means of communication systems.
- **Language complexity** causes speech recognition to be an extremely difficult job. So far, the duty of speech recognizers is simplified by reducing the number of utterances by the imposition of semantic restraints. However, developers should note the multi-disciplinary natures of speech signals because speech is a natural activity of human beings.
- **Environmental condition** is also a fundamental concern of speech recognizers. It includes type of noise, signal/noise ratio, and working conditions.

### **3.1.3 Developing ASR**

Steps to develop and improve a general automatic speech recognition system have been observed as [11]:

- i. First, the speech is acquired through a unidirectional and noiseless microphone.
- ii. Signal parameterization using an appropriate feature extraction technique.
- iii. Acoustic analysis: The training waveforms are converted into a series of coefficient vectors.
- iv. Definition of the models: A prototype of a model is defined for each element of the task vocabulary.
- v. Training of the models: Each model is initialized and trained with the data.
- vi. Definition of the task: Here the grammar for the speech recognizer is defined.
- vii. Recognition: The unknown input speech signal is recognized at this level.

There have been some changes at one or more steps applied due to improvement in current techniques, appearance of new techniques, and alterations in acoustic structure of different languages from time to time.

### **3.2 Speech Recognition of Children with Down Syndrome**

Down syndrome, also known as trisomy 21, is a genetic disorder caused by the presence of all or part of a third copy of chromosome 21 [10]. It is typically associated with physical growth delays, characteristic facial features, and mild to moderate intellectual disabilities [12].

It affects 1 of 800-1000 newborns [13]. Children with Down syndrome have been examined to have delays and shortfalls in language proficiency, speech intelligibility, and speech fluency. Even though delay in language ability might have affected the speech intelligibility, Hamilton also focused on the poor speech clarity as the main obstruction to effective communication in the population with Down syndrome [14]. Basically, people with Down syndrome were reported to have more hypoplastic facial middle third, reduced nasal protrusion, and a reduced mandible size than their normal counterparts. Some researchers hypothesized that Macroglossia causes unintelligibility of speech in children with Down syndrome, hence the suggestion for tongue reduction surgeries [15]. Macroglossia is a medical concept for an unusually large tongue. Serious enlargement can lead to functional difficulties in speaking. It is uncommon and usually occurs in children. Recent studies claim another hypothesis, which advocated that the vocal tracts of children with Down syndrome are smaller than their normal comrades, so their tongues may appear large in their oral cavity. Such hypotheses have been braced by many modern theorists [16].

These children are delayed in the use of meaningful speech and are slow to obtain a fruitful vocabulary. In some cases, their speech remains unintelligible throughout childhood and adolescence, making it difficult to communicate with those around them [17]. The purpose of this thesis is to develop a speech recognition system for children with Down syndrome to ease their and their parents' lives and communicate with others almost normally.



### **3.2.1 The Phonology of Single Words**

Generally, word productions of children with Down syndrome have the same phonological characteristics as those of children with typical development [18]. Specifically, stop, nasal, and glide consonants tend to be produced accurately while fricatives, affricates, and liquids are often in error [19]. Phonological process analyses have also emphasized similarities between children with Down syndrome and those with typical development with the following patterns occurring frequently: (1) consonant clusters are produced as singleton consonants; (2) word-final consonants are omitted; (3) target fricatives and affricates are produced as stops; (4) aspirated voiceless stops in initial position are deaspirated; (5) word-initial liquids are produced as glides and word-final liquids are produced as vowels or are omitted; and (6) word-final voiced obstruent sounds are devoiced [20].

# Chapter 4 Voice Recolonization System Modeling and Simulations

## 4.1 Data Collection

In the data collection stage, two children (one with Down syndrome and one with typical development) were asked to pronounce two English words. Their speech was recorded using a mobile phone first, then a computer microphone with sampling frequency of 48kHz, 24-bit per sample, one channel WAV format was used. Each child read every word with a recording time of 3 seconds. The words tested were apple and cat. Each child read every word twelve times to be used later during process.

The two children were both males at 10 years of age. They were chosen at this age because at this point, their abilities in communicating with others are better than those of a younger age. They were selected randomly from a school and their identity was not collected or reported to protect their privacy and identity. Moreover, the voice recording procedure was conducted in a school setting.

## 4.2 Software

Two software programs were used during the development of the recognition system:

- MATLAB R2015b (**matrix laboratory**): is used in writing the codes of the system. MATLAB is a very high-performance language for technical computing. It is a fourth-generation programming language developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and information, utilization of algorithms, development of user interfaces,

and interaction with programs and scripts written in other languages including C, C++, Java, VHDL, and Python [21]. It integrates computation, visualization, and programming in an easy-to-use environment where issues and their solutions are expressed in known mathematical notation.

- Altera Quartus II 15.0: is a programmable logic device design software generated and designed by Altera®. It allows analysis and synthesis of HDL design, which enables the developers to compile and run their designs and perform timing analysis. Quartus includes an implementation of VHDL for hardware description, visual editing of logic circuits, and vector waves compilation and simulation.

Quartus II software has many features including: SOPC Builder, SoCEDDS, external memory interface toolkit, and Qsys. Qsys is a system integration tool that is the next generation of a SOPC Builder. It uses an FPGA-optimized network-on-chip-architecture. In this work, Qsys was used to generate Altera® Fast Fourier Transform II (FFT MegaCore) that includes Bit-accurate MATLAB models. The FFT MegaCore function is a high-performance, highly-parameterizable Fast Fourier Transform processor. More details about this core will be discussed in coming sections. Fortunately, MATLAB codes can be converted into VHDL codes by what is so called a HDL coder toolbox. HDL coder generates portable, synthesizable VHDL code from MATLAB functions. The generated code can be utilized for the FPGA programming. HDL coder offers a workflow advisor that automates the programming of Altera® FPGA. Developers can control HDL architectures and implementation, highlight critical paths, and generate hardware resource

utilization estimates. HDL coder provides traceability between Simulink model and the generated VHDL code, enabling code verification for high-integrity applications.

### **4.3 System Datasets and Parameters**

The dataset used in the testing phase consisted of 48 samples (2 children \* 2 words \* 12 repetitions) recorded with the same training children in a clean environment, where sounds were recorded using laptop microphone with sampling frequency of 48 kHz, 24-bit, Waveform Audio File Format (WAV). Child C is with Down syndrome, while child 1 is with typical development. The speech recognition system is implemented and tested using MATLAB R2015b also a MATLAB testbench generated by Quartus II. The laptop used in recording of sounds and computation of the system is a TOSHIBA laptop with these specifications: AMD Athlon™ II P320 Dual-Core Processor at 2.10 GHz, 3 GB RAM, 64-bit operating system, and Windows 8.1.

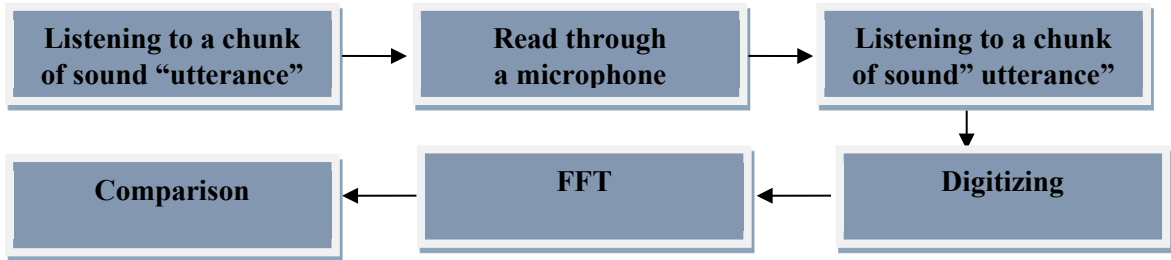
### **4.4 System Block Diagram**

What is known about human speech processing is still very limited [22]. The recognition process, in general, consists of multiple steps [23].

The first step is to listen to a chunk of sound (utterance), and then this utterance is captured and recorded through a microphone. In the digitizing step, the sound is digitized by a piece of hardware or software called Analog to Digital Converter (ADC) and in this thesis, MATLAB performs this task. ADC converts the analog (continuous-time, continuous-amplitude) speech signal into a digital (discrete-time, discrete-amplitude) speech signal. A speech recognition system doesn't need a Digital to Analog Converter (DAC) because the results of processing remain in digital form. After digitizing the input

signal, it will be converted from time domain into frequency domain using Fast Fourier Transform (FFT) algorithm to extract and identify the features of the spoken word.

In the comparison stage, the spoken word is compared with the word suggested in the dictionary. Speech recognition steps are summarized in Figure 3.



**Figure 3 Voice recognition process**

The speech recognition system includes two stages; a training stage and a recognition stage, both of which have common blocks that are wave recording, speech pre-processing and feature extraction. In the training stage, the system is trained by building its own dictionary. The word must be recorded many times. For example, repeating the word "Apple" many times with a pause between each utterance. The output of the training stage is a reference model.

In the recognition stage, the extracted features are compared with the reference model in the previous stage. Also, the word that has the best match will be displayed as an output. Figure 4 shows a block diagram of the overall system.

