WEB-BASED SPEECH RECOGNITION SYSTEM FOR KADAZAN LANGUAGE

ARFIVEYINA ARMIDALE

FACULTY OF COMPUTING AND INFORMATICS

UNIVERSITI MALAYSIA SABAH

2022

UMS
UNIVERSITI MALAYSIA SABAH

WEB-BASED SPEECH RECOGNITION SYSTEM FOR KADAZAN LANGUAGE

ARFIVEYINA ARMIDALE

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

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

2022

NAME : ARFIVEYINA ARMIDALE

MATRIC NUMBER : BI18110126

TITLE : WEB-BASED SPEECH RECOGNITION SYSTEM FOR KADAZAN LANGUAGE

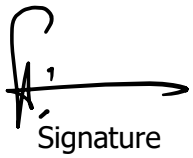
DEGREE : BACHELOR OF COMPUTER WITH HONOURS (NETWORK ENGINEERING)

VIVA'S DATE : 25 JANUARY 2022

CERTIFIED BY;

1. SUPERVISOR

DR. SAMRY @ MOHD SHAMRIE SAININ



Signature

DECLARATION

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

24 JANUARY 2022

ARFIVEYINA ARMIDALE

BI18110126

ACKNOWLEDGEMENT

First and foremost, I would want to praise and thank God, the Almighty, who has blessed me with numerous knowledge, and opportunities, that allow me to finally complete my thesis. I owe a great debt of gratitude to my supervisor, Dr. Shamrie Bin Sainin, for his advice and regular supervision, as well as for providing vital project information and for his assistance in finishing the project. Also, thanks to my examiner, Prof. Dr. Jason Teo Tze Wi, and my panel, Dr. Salmah Binti Fattah for giving suggestions and good feedback that help me improve my project.

Aside from that, I'd want to express my gratitude to my loving family, who have prayed for me, guided me, and continue to encourage me to complete this Final Year Project report. Not to mention my greatest friends who have encouraged me to stay tough and never give up on completing this project report.

ARFIVEYINA ARMIDALE

BI18110126

ABSTRACT

Because Kadazan speech contains unique traits not seen in other languages, there is currently no system that provides common information and tools for Kadazan speech recognition. In this study, the implementation of Mel Frequency Cepstral Coefficient (MFCC) and Neural Network is explored as a method for the proposed system which is a Web-Based speech Recognition System for Kadazan Language. The objectives of this project are 1) to prepare the requirement and analysis for the Kadazan language speech recognition web-based system, 2) to develop the web-based application for the Kadazan language speech recognition system, and 3) to evaluate Kadazan language speech recognition and functionality of the web-based application. The prototype contains 10 keywords that are used to decide how the users pronounce each of the keywords. The speech recognition technology is incorporated into a web-based system using PHP and Python after extraction, training, and test of the data complete, to create the working implementation that can detect the user's pronunciation accuracy. The output from this project is the system is sometimes unable to predict the words spoken but still gives accuracy. The finding in this project is the MFCC and Neural Network are good feature extraction and classifier. However, to approach the limitation in this project, different feature extraction approaches and the study of additional classifiers, as well as researching by training the model with a larger dataset and using word phonemes are needed.

ABSTRAK

Oleh kerana pertuturan Kadazan mengandungi ciri unik yang tidak dilihat dalam bahasa lain, pada masa ini tiada sistem yang menyediakan maklumat dan alat biasa untuk pengecaman pertuturan Kadazan. Dalam kajian ini, pelaksanaan Mel Frequency Cepstral Coefficient (MFCC) dan Neural Network dikupas sebagai kaedah untuk sistem yang dicadangkan iaitu Sistem Pengecaman Pertuturan Berasaskan Web bagi Bahasa Kadazan. Objektif projek ini adalah 1) untuk menyediakan keperluan dan analisis untuk sistem berasaskan web pengecaman pertuturan bahasa Kadazan, 2) untuk membangunkan aplikasi berasaskan web untuk sistem pengecaman pertuturan bahasa Kadazan, dan 3) untuk menilai pertuturan bahasa Kadazan pengiktirafan dan kefungsian aplikasi berasaskan web. Prototaip mengandungi 10 kata kunci yang digunakan untuk menentukan cara pengguna menyebut setiap kata kunci. Teknologi pengecaman pertuturan digabungkan ke dalam sistem berasaskan web menggunakan PHP dan Python selepas pengekstrakan, latihan dan ujian data selesai, untuk mencipta pelaksanaan kerja yang boleh mengesan ketepatan sebutan pengguna. Output daripada projek ini ialah sistem kadangkala tidak dapat meramal perkataan yang diucapkan tetapi masih memberikan ketepatan. Penemuan dalam projek ini ialah MFCC dan Rangkaian Neural adalah pengekstrakan dan pengelas ciri yang baik. Walau bagaimanapun, untuk mendekati had dalam projek ini, pendekatan pengekstrakan ciri yang berbeza dan kajian pengelas tambahan, serta penyelidikan dengan melatih model dengan set data yang lebih besar dan menggunakan fonem perkataan diperlukan.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The Kadazan people of Borneo speak the Kadazan language, which starts in the Nosoob-Kepayan area and continues through Penampang-Putatan and Papar, Sabah. Kadazan has its characteristics, grammatical structures, and rules, much like every other language. Language changes in multi-ethnic speech cultures may result in the mother tongue no longer being spoken within the group itself. In the family realm, Sabah local Malay and/or English have taken over some of Kadazan's communicative roles. Second, the degree to which members of the speech group have modified their habitual language usage is based on several factors including schooling, intermarriage, reputation, migration, and language touch. These aspects hasten the transition away from Kadazan and toward other languages. The speech recognition technology can be a tool to help speed up and sustain the speech process of native speakers Kadazan language.

Speech recognition is interdisciplinary computer science and computer linguistics subfield that advances methodologies and technology that make it possible for computers to understand and translate spoken language into text. Over the years with the recent emergence of technology, the growing contact between humans and machines or automated systems has made it an important and integral part of our lifestyle. The theory of speech recognition originated back in the early 1950s and it was all about activation and recognition of voices. The National Science Foundation later sponsored the experiment during the 1980s and used the technology on individuals with disabilities and victims of sclerosis and cerebral paralysis. Speech

recognition programs were very complicated and constrained during this period. Users will use tiny vocabulary or any person can use it. There was a significant increase in speech recognition by the end of the 1990s and it was extended to every possible sector. Most of the computers, gadgets, and cars today contain. Facility for voice recognition. For their smartphones, famous businesses such as Apple, Microsoft, and Google have their speech assistants.

1.1 Problem To Be Solved

In my research, the student who studies the Kadazan language will face problems with pronunciation, most of them when the student is not a local from Sabah. By observing that a speech recognition system will be a good idea for guiding