WEB-BASED ARABIC SPEECH RECOGNITION SYSTEM

NURUL ASHIEKIN BINTI CHE ROSLAN

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

2022

WEB-BASED ARABIC SPEECH RECOGNITION SYSTEM

NURUL ASHIEKIN BINTI CHE ROSLAN

THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF BACHELOR OF COMPUTER SCIENCE WITH HONORS (NETWORK ENGINEERING)

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

2022

UNIVERSITI MALAYSIA SABAH

NAME	:	**NURUL ASHIEKIN BINTI CHE ROSLAN**
MATRIC NUMBER	:	**BI18110183**
TITLE	:	**WEB-BASED ARABIC SPEECH RECOGNITION SYSTEM**
DEGREE	:	**COMPUTER SCIENCE WITH HONOR (NETWORK ENGINEERING)**
VIVA DATE	:	**25th JANUARY 2022**

CERTIFIED BY

Signature

SUPERVISOR

DR. SAMRY @ MOHD SHAMRIE SAININ

UNIVERSITI MALAYSIA SABAH

DECLARATION

I, hereby declare that the materials in this project report is my own except for quotations, equations, summaries, and references, which have been duly acknowledged.

21 JANUARY 2022

……………*ashiekin*………………

NURUL ASHIEKIN BINTI CHE ROSLAN

BI18110183

ACKNOWLEDGEMENT

Firstly, thanks to you to Allah SWT for giving me chance and make my work ease in many ways. I am highly indebted to my supervisor Dr. Samry @Mohd Shamrie Sainin for their guidance and constant supervision as well as for providing necessary information regarding the project and for their support in completing the project.

I would like to express my gratitude towards my family for their kind co-operation and encouragement which help me in the completion of this project. I would like to express my special gratitude and thanks to speakers for giving me such attention and time. My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities. Lastly, I want thanks to myself because manage to finish this project even with tears.

NURUL ASHIEKIN BINTI CHE ROSLAN

21 JANUARY 2022

ABSTRACT

Speech recognition, also known as automated speech recognition (ASR), computer speech recognition, or speech-to-text, is a feature that allows a computer software to convert human speech into written text. ASR is a study of how computers can accept voice information from humans and translate it with the greatest probability of accuracy. Capturing and digitizing sound waves, translating them to simple language units or phonemes, creating vocabulary from phonemes, and contextually interpreting the word are all part of speech recognition. Speech recognition poses some interesting challenges such as varying acoustic conditions, dialects, and articulation at word's boundaries. Students may find it difficult to learn a new language, particularly the communication aspect, because pronunciation accuracy can be challenging. The suggested project is aimed to the students at University Malaysia Sabah who are currently studying Arabic language at the beginner level. The objectives of this project are to investigate the Arabic language speech recognition for beginner using Mel-Frequency Cepstrum Coefficient (MFCC) and Artificial Neural Network (ANN), to develop the web-based application for Arabic speech recognition using the Python and PHP and to evaluate the Arabic speech recognition and functionality of the web- based system. The speech samples that will be used in this project from the expert or someone can speak fluently in Arabic. The feature extraction, the Mel-Frequency Cepstrum Coefficient (MFCC) is used to improved detection accuracy. In this project, the implementation of Artificial Neural Network (ANN) is explored as a machine learning model in this project using a python. The project will be performed in waterfall model with the phases such as requirements, design, implementation, testing and operation. This project's result is occasionally unable to forecast the words uttered but still provides accuracy. Findings for this project is MFCC is that it is good in error reduction and able to produce a robust feature when the signal is affected by noise and ANN is a suitable classifier to use. However, a larger dataset is needed to get accurate prediction.