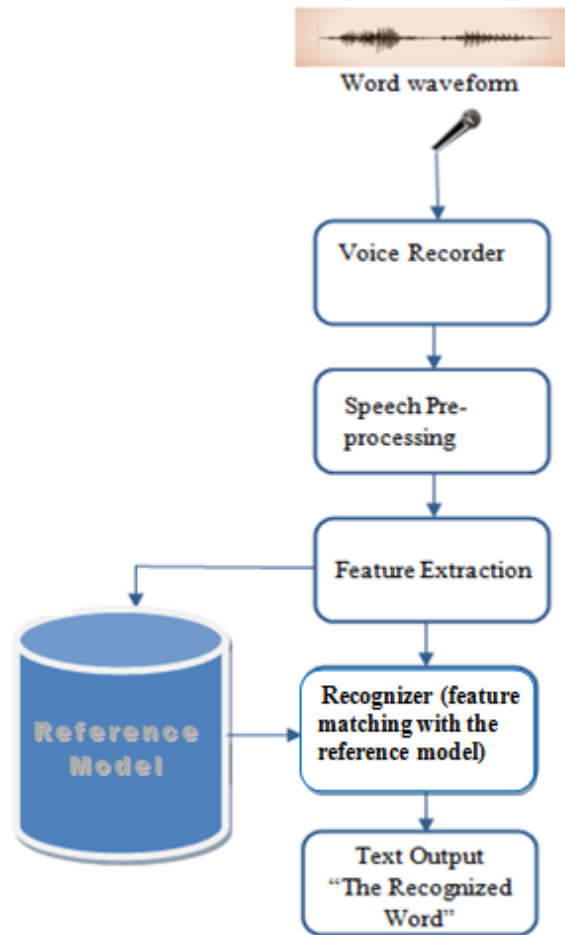


Figure 4 Voice recognition flow diagram

#### 4.5 Voice Recording

Sound is what the ear anticipates and the human ear is sensitive to acoustic pressure waves typically between 15 Hz and about 20 kHz with some sensitivity variation in that range [24]. As mentioned before, every child has to complete a voice recording session of two words repeated twelve times in order to use them in the training stage. These words were recorded by a system consisting of a microphone which converts air pressure variation

into a continuous-time voltage signal. In addition, this microphone converts sound waves into electrical audio signals that will be amplified to a usable level.

The sampling rate frequency was chosen to be 48000 Hz because the maximum frequency the human ear can hear is 20 kHz, and according to the Nyquist theorem, which states that the signal should be sampled for all time at a rate more than twice the highest frequency at which its Continuous Time Fourier Transform(CTFT) is non-zero.

Now the question is “can the original signal be reconstructed?” And to answer, yes it can, but only if the sampling frequency is greater than twice the maximum frequency of the signal being sampled, or equivalently:

$$\text{sampling Rate} \geq 2 \times f_{max}$$

Audio waveforms are typically sampled at 44.1 kHz (CD), 48 kHz, 88.2 kHz, or 96 kHz [25].

## **4.6 Pre-processing Stage**

Pre-processing is the fundamental signal processing applied before extracting features from speech signal, for the purpose of improving the performance of feature extraction algorithms also to improve recognition accuracy.

### **4.6.1 DC Offset Removal**

The initial speech frequency signal often has a constant component, i.e. a non-zero mean. This is typically due to DC bias within the recording instruments. DC offset occurs when hardware, such as a sound card, adds DC current to a recorded audio signal. This current produces a recorded waveform that is not centered on the baseline. Therefore, removing this DC offset is the process of forcing the input signal mean to the baseline [26].

It is important to remove the DC component to decide the boundary of the spoken word. The DC component can be easily removed by subtracting the mean value from all samples within an utterance.

#### **4.6.2 Amplitude Normalization**

Recorded signals often have varying energy levels due to child's volume and microphone distance. Amplitude normalization can cancel the inconsistent energy level between signals, thus can enhance the performance in energy-related features. There are several methods to normalize signal's amplitude. One of them is to divide the speech signal by the maximum of absolute value of the signal, so that the dynamic range of the signal is constrained between -1.0 and +1.0.

#### **4.6.3 Butterworth filter**

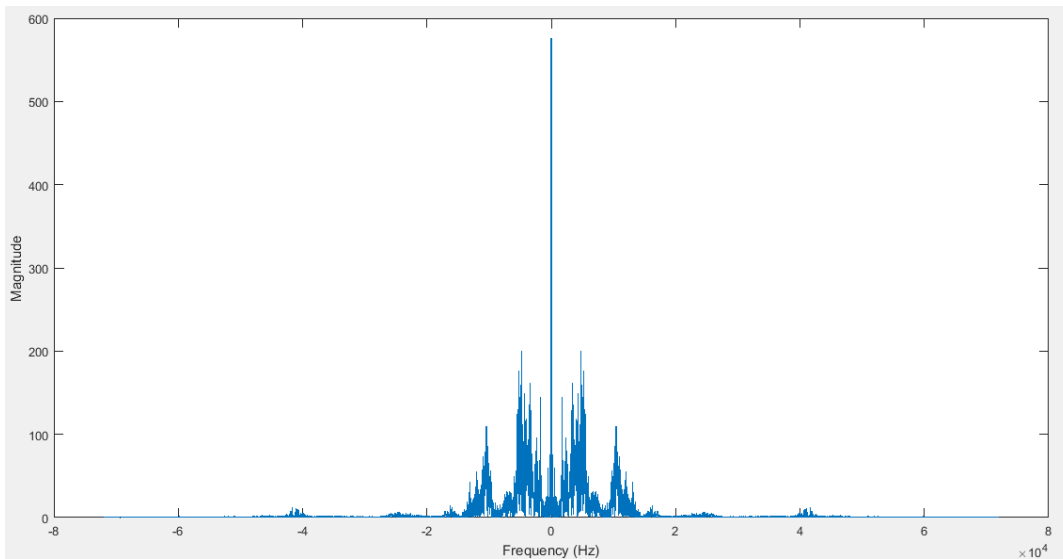
In this work, a low-pass Butterworth filter was used with the speech signals of child with Down syndrome to eliminate the noise added to child's speech signals by many factors. This filter designed to have as flat a frequency response as possible in the passband. Also, it keeps the frequencies in the chosen range only, so if there are noises in the speech signal, this filter will remove them [27]. Generally, the normalized cutoff frequency, for a Butterworth filter, must be between 0 and 1 with respect to the Nyquist frequency (sampling frequency/2).

In this case, the filter cuts off about 1200 Hz. So, the normalized cutoff frequency is calculated using this formula:

$$\text{Normalized cutoff frequency} = \frac{\text{Desired cutoff frequency}}{\left(\frac{\text{Sampling frequency}}{2}\right)}$$

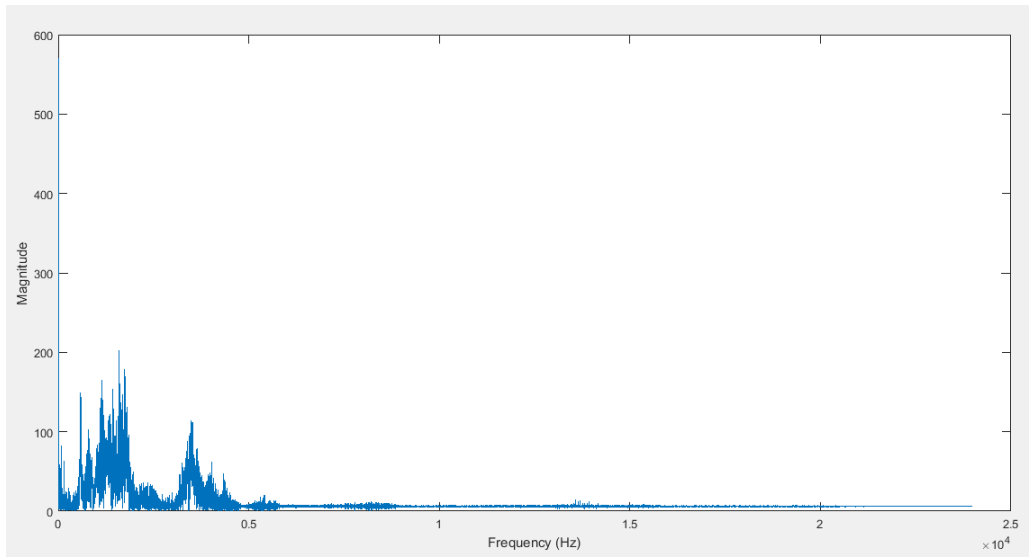


Consequently, the normalized cut off frequency is 0.05. In MATLAB, use `butter()` and `filter()` to setup a low-pass Butterworth filter. In this thesis, one child with Down syndrome results were recorded. Figure 5 shows a shifted FFT signal of child C saying the word “Apple” before using a Butterworth filter.