Keywords: Speech recognition, Machine Learning, Mel-Frequency Cepstrum Coefficient, Artificial Neural Network, Arabic language, Python, PHP

ABSTRAK

Pengecaman pertuturan, juga dikenali sebagai pengecaman pertuturan automatik (ASR), pengecaman pertuturan komputer atau pertuturan ke teks, ialah ciri yang membolehkan perisian komputer menukar pertuturan manusia kepada teks bertulis. ASR ialah kajian tentang bagaimana komputer boleh menerima maklumat suara daripada manusia dan menterjemahkannya dengan kebarangkalian ketepatan yang paling besar. Menangkap dan mendigitalkan gelombang bunyi, menterjemahkannya kepada unit atau fonem bahasa mudah, mencipta perbendaharaan kata daripada fonem, dan mentafsir perkataan secara kontekstual adalah sebahagian daripada pengecaman pertuturan. Pengecaman pertuturan menimbulkan beberapa cabaran menarik seperti keadaan akustik yang berbeza-beza, dialek dan artikulasi pada sempadan perkataan. Pelajar mungkin mendapati sukar untuk mempelajari bahasa baharu, terutamanya aspek komunikasi, kerana ketepatan sebutan boleh mencabar. Projek yang dicadangkan ini disasarkan kepada pelajar Universiti Malaysia Sabah yang kini sedang mempelajari bahasa Arab di peringkat permulaan. Objektif projek ini adalah untuk menyiasat pengecaman pertuturan bahasa Arab untuk pemula menggunakan Mel-Frequency Cepstrum Coefficient (MFCC) dan Artificial Neural Network (ANN), untuk membangunkan aplikasi berasaskan web untuk pengecaman pertuturan Arab menggunakan Python dan PHP dan untuk menilai pengecaman pertuturan Arab dan kefungsian sistem berasaskan web. Contoh ucapan yang akan digunakan dalam projek ini daripada pakar atau seseorang yang boleh bertutur dalam bahasa Arab dengan fasih. Pengekstrakan ciri, Mel-Frequency Cepstrum Coefficient (MFCC) digunakan untuk meningkatkan ketepatan pengesanan. Dalam projek ini, pelaksanaan Rangkaian Neural Buatan (ANN) diterokai sebagai model pembelajaran mesin dalam projek ini menggunakan ular sawa. Projek ini akan dilaksanakan dalam model air terjun dengan fasa seperti keperluan, reka bentuk, pelaksanaan, ujian dan operasi. Hasil projek ini kadangkala tidak dapat meramalkan perkataan yang diucapkan tetapi masih memberikan ketepatan. Penemuan untuk projek ini ialah MFCC adalah baik dalam pengurangan ralat dan mampu menghasilkan ciri yang mantap apabila isyarat dipengaruhi oleh hingar dan ANN adalah pengelas yang sesuai digunakan. Walau bagaimanapun, set data yang lebih besar diperlukan untuk mendapatkan ramalan yang tepat.

***Kata kunci**: Pengecaman pertuturan, Pembelajaran Mesin, Pekali Cepstrum Frekuensi Mel, Rangkaian Neural Buatan, Bahasa Arab, Python, PHP*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Chapter Overview

This chapter describes the idea and investigates to development Web-Based Arabic Speech Recognition. It also includes its problem background, problem statement, project objectives, and project scope.

1.2 Problem Background

Speech recognizer is a application that automatically transcribes speech into text based on some finite vocabulary that restricts the words being printed out. The recognizer needs to segment the speech signals into successive phones and identify the phones corresponding to segments before transcribing the phone strings to text. Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the dictation, Interactive Voice Response, it can be used to learn a foreign language and an active area of study allowing the communication between human and machine. It is the process of understanding the human speech by a computer.

ASR is a technology which makes life easier and very promising. An effective system can replace or reduce the reliability on standard keyboard input. Attempts to build automatic speech recognition (ASR) systems were first made in the 1950s. These early speech recognition systems tried to apply a set of grammatical and syntactical rules to identify speech. If the spoken words adhered to a certain rule set, the system could recognize the words. However, human language has numerous exceptions to its own rules. The way words and phrases are spoken can be vastly altered by accents, dialects, and mannerisms.

The implementation of these kinds of systems requires a particular process for the speech signal to provide reliable features that can recognize properly the input spoken word. The basic sounds and grammatical forms in most languages are inherently identical. For anyone learning a language other than their mother tongue, however, pronunciation can be difficult. Hearing and copying sounds that are not in your native language can be difficult.

Speech recognition for language learning has proven to be a powerful tool for students learning a new language. This is how we ensure that the student is practicing proper intonation and pronunciation. This technology provides a wonderful opportunity to bring the progress of and student's individual pronunciation into training. Furthermore, it encourages each student in learning to focus more on improving their oral language in the future.

1.3 Problem Statement

Problem statement is a concise description of an issue to be addressed or a condition to be improved upon a project. In this project, Arabic is considered a 'truly foreign language', the pronunciation for the Arabic is challenging and it will be difficult for the students to really hear what they are doing wrong especially for the beginner which is student first level in the University Malaysia Sabah.

Therefore, it is so important to have an application that can identify problems in a meaningful way and that can determine individually what each student needs to work on. However, speech recognition applications provide a flexible study option. The student will be able to practice speaking when student wishes without established schedules and without the need to schedule a specific time or day with the lecturer. This allows flexibility for the student to practice speaking the language that they are learning at any time of the day.

And this project will help student to study and recognize the proper intonation and pronunciation the basic words or sentences in Arabic and translate it into English or Bahasa Melayu.

1.4 Projects Objectives

1. To investigate the Arabic language speech recognition for beginner using Mel-Frequency Cepstrum Coefficient (MFCC) and Artificial Neural Network (ANN).
2. To develop the web-based application for Arabic speech recognition using the Python and PHP.
3. To evaluate the Arabic speech recognition and functionality of the web- based system.

1.5 Project Scope

The project scope is also being the part of project planning that involves determining and documenting a list of specific project goals, deliverable, features, functions, tasks, deadlines, and ultimately costs. The project scope is as following:

- To help the beginner’s which is level 1 student in UMS learn about the basic Arabic pronunciation.
- To learn Arabic language using speech recognition method.
- Evaluation of the system functionality and its performance.

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter Overview

This chapter describes the literature review that relates to this research. Before system development started, research and study on various aspects are needed. For example, study on users' requirements gathering techniques can ensure clear and detail requirements are gathered. Besides, tools and techniques need to be studied too. It is better understanding on available choices of tools and techniques are important to make the best decision for tools and techniques to be used. Here are studies and research that have been done.

2.2 Mel-Frequency Cepstrum Coefficient (MFCC)

The mel-frequency cepstrum (MFC) is a representation of a legitimates fast- term power spectrum based entirely on linear cosine change on a non-linear mel scale of frequency in sound processing. Mel-frequency cepstral coefficients (MFCC) are coefficients that make up a Mel-frequency cepstral coefficient (MFC). They're based on a cepstral representation of the audio sample (a nonlinear "spectrum-of-aspectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that the MFFC's frequency bands are evenly separated on the Mel scale, which more closely approximates the human hearing system's response than the normal cestrum's linearly spaced occurrence bands. This frequency warp can allow for improved sound representation, such as in audio compression.

The following is a typical way to make MFCCs:

(i) Perform a Fourier transform on a signal (or a windowed snippet of a signal). Using triangular overlapping windows, map the power of the spectrum acquired on top of onto the Mel scale.

(ii) Record the power logs for each of the Mel frequencies.

(iii) As if the list of Mel log powers were a signal, compute the discrete cosine transform.

(iv) The MFCCs are the resultant spectrum's amplitudes.

Because MFCC values aren't particularly stable in the presence of additive noise, it's not surprising to standardize their values in speech popularity structures to reduce noise's influence. To improve robustness, some researchers recommend adjusting the basic MFCC algorithm, such as increasing the log-mel- amplitudes to a reasonable energy (about 2 or three) before performing the DCT (discrete cosine transform), which reduces the impact of low-power additions. Since the logarithmic outcome of all clean out financial institution energy was dealt with using discrete cosine transform mathematical equipment in MFCC, a couple of-14 out of n DCT samples or coefficient were stored while the remainder was deleted.