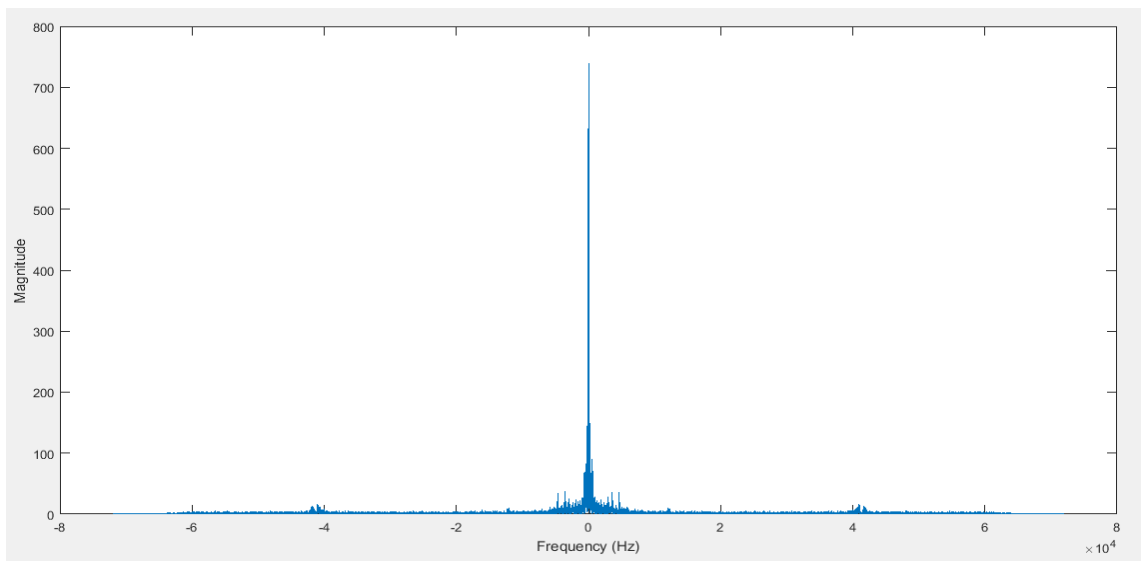


**Figure 5 Shifted FFT signal of child C saying the word “Apple”**

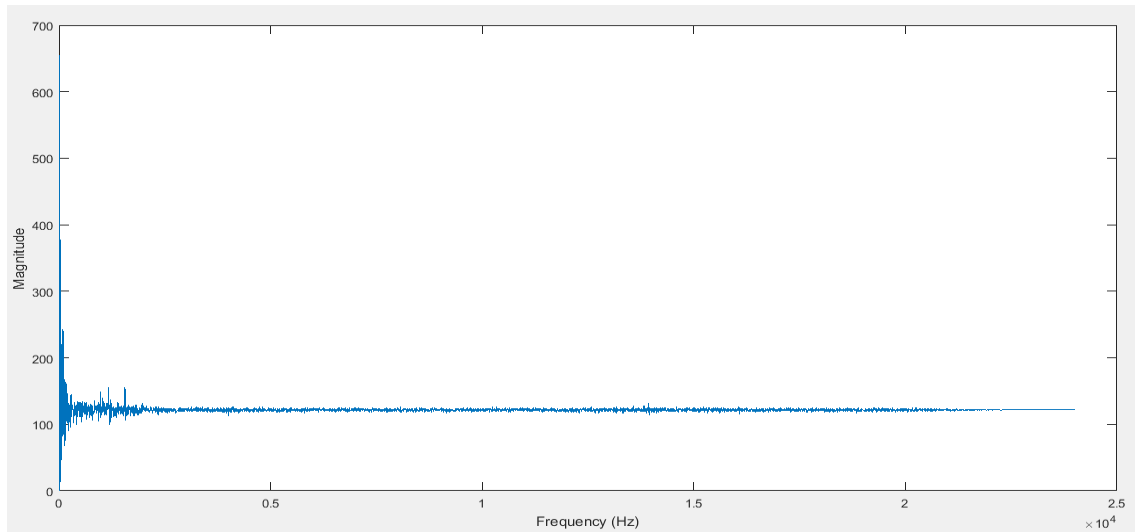
By using `fftshift` in MATLAB, the frequency center was shifted to 0 Hz. The left frequency span is from 0 to -80,000 Hz, while the right frequency span is from 0 to 80,000 Hz. Ideally, the frequency distribution for a negative frequency should equal the positive frequency. Only the positive frequency range was focused, since FFT plot is symmetric. Also, the frequency axis was scaled. Figure 6 shows noisy single-sided shifted FFT for child C saying the word “Apple” before applying a Butterworth filter. From Figures 7 and 8, it can be observed that Butterworth filter suppresses noise from the speech signal by the specified cutoff frequency.



**Figure 6 Single-sided FFT signal of child C saying the word “Apple”**



**Figure 7 Double-sided scaled and filtered signal**



**Figure 8 Single-sided scaled and filtered signal**

#### **4.7 Endpoint Detection (Word Boundary Detection)**

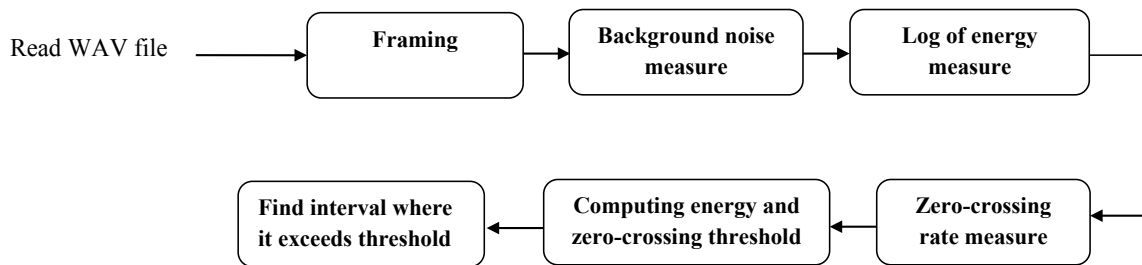
The performance of speech recognition system is sometimes degraded in adverse environments. Accurate speech endpoint detection is crucial for robust speech recognition. The speech endpoint detection is used to determine the speech and non-speech phases, it plays the key role in speech signal processing. Imprecise endpoint detection reduces the recognition ratio and increases the computation for processing the speech. During the last decades, several endpoint detection methods have been developed. Those methods can be categorized, approximately, into two classes; one is based on thresholds and the other is pattern-matching method. In general, the first kind extracts the acoustic features for each frame of signals and then compares these values of features with preset thresholds to classify each frame. Pattern-matching method requires estimation of the model parameters of speech and noise signal. The method that is based on pattern-matching is a high-accuracy method, but the disadvantages of using it are model dependency, high complexity and excessive computation. It is difficult to apply for the real-world speech signal processing system. Compared with pattern-matching method, threshold-based method is simpler and

faster since it does not need to keep much training data and train models. Common short-time energy and zero-crossing rate method is part of this sort, but it is sensitive to different types of noise and cannot fully identify the characteristics of a speech signal [28].

In this work, endpoint detection was used to extract the spoken word features and remove background noise and silence at the beginning and end of the spoken word. Endpoint detection improves ASR system performance in terms of accuracy and speed. Classification of speech into silence, voiced or unvoiced sounds, offers a useful basis for subsequent processing [29].

- Silence: When no speech is produced.
- Unvoiced: Vocal cords are not vibrating (do not entail the use of the vocal cords) shaping in a periodic or aperiodic speech waveform.
- Voiced: Vocal cords are vibrating during the pronunciation of a phoneme because they are tensed, resulting in speech waveforms that is quasi-periodic. Quasi-periodic means that the speech waveform can be noticed as periodic over a short period of time (5-100 ms) during which it is stationary.

Speech endpoint detection has multiple steps, which are illustrated in Figure 9.



**Figure 9 Endpoint detection block diagram**

## 1<sup>st</sup> Step: Framing and Removing Silence

In order to have a stationary sound we need to subdivide the sound wave into small frames. Typically, ASR systems use a frame size between 10 ms and 30 ms with 50% frame overlap [30]. In this system, a frame size of 10 ms with 50% overlap and the following parameters was used:

- $F_s=48000$  Hz
- *Record duration =3 seconds*
- *Signal length=Record duration  $\times$   $F_s=144000$  samples*
- *Frame duration=10 ms =0.01 seconds*
- *Number of frames=Signal length  $\div$  Frame Length =300 frames*

After analyzing the signals in time domain, it was found that the maximum amplitude of the silent sound is 0.03, so the threshold value was chosen to be 0.05. A code was written to calculate the maximum for each frame and compare it to the threshold value to decide, through an IF statement, whether it is a silent or non-silent frame. Moreover, this code computes and shows the number of the non-silent utterances in the WAV file.

## 2<sup>nd</sup> Step: Background Noise Measure

A noise-estimation algorithm is proposed for highly non-stationary noise environments [31]. Noise estimate is a crucial part and it is important for speech enhancement algorithms. If the noise estimate is too low, annoying residual noise will be audible and if the noise estimate is too high, speech will be distorted resulting possibly in eligibility loss. A very well-known approach of noise estimation is to estimate and update the noise spectrum during the silent (pauses) segments of the signal [32].