2.3 Artificial Neural Networks (ANN)

An artificial neural network (also known as a neural network) is an adaptive system that learns by integrating linked nodes or neurons in a layered structure like that of the human brain. A neural network may be trained to identify patterns, organize the data, and anticipate future occurrences by learning from data. A neural network divides input into layers of abstraction. It, like the human brain, can be trained using many examples to detect patterns in voice or visuals.

Its behaviour is characterised by the way its constituent pieces are linked, as well as the strength, or weights, of those connections. During training, these weights are automatically modified according to a predefined learning rule until the artificial neural network properly executes the target job. Neural networks are a form of machine learning technique that is inspired by how neurons communicate in the human brain.

Neural networks are particularly well-suited to modelling non-linear interactions, and they are commonly employed in speech, vision, and control systems to conduct pattern recognition and categorise objects or signals. Neural networks have become well-known for their ability to do complicated identification tasks such as facial recognition, text translation, and voice recognition. These techniques are a primary driver of innovation in advanced driver assistance systems and tasks such as lane categorization and traffic sign recognition. A neural network, which is based on organic nerve systems, integrates numerous processing levels by employing basic parts that operate in parallel.

The network is made up of three levels which is an input layer, one or more hidden layers, and an output layer. Each layer has numerous nodes, or neurons, and the nodes in each layer utilise the outputs of all nodes in the preceding layer as inputs, allowing all neurons to communicate with one another across several levels. Each neuron is often allocated a weight, which is altered during the learning process, and changes in the weight affect the intensity of that neuron's output.

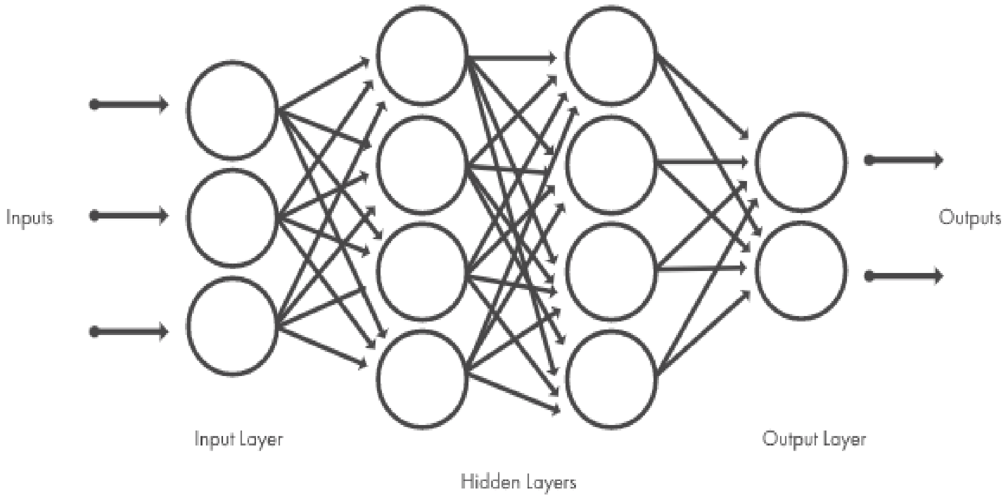

Figure 2.1: Typical neural network architecture

2.4 Related works on speech recognition

The aim of speech recognition is to enable machines to accept sounds and act based on it. Automatic speech recognition is the ability for a machine to recognize "receive and interpret" the speech and convert it into readable form or text and performing an action based on the instructions defined by the human. Authors Yu and Deng dissected the AASR system into four stages. First, pre-processing stage. Second, Feature extraction stage. Third, decoding using Acoustic model, Language model, and pronunciation dictionary. Fourth, Post-processing results where the best hypothesis is produced. The stages work as following: First, speech waveform used as input in the pre-processing stage. Then, the output is processed speech waveform, and this is used as input in feature extraction stages where we have the feature vector as output and use it as input in the next stage, the decoding stage. In this stage, the Acoustic model, is employed along with a pronunciation dictionary. After that, the n-best hypothesis is the output of the pronunciation dictionary stage is used in post-processing as input. As a result, the best hypothesis is produced from this work operation.