### **3<sup>rd</sup> Step: Log of Energy Measure**

The energy associated with speech is time varying in nature [33]. By the nature of speech production, the speech signal consists of voiced, unvoiced and silence regions. As mentioned before, speech signal may be stationary when it is viewed in blocks of 10-30 ms. Thus, to process speech signal by different signal processing techniques and tools, it is viewed in terms of 10-30 ms. Such a processing is described as Short Term Processing (STP). The concern for any automatic processing of speech is to know how the energy is varying with time and to be more particular, energy combined with short term region of speech. Short-term energy is the dominant and most natural feature that has been used in this system. It is noticed that short-term energy is the most effective energy criterion for this task. Voiced speech has most of its energy recorded in the lower frequencies, whereas most energy of the unvoiced speech is collected in the higher frequencies. The feeling of the sound intensity recognized by human ears is not linear but rather logarithmic. Hence, it is better to express the energy function in logarithmic form [34, 35].

In the endpoint detection, the zero crossing and the log of energy were used to find the boundary of the words as they are simple, speedy, and authentic. The energy of unvoiced sounds is usually lower than voiced sounds, but higher than silence. So, short-term energy can be used for voiced, unvoiced and silence categorization of speech signal [36].

### **4<sup>th</sup> Step: Zero-crossing Rate Measure**

Zero-crossing rate provides information about the number of zero-crossings present in a given signal. The number of zero crossings alludes to the number of times

speech samples change sign in each frame. The zero-crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced because of excitation of the vocal tract by the periodic flow of air at the glottis, and usually shows a low zero crossing count. Unvoiced speech is produced due to excitation of the vocal tract by the noise-like source at a point of constriction in the interior of the vocal tract and shows a high zero crossing count. The zero-crossing count of silence is expected to be lower than that of unvoiced speech, but quite comparable to that for voiced speech. The following equation can be used to calculate the zero-crossing rate [36].

$$Zcr(m) = \sum_{n=1}^N \frac{|sgn(S_m(n)) - sgn(S_m(n-1))|}{2} \quad (4.1)$$

Where:

$Zcr(m)$  is the zero-crossing rate of the frame  $m$

$$sgn(S_m(n)) \text{ is the sign function } = \begin{cases} 1, & S_m(n) \geq 0 \\ -1, & S_m(n) < 0 \end{cases} \quad (4.2)$$

$S_m(n)$  is the speech signal in the sample number  $n$  in the frame  $m$ .

$N$  is the frame size.

By combining the energy and the zero crossing rates, the type of speech can be approximately concluded. Considering the voiced sound, the voiceless sound and the silent part, the short-time energy average per frame of voiced sound is the biggest and the short-time zero rate is the lowest; the short-time energy average per frame of voiceless sound comes in between, but the short-time zero rate is the highest; the short-time energy average per frame of the silent part is the lowest and the short-time zero rate comes between. This kind of comparison is relative, but has no actual value relations [37].

### 5<sup>th</sup> Step: Calculating Energy and Zero-crossing Threshold

The utterance should be determined at which it begins and where it ends. For that, we need to find and compute the energy threshold. The energy threshold can be described as:

$$T_E = \mu_E + \alpha \times \sigma_E \quad (4.3)$$

Where:  $\mu_E$  is the mean and  $\sigma_E$  is the standard deviation of the energy of the frames. The  $\alpha$  term is a constant that must be fine-tuned according to the characteristics of a signal. Several values of  $\alpha$  were tested in the range from zero to one and it was found that the best word boundary detection and system accuracy are with  $\alpha=0.5$  [36].

Zero-crossing threshold can be obtained by using the following equation:

$$\mu_Z + \beta \times \sigma_Z \quad (4.4)$$

Where  $\mu_Z$  is the mean and  $\sigma_Z$  is the standard deviation of the zero-crossing rates of the frames and  $\beta$  is a parameter which is found through past experiments. In this thesis, it was found that the most decent value for the threshold factor is  $\beta = 0.5$  based on multiple calculations [38]. Moreover, based on much research, the zero-crossing rate of speech signals should be greater than 25 zero-crossings per frame [38]. In this system, crossing thresholds were set to:

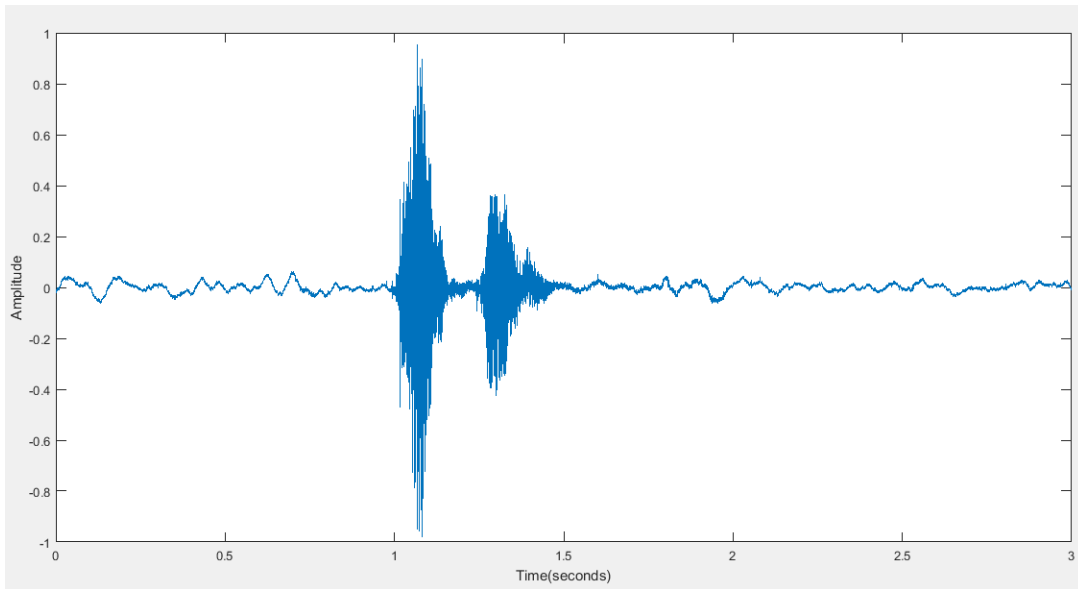
$$T_Z = \max(\mu_Z + (\beta \times \sigma_Z), 25) \quad (4.5)$$

### 6<sup>th</sup> Step: Find Interval where it Exceeds Threshold

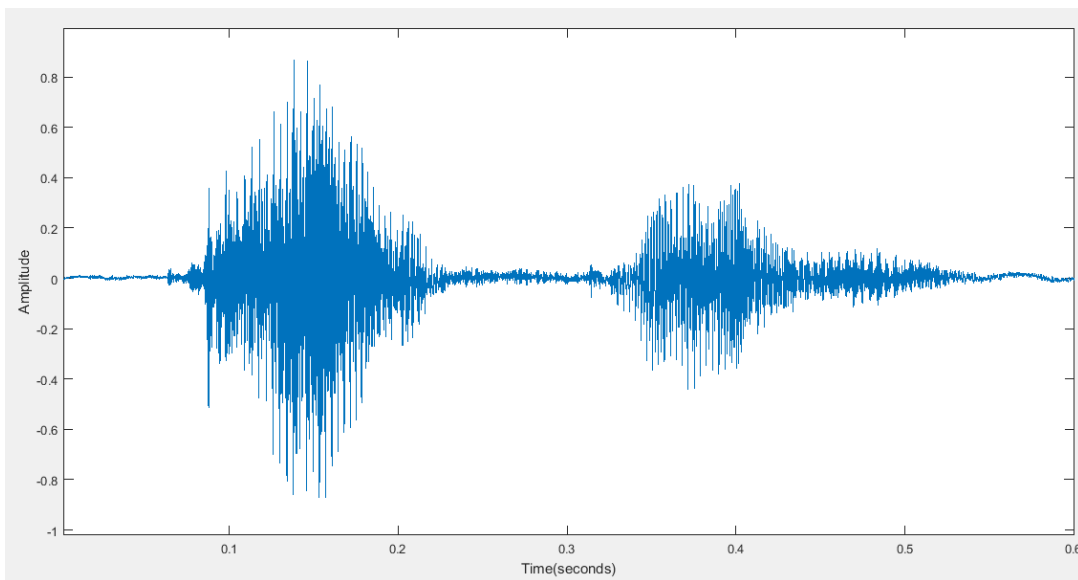
By finding the energy and zero-crossing rate thresholds, the resultant threshold can be subtracted from the original speech. In this step, we test each frame by comparing its energy and zero-crossing rates with the thresholds. Figure 10 and Figure 11 are time-



domain plots of the signal of a spoken word “Apple” by child C before and after applying endpoint detection.