Turab, Khatatneh, and Odeh in discussed the phoneme recognition as it is related to speech recognition. The techniques used are as follows Gaussian Low Pass filtering algorithm along with the neural network in the pre-processing stage to have an improvement on the results. Furthermore, the stages of phoneme recognition are catching a signal, sampling, quantization and setting energy. After that, a neural network is used to enhance the results. Moreover, this paper shows the enhanced impact in results after applying the Gaussian Low Pass filter in voice signals hence, the noise was reduced. After that, in the training phase, the neural network has been used to train the system to recognize the speech signals.

Ahmed and Ghabayen proposed three approached to enhance the AASR. The paper started with the first approach which is the punctuation modelling, in this approach Ahmed and Ghabyaen proposed a decision tree with variant pronunciation generation. After that, a hybrid approach proposed and used to adapt the native acoustic model with another native acoustic model. Finally, the language model is enhanced and improved using a processed text. The model efficiency was measured by Word Error Rate (WER) which is a metric to measure the performance of speech recognition and calculates misrecognitions at the word level. Consequently, the pronunciation model reduced WER by 1%, The acoustic modelling reduced the WER by 1.2% and the language model reduced WER by 1.9%.

Emami and Mangu, examine the neural network usage for Arabic speech recognition using a distributed word representation. Furthermore, the model of the neural network allows robust generalization and enhance the ability to fight the data sparseness problem. Also, the investigation process includes different configuration neural probabilistic model, n-gram order parameter experiment, output vocabulary, the method of normalization, model size and parameters. The experiment has been done on the Arabic news broadcast, and conversation broadcast. As a result, some improvement has been achieved using the optimized neural network model over the 4-gram baseline model resulting in up to 0.8% absolute reductions and 3.8% relative WER. However, different parameters do not have a significant impact on model performance. The paper was based on analysing first. Then, feature extraction. After that, modelling and finally, testing.

Based on Desai, Dhameliya, and Desai, the proposed speech recognition system contains four stages. First, feature extraction. Second, database. Third, network training. Fourth, testing or decoding. Anzi and AbuZeina used WER metric to evaluate the performance of isolated-word recognition and continuous speech recognition. The evaluation of continuous speech was presented for seven papers based on the improvement of WER. The results were as follow Kirchhof, Bilmes and Stolcke in performed performance evaluation using a language model for morphology and the improvement of WER for two different test sets is 1.8% and 1.5% respectively. Emami, Ahmad and Lidia use two different configurations of neural probabilistic models and the improvement of WER is 0.8% and 3.8% respectively. The authors used broadcast news corpus and improved the WER by 13.66%. Hyassat and Abu Zitar used the holy Quran corpus and WER improved by 46.182%. Elmahdy and Mohamed used Egyptian Colloquial Arabic and reached 99.34% of recognition accuracy. Selouani, Sid Ahmed and Malika Boudraa used MSA continues speech corpus and reach an accuracy rate of 91.65%. The authors Jurafsky and Martin used MSA continues speech corpus and the improvement of WER using diacritical marks and without resulting 11.27% and 10.07% respectively

2.5 Conclusion

Based on the review in chapter 2, speech recognition system that has many methods to approach to improve system. And the neural network is one of the famous classifiers which useful by remembers each information through time.

CHAPTER 3

METHODOLOGY

3.1 Chapter Overview

Chapter 3 contains the descriptions overall methodology to develop this system. This chapter describes the method and technique used to develop this system. Other than that, it also provides the hardware and software requirement used to run the project.

3.2 Project Development Methodology

To develop FYP project successfully, a methodology will be chosen which is waterfall development. Waterfall will be used as the main methodology to develop the system and prototype will be used to support implementation phase. Critical issue of the waterfall development that always been discussed which is the developed system may not fulfil users' requirements. Therefore, waterfall development that does not review users' requirements from time to time may not realize the changes. Waterfall development with the virtue of consistency and completeness of documentation are the main reasons why it has been chosen as the main methodology in this system development. Step-by-step approach of waterfall development has provided clear guidelines to developers to develop a complete functional system. Moreover, completeness of documentations that produce in waterfall development can help further upgrading tasks and study of this system. Lastly, MFCC will be use as a feature extraction and ANN as a classifier in this project.