**Figure 10** The waveform of the word “Apple” before applying endpoint detection

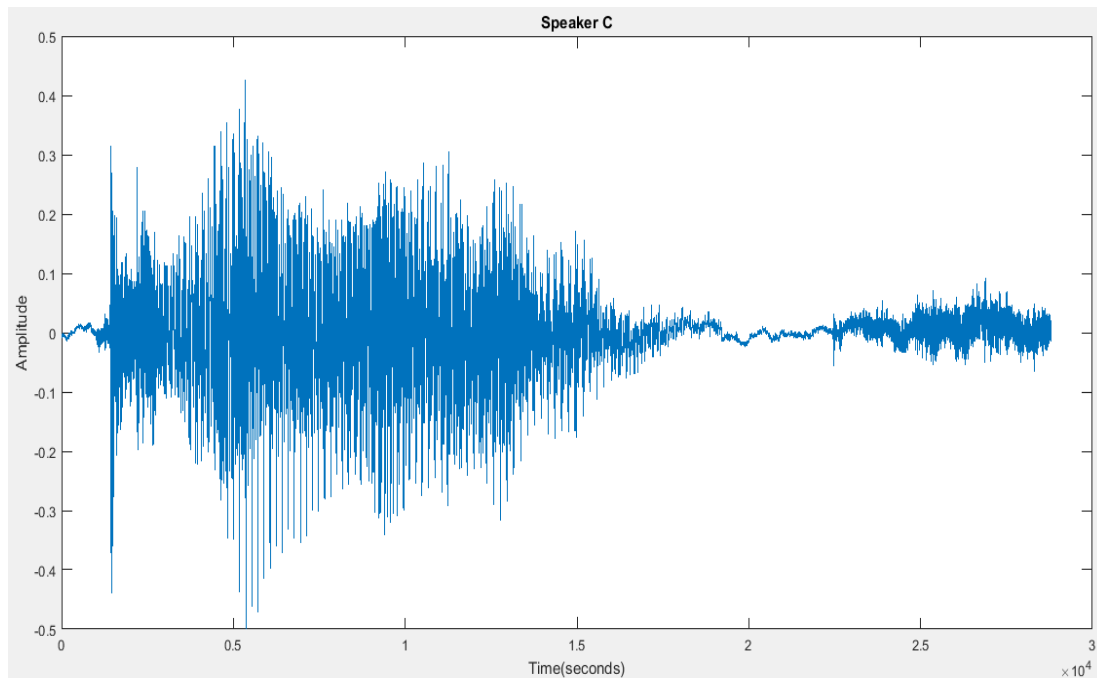


**Figure 11** The waveform of the word “Apple” after applying endpoint detection

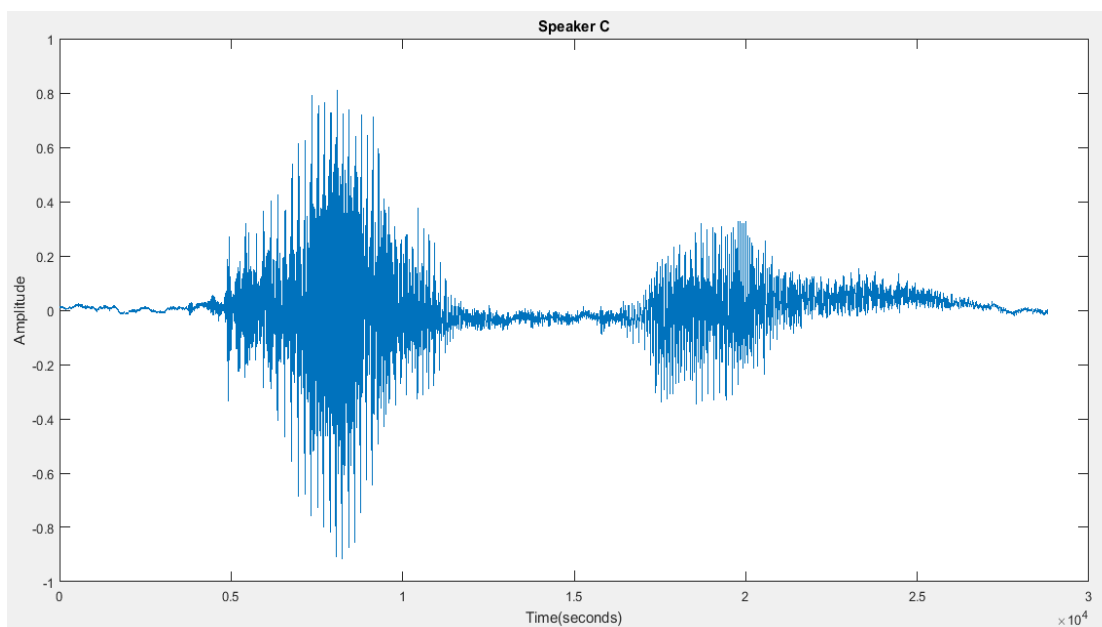
Figure 10 illustrates the time-domain signal of word “Apple” by a child with Down syndrome, the signal length is 3 seconds or 144000 samples. In Figure 11, the signal length of the word “Apple” spoken by a child with Down syndrome child applying the endpoint detection becomes to 23999 samples or 0.499 seconds, where silence at the beginning and end of the word is removed. This will decrease the computation time of the feature extraction and recognition in the system to about 83%  $\left(\frac{3\text{sec}-0.499\text{sec}}{3\text{sec}} \times 100\%\right)$  approximately.

#### **4.7.1 Time-domain waveforms**

To compare the time waveforms of the two different spoken words by different children, Figure 12 shows the time-domain waveform of a child with Down syndrome pronouncing the word “Cat”. However, Figure 13 shows the time-domain waveform of a child with Down syndrome pronouncing the word “Apple”.



**Figure 12** Voice recording of “Cat” for child c with Down syndrome



**Figure 13** Time-domain plot of a child with Down syndrome of the word “Apple”

Overall, the time-domain plots in Figure 12 and in Figure 13 are similar at some time intervals, and it is difficult to distinguish which word was spoken from the time-domain plot. Children voices were recorded and plotted by MATLAB commands. There are differences in the magnitude of the waveforms. The amplitude represents the loudness of the sound which gives irrelevant information about which word was spoken. Therefore, the amplitudes of the waveforms were not compared.

By visual inspection, it is ineffective to identify individual frequency components from the time-domain plots. The frequency plot is preferred to analyze the frequency components of the waveforms. To get the frequency components of the waveforms, FFT should be applied on the time-domain signals to analyze them in frequency domain. In this thesis, FFT was used as a feature extraction technique.

#### **4.8 Feature Extraction**

The goal of feature extraction is to represent any speech signal by a finite number of measures (or features). This is because the entirety of the information in the acoustic signal is too much to be processed, and not all information is appropriate for specific tasks. In ASR systems, the approach of feature extraction has generally been to find a representation that is relatively stable for different examples of the same speech sound, despite differences in the speaker or environmental characteristics, while keeping the part that represents the message in the speech signal relatively undamaged. FFT was used in this research as a feature extraction technique.

A key assumption made in the design of most speech recognition systems is that the segment of a speech signal can be considered as stationary over an interval of few milliseconds. Therefore, the speech signal can be divided into blocks which are usually

called frames. That is, the frames are overlapping to provide longer analysis windows. Within each of these frames, some feature specifications characterizing the speech signal are extracted. These feature parameters are then used in the training and in the recognition stage [39]. Feature Extraction and Classification are the two dominant stages of speech processing and among speech recognition stages; feature extraction is a key, because a better feature is good for improving recognition ratio.

#### 4.8.1 Fast Fourier Transform (FFT)

The FFT is an effective way of computing the Discrete Fourier Transform (DFT). DFT is equivalent to Fourier Transformation in continuous signals. It is, however, used to transform a discrete signal to a discrete frequency spectrum [40].

The Fourier Transform has countless applications ranging from speech recognition to astronomy, from Radar to mobile phones. So, it is one of the most fruitful mathematical techniques ever invented and created. For transforming the discrete-time signal from time domain into its frequency domain, the FFT is nothing but the DFT, except the difference is that the FFT is faster, more efficient, and more accurate on computation. So, it is convenient to investigate FFT by firstly considering the N-point DFT equation which is given by:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{j\omega_k n} \quad (4.6)$$

where  $x(n)$  is the input,  $e^{j\omega_k n}$  is the phase factor,  $n$  and  $k$  are integers from 0 to  $N-1$  [41]. First of all, we must separate  $x(n)$  into two parts:  $x(\text{odd})=x(2m+1)$  and  $x(\text{even})=x(2m)$ , where  $m=0, 1, 2, \dots, N/2-1$ . Then the N-point DFT equation also becomes two parts, and ends up with the following equation:

$$\begin{aligned}
X(k) &= \sum_{m=0}^{\left(\frac{N}{2}\right)-1} x_1(m) W_{N/2}^{mk} + W_N^{mk} \sum_{m=0}^{\left(\frac{N}{2}\right)-1} x_2(m) W_{N/2}^{mk} \\
&= X_1(k) + W_N^k X_2(k) \quad , k=0, 1, \dots, N/2
\end{aligned} \tag{4.7}$$

$$X\left(k+\frac{N}{2}\right) = X_1(k) - W_N^k X_2(k) \quad , k=0, 1, \dots, N/2$$

Where  $e^{jw_k n} = W_N^{kn}$ . So here the N-point DFT is separated into two N/2-point DFT. For original N-point DFT Eq. (4.6), it has  $(N^2)$  complex multiplications and N/2-point DFT Eq. (4.7) has  $(N^2/2) + (N/2)$  multiplications. This is the operation for reducing the calculations from N points to N/2 points. This signal for N point DFT is continuously subdivided until the final signal sequence is reduced to the one point sequence. So, the total number of complex multiplications will be approximately reduced to  $(N/2) \log_2(N)$  [41].

The FFT returns a set of complex numbers (amplitude and phase), with exception of the spectral components at  $f=0$  and  $f=fs/2$  (*Nyquist Rate*). DFT has linearly spaced frequency bands, as its bins are spaced at intervals of  $(Fs/N)$ , where N is the length of the DFT vector (number of points) and Fs is the sampling rate. The MATLAB command that was used to obtain the FFT is `fft(x)`. This command computes the DFT of x using a Fast Fourier Transform algorithm. To determine the magnitude values of the FFT output, `abs(fft(x))` will return the magnitude only, and that is what we need to start analyzing the speech signals. It was performed 512-point DFT on the time-domain samples using MATLAB and Altera® Quartus II (FFT MegaCore function).

The FFT MegaCore function implements a complex FFT or inverse FFT (iFFT) for high-performance applications. The FFT MegaCore function implements the following architectures:

- Fixed transform size architecture.
- Variable streaming architecture.

The FFT MegaCore function is included in Altera® Quartus II libraries under the DSP library and it is powered by Qsys. The proposed system specifications are showed in Table 1.

**Table 1 FFT MegaCore function Parameters**

<b>Function Parameters</b>	<b>Values</b>
Transform Length	512 points
Transform Direction	Forward
I/O Data Flow	Variable Streaming
Input Order	Natural
Output Order	Natural
Data and Twiddle Representation	Single Floating Point
Calculation Latency	512 cycles
Throughput Latency	1024 cycles

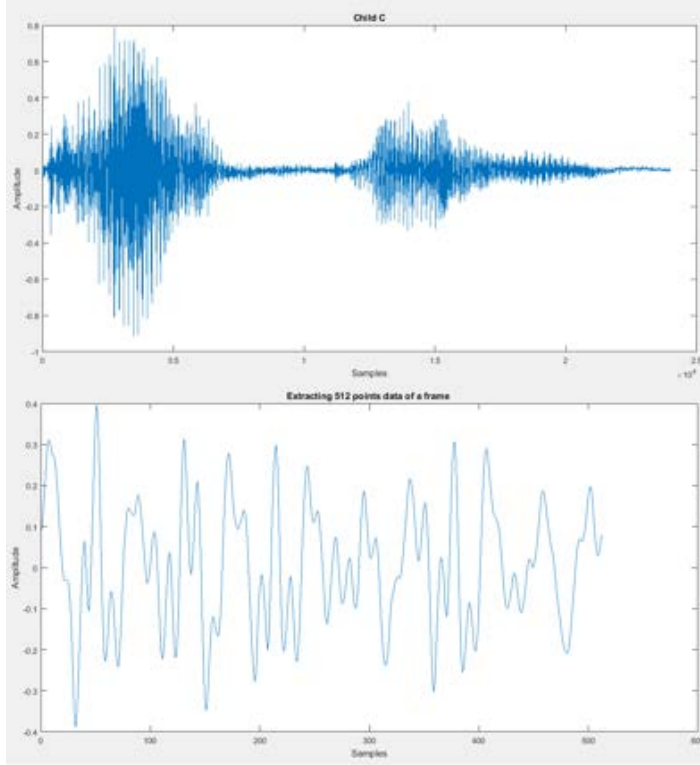
The streaming I/O data flow FFT architecture allows continuous processing of input data, and outputs a continuous complex data stream without the requirement to cutoff the data flow in or out of the FFT function. The variable streaming architecture FFT implements a radix-2<sup>2</sup> single delay feedback architecture, which developer can configure during runtime to perform the FFT algorithm for transform lengths of 2<sup>m</sup> where 4 ≤ m ≤ 18. This architecture utilizes either a fixed-point representation or a single precision floating point representation. In this research, the architecture used a single precision floating point that allows a large dynamic range of values to be represented while controlling a high Signal to Noise Ratio (SNR) at the output. Radix-2<sup>2</sup> single delay feedback architecture is a fully pipelined architecture for computing the FFT of incoming data. It is like radix-2 single delay feedback architectures. There are  $\log_2(N)$  stages with each stage containing a single butterfly unit and a feedback delay unit that delays the incoming data by a specified number of cycles, halved at every stage where N is the transform length [42].

Quartus II software allows the user to generate a Testbench system, which is a new hardware system that instantiates the system under test, and gives functional models to drive the top-level interfaces. Once generated, the bus functional models can interact with the system in the simulator. The Testbench compiles and runs easily in MATLAB environment. This MATLAB Testbench exercises the Altera® FFT Model, and is generated by Altera's FFT MegaCore to output results to the attached text files that are in the same directory with the Testbench files.

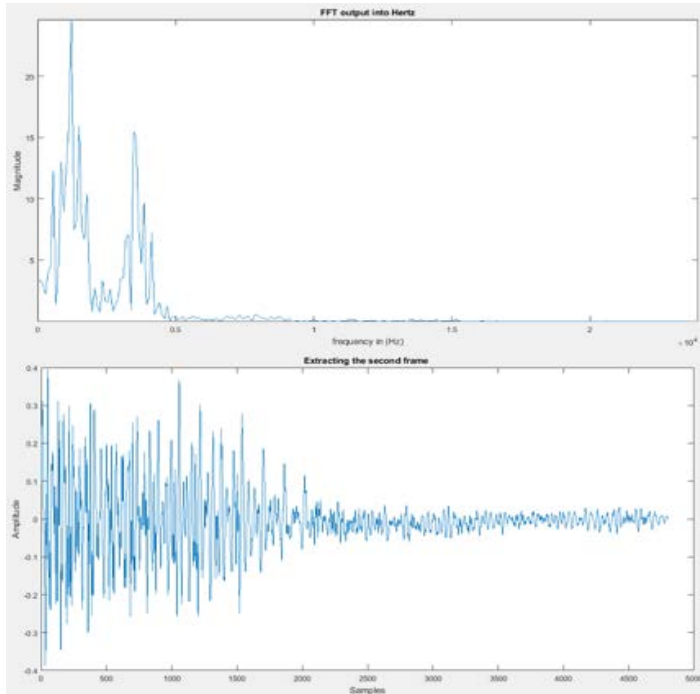
Altera's FFT MegaCore Testbench takes the input data as an array of points, so the speech signal should be plotted and subdivided into frames in order to choose any frame



and extract 512 points data from it. Child C was chosen to do some experiments on his time-domain speech signal of the word “Apple”. First, the time-domain speech signal was shifted up by adding a certain scale, and picked the second frame from the resulting output. After that, 512 points were extracted from it to be fed to MATLAB interfaces. In Figure 14, child C voice pronouncing the word “Apple” was recorded, then plotted the time-domain signal to extract any frame from it. The second frame was chosen and extracted to simplify the computation of FFT and get 512 points data. The frequency domain data was plotted by MATLAB and the results are shown in Figures 14 and 15, each figure consists of two plots. In Figure 14, the first plot illustrates time-domain speech signal of the word “Apple”. The signal was broken into frames, and the second frame was chosen. The process of breaking the signal into frames and extracting a specific frame, was done through MATLAB code. The extracted 512 points data are shown in the first plot of Figure 15. Altera’s Testbench take the input as numbers stored in text file, so the 512 points data should be recorded from MATLAB, stored in the input text file and used in the test bench to find the FFT plot. Single-sided FFT is illustrated in the second plot of Figure 15 and shows the speech signal in frequency domain.



**Figure 14** Extracting a frame from time-domain signal



**Figure 15** Extracting 512 points of a frame

### 4.8.2 Formants

Formants are groups of frequencies near the harmonics. Harmonics are multiples of fundamental frequency. For example, if the fundamental frequency equals 100Hz then the harmonics will be 200, 300, 400, 500, ... Hz. The fundamental frequency is the first harmonic (highest frequency). Figure 16 illustrates how formants and the harmonics spikes can be determined.