3.2.1 Phases In Waterfall Model

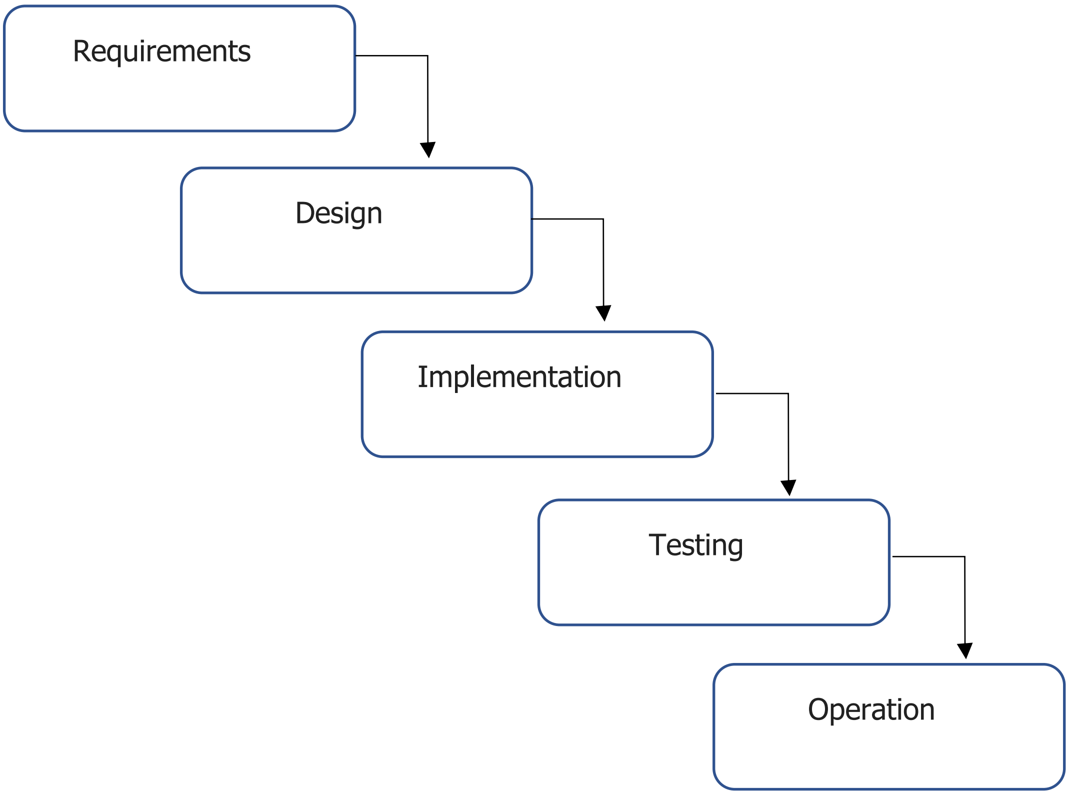

Figure 3.1: Waterfall Model

3.2.2 Requirements Phase

1. System: all requirements of the product/system done through interview, observation, etc. functional, usability, technical (software, platform, database).
2. Research embedded: all requirements of the proposed research element – respondent, data, algorithm, software, platform.

3.2.3 Design Phase

1. System: software architecture, use case, context diagram, DFD, Data dictionary, user interface.
2. Research embedded: Data sampling, architecture, parameters, training, and testing.

3.2.4 Implementation Phase

1. Research embedded: Model generation (best model from training and testing)
2. System: coding and integration with (learning model).

3.2.5 Testing Phase

1. System: functionality, unit, usability.
2. Research embedded: performance(recognition/classification) of the integrated learning model.

3.2.6 Operation Phase

1. System: operation of the system (e.g., live server)
2. Documentation: report.

3.3 MFCC Feature Extraction

MFCC feature extraction is a method that is widely used in speech recognition. Figure below flowchart of MFCC feature extraction.