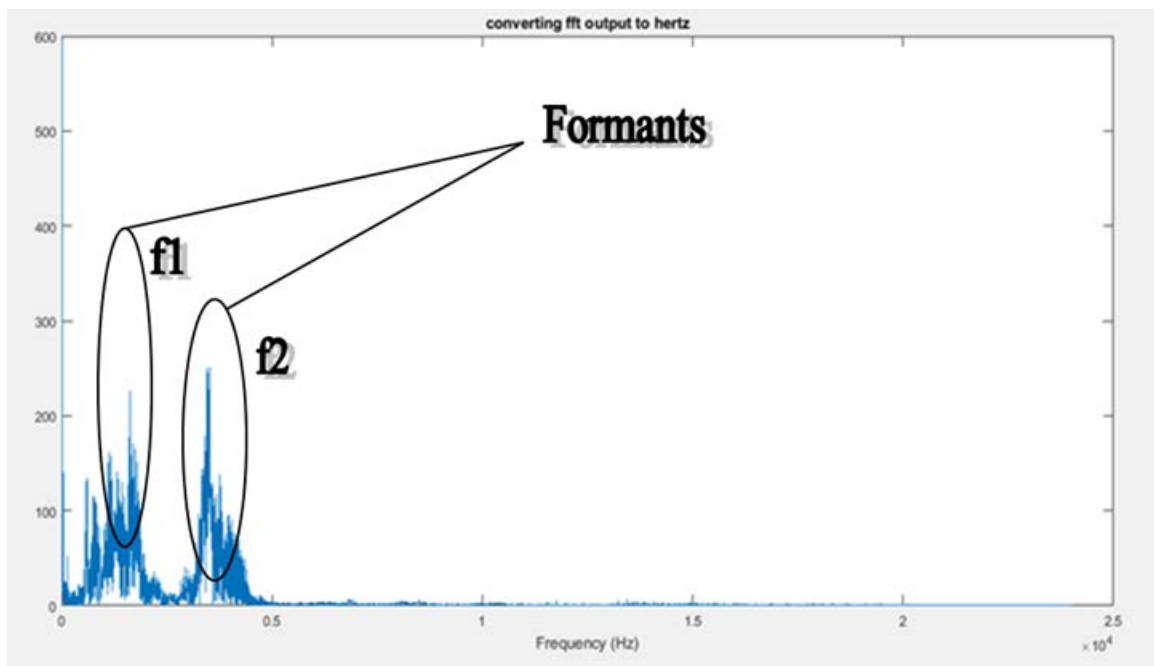
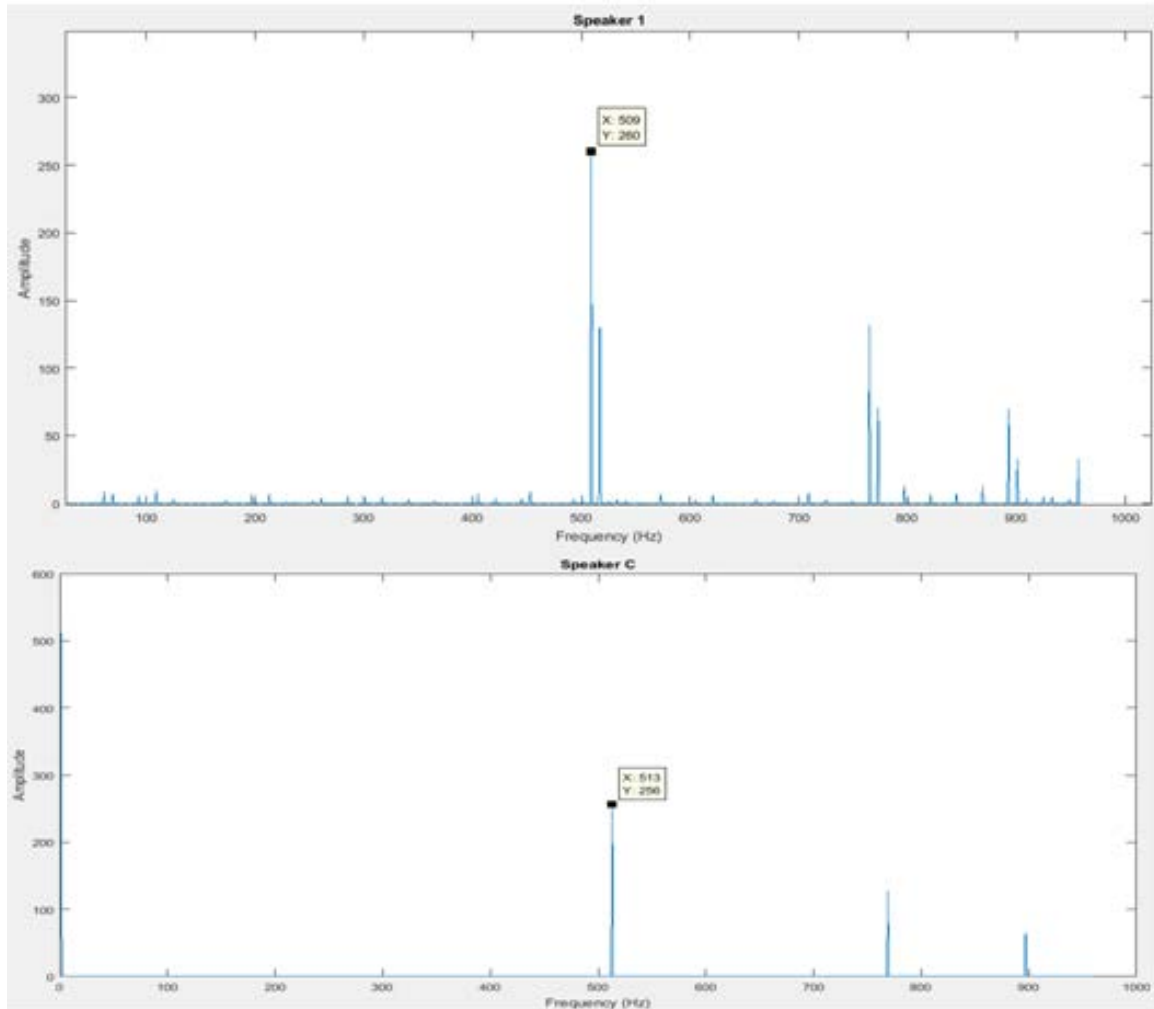


Figure 16 Formants Identification

The harmonic spikes are labeled as f1, and f2 respectively. The formants are circled in the plot. The formant frequencies, at which the spectral peaks are located, are the most important. After specifying the formants, formant frequencies can be easily specified and developer can determine if, or if not, are included in the audible range. Developers usually determine the first three formants because of their importance in detecting the spoken word.

Most of the words spoken across different accents around the world have some common frequency domain features that can be used to detect the spoken word. Figure 17 is a FFT plot of two children pronouncing the same word, each plot was generated by a MATLAB code to identify and determine the first three formants, and it illustrates the similarities between frequency-domain speech features of the two signals. The first plot shows the first three formants of a child with typical development. However, the second plot illustrates the first three formants of a child with Down syndrome. Formant frequencies were observed, from plots, as following:

- For child 1, the first formant frequency is 509 Hz, the second formant frequency is 765 Hz and the third formant frequency is 893 Hz.
- For child C, the first formant frequency is 513 Hz, the second formant frequency is 769 Hz and the third formant frequency is 897 Hz.



**Figure 17 FFT plots of two children pronouncing the same word**

It is observed that the first, second and third formants of the two children are within the audible range (15 Hz to 20 kHz) and the human ear can hear them. This observation is important and helps, in the future, to build a hardware detection system that can detect the spoken word and display it to the user, since MATLAB easily converts its code to VHDL using HDL coder toolbox. Then VHDL will be used to build the hardware and simulate it.

## **Chapter 5 Conclusion and Future Research**

This chapter summarizes the results of isolated word speech recognition for children with Down syndrome and gives some future directions that could make this system more flexible and effective.

### **5.1 Summary and Closing Remarks**

In this thesis, an isolated speech recognition system for children with Down syndrome was designed, based on a combination FFT as a feature extraction technique combined with formants. First, I used the pre-processing step as a word boundary detector using the energy and the zero crossing rates to automatically remove silences and determine the start and the end of the word in the input signal. Then some preparation steps to the speech signal were performed before extracting the features to improve the accuracy of the recognition and to make the system more sensitive to noise.

The system was implemented in MATLAB using an AMD Athlon™ II P320 Dual-Core Processor laptop. The dataset used in this system included two English words recorded in a calm environment with two different children using a laptop microphone. Each child read each word twelve times. The datasets of the system were recorded in a calm environment and the words were spoken with same style and same distance to microphone to have almost the same loudness. In this thesis, I tested and showed that the first three formants of the speech signal have frequencies within the Audible Range that the human ear can hear (15Hz to 20 kHz). From here I can build, in the future, a hardware that can detect the spoken words and display them to the user. The performance could be improved by training the system with large datasets.

Finally, this system is one of the emerging systems that can improve speech recognition for children with Down syndrome and will have an important role in the future of speech technology research.

## **5.2 Future Work and Recommendations**

In the future, I should try to improve the system to move to the general case of an independent speaker with a large vocabulary in English using a continuous speech recognition system that is sensitive to noise and to the differences in speech style and loudness. Also, I should try the following techniques:

- Developing the system by using other noise cancellation techniques.
- Improve the performance of the system by using large training datasets.
- Extend the work to include more words.
- Select other feature extraction techniques such as Mel-frequency cepstral coefficients (MFCCs) and the image of the spectrum of speech, using image techniques to extract the features of the speech spectrum.
- Combine other different methods to improve accuracy in trade off computation time.
- Build hardware that detects the spoken words and displays them on a screen, since MATLAB deals perfectly fine with Altera®, VHDL codes, HDL coder, and Testbench.