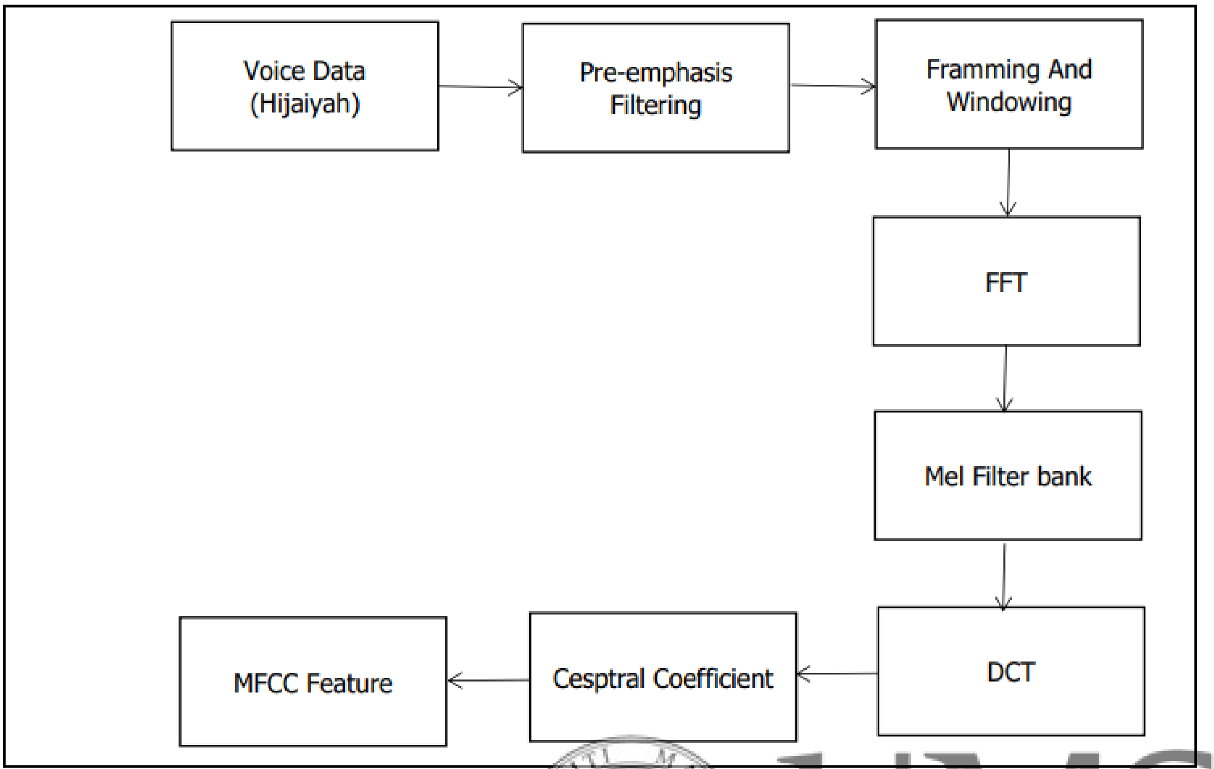

Figure 3.2: MFCC feature extractio

(i) **Pre-emphasis Filtering**

Pre-emphasis filtering was applied to increase the high frequency energy and decrease the low frequency energy.

(ii) **Framming and windowing**

Speech signal is nonstationary, but usually the speech signal has stationary at a certain time range (20-40 ms), namely short windows or frame. In the framing process speech signal will be divided into several frames and then processed.

(iii) **Fast Fourier Transform (FFT)**

At this stage, each frame was converted from time domain form into frequency domain form. FFT is a computational algorithm of Discrete Fourier Transform (DFT).

(iv) **Mel Filter bank**

At this stage bank-filter analysis was used to perform linear predictions.

(v) **Descrete Cosine Transform (DCT)**

To bring back the signal in time domain form, then at this stage is done by Descrete Cosine Transform (DCT). DCT transforms the cosine component only.

(vi) **Cepstral Coefficient**

The result for Cepstral Coefficient that will be used as a feature in speech recognition.