Also, I recommend building a common English database, to benefit researchers and to reduce time in recording and gathering data, and also to be able to compare their various approaches with a single database.

## Bibliography

- [1] M.A. Anusuya and S.K. Katti, " Speech Recognition by Machine: A Review", *International Journal of Computer and Information Security*, Vol. 6, No. 3, 2009.
- [2] Hemdal, J.F. and Hughes, G.W., "A feature based computer recognition program for the modeling of vowel perception, in Models for the Perception of Speech and Visual Form", W. Ed. MIT Press, Cambridge, MA. 1964.
- [3] D. Tran, M. Wagner, "A robust clustering approach to fuzzy Gaussian mixture models for speaker identification", *Knowledge-Based Intelligent Information Engineering Systems 1999. Third International Conference*, pp. 337-340, 1999.
- [4] R.K. Moore, "Twenty things we still don't know about speech", *Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research and Technology*. 1994.
- [5] A. Samouelian, "Knowledge based approach to consonant recognition," *Acoustics, Speech, and Signal Processing, ICASSP-94., IEEE International Conference on*, Adelaide, SA, pp. I/77-I/80, 1994.
- [6] Tripathy, Hrudaya Ku, B. K. Tripathy, and Pradip K. Das. "A knowledge based approach using fuzzy inference rules for vowel recognition." *Journal of Convergence Information Technology* 3.1, 51-56. Vol. 3 No 1, March 2008.
- [7] L. R. Rabiner and B.-H. Juang, "Fundamentals of speech recognition", Pearson Education, 2005.
- [8] Joseph P. Campbell, JR., "Speaker recognition: A tutorial". *Proceedings of IEEE* Vol. 85. No. 9, September 1997.



- [9] "The Inner Ear," American Speech-Language-Hearing Association. [Online]. Available: <http://www.asha.org/public/hearing/Inner-Ear/>. [Accessed: 23-Apr-2017].
- [10] D. Patterson, "Molecular genetic analysis of Down syndrome," *Human Genetics*, vol. 126, no. 1, pp. 195–214, 2009.
- [11] W. Ghai and N. Singh, "Literature Review on Automatic Speech Recognition," *International Journal of Computer Applications*, vol. 41, no. 8, pp. 42–50, 2012.
- [12] M. Weijerman and J. de Winter, "Clinical practice", *European Journal of Pediatrics*, vol. 169, no. 12, pp. 1445-1452, 2010.
- [13] G. Laws, "Contributions of phonological memory, language comprehension and hearing to the expressive language of adolescents and young adults with Down syndrome", *Journal of Child Psychology and Psychiatry*, vol. 45, no. 6, pp. 1085-1095, 2004.
- [14] C. Hamilton, "Investigation of the articulatory patterns of young adults with Down syndrome using electropalatography", *Down Syndrome Research and Practice*, vol. 1, no. 1, pp. 15-28, 1993.
- [15] C. Sforza, C. Dellavia, M. Goffredi and V. Ferrario, "Soft-tissue Facial Angles In Individuals With Ectodermal Dysplasia: A Three-dimensional Non Invasive Study", *The Cleft Palate-Craniofacial Journal*, 2005.
- [16] C. Guimaraes, L. Donnelly, S. Shott, R. Amin and M. Kalra, "Relative rather than absolute macroglossia in patients with Down syndrome: implications for treatment of obstructive sleep apnea", *Pediatric Radiology*, vol. 38, no. 10, pp. 1062-1067, 2008.

- [17] C. Stoel-Gammon, "Down syndrome phonology: Developmental patterns and intervention strategies", *Down Syndrome Research and Practice*, vol. 7, no. 3, pp. 93-100, 2001.
- [18] H. Weyerts, S. Rosenberg and L. Abbeduto, "Language and Communication in Mental Retardation: Development, Processes and Intervention", *Language*, vol. 71, no. 3, p. 618, 1995.
- [19] K. Bleile and I. Schwarz, "Three perspectives on the speech of children with Down's syndrome", *Journal of Communication Disorders*, vol. 17, no. 2, pp. 87-94, 1984.
- [20] J. Borsel, "Articulation in Down's syndrome adolescents and adults", *International Journal of Language & Communication Disorders*, vol. 31, no. 4, pp. 415-444, 1996.
- [21] "MATLAB", En.wikipedia.org, 2017. [Online]. Available: <https://en.wikipedia.org/wiki/MATLAB>. [Accessed: 26- Apr- 2017].
- [22] F. Chen and K. Jokinen, *Speech technology*, 1st ed. New York: Springer, 2010.
- [23] R. Rezai, *Brain: computer interface system: recent progress and future prospects*, 1st ed. Croatia: InTech, 2013.
- [24] M. Roberts, *Fundamentals of signals and systems*, 1st ed. Boston: McGraw-Hill Higher Education, 2008.
- [25] D. Self, "Audio engineering explained", 1st ed. Oxford: Focal Press, 2010.
- [26] M. D. Al-Hassani, "Identification Techniques using Speech Signals and Fingerprints", Ph.D. Thesis, Department of Computer Science, Al-Nahrain University, Baghdad, Iraq, September 2006.

- [27] S. Butterworth, "On the Theory of Filter Amplifiers," *Wireless Engineer*, vol. 7, Oct. 1930.
- [28] K. Li, M. Fei, G. W. Irwin, S. Ma, "Bio-Inspired Computational Intelligence and Applications", International Conference on Life System Modeling and Simulation, LSMS, Shanghai, China, 2007.
- [29] H. Deng and D. O'shaughnessy, "Voiced-Unvoiced-Silence Speech Sound Classification Based on Unsupervised Learning," *Multimedia and Expo, 2007 IEEE International Conference on*, 2007.
- [30] D. Eringis and G. Tamulevičius, "Improving Speech Recognition Rate through Analysis Parameters", *Electrical, Control and Communication Engineering*, vol. 5, no. 1, 2014.
- [31] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.
- [32] J. Ramirez, J. M., and J. C., "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," *Robust Speech Recognition and Understanding*, Jan. 2007.
- [33] R. E. Remez and P. E. Rubin, "On the perception of speech from time-varying acoustic information: Contributions of amplitude variation," *Perception & Psychophysics*, vol. 48, no. 4, pp. 313–325, 1990.
- [34] D. Shete, P. S. Patil, and P. S. Patil, "Zero crossing rate and Energy of the Speech Signal of Devanagari Script," *IOSR journal of VLSI and Signal Processing*, vol. 4, no. 1, pp. 01–05, 2014.
- [35] W. A. Yost, "Spatial Hearing: The Psychophysics of Human Sound Localization, Revised Edition," *Ear & Hearing*, vol. 19, no. 2, p. 167, 1998.

- [36] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," *Advanced Techniques in Computing Sciences and Software Engineering*, pp. 279–282, 2009.
- [37] L. Rabiner and R. Schafer, *Digital processing of speech signals*, 1st ed. Englewood Cliffs: Prentice-Hall, 1978.
- [38] A. O. A. Noor, S. A. Samad, and A. Hussain, "Development of a Voice Activity Controlled Noise Canceller," *Sensors*, vol. 12, no. 12, pp. 6727–6745, 2012.
- [39] C. Yılmaz, "A Large Vocabulary Speech Recognition System for Turkish", MS Thesis, Bilkent University, Institute of Engineering and Science, Ankara, Turkey, 1999.
- [40] V. Ingle and J. Proakis, *Essentials of digital signal processing using MATLAB®*, 1st ed. [S.l.]: Cengage Learning, 2012.
- [41] T. Yang, "The Algorithms of Speech Recognition, Programming and Simulating in MATLAB." Student Thesis, University of Gävle, Faculty of Engineering and Sustainable Development, 2012.
- [42] S. He and M. Torkelson, "A new approach to pipeline FFT processor," *Proceedings of International Conference on Parallel Processing*, Honolulu, HI, 1996, pp. 766-770.1996.

June 1, 2017

Dr. Frank Li, Principal Investigator  
Ms. Janah Emeeshat, Co-investigator  
Department of Electrical & Computer Engineering  
UNIVERSITY

RE: HSRC PROTOCOL NUMBER: 195-2017  
TITLE: Isolated Word Speech Recognition System for Children with Down  
Syndrome

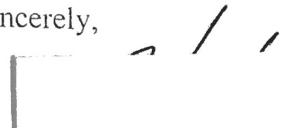
Dear Dr. Li and Ms. Emeeshat:

The Institutional Review Board has reviewed the abovementioned protocol and determined that it is exempt from full committee review based on a DHHS Category 1 exemption.

Any changes in your research activity should be promptly reported to the Institutional Review Board and may not be initiated without IRB approval except where necessary to eliminate hazard to human subjects. Any unanticipated problems involving risks to subjects should also be promptly reported to the IRB.

The IRB would like to extend its best wishes to you in the conduct of this study.

Sincerely,

  
Mr. Michael A. Hripko  
Associate Vice President for Research  
Authorized Institutional Official

MAH:cc

c: Dr. Jalal Jalali, Chair  
Department of Electrical & Computer Engineering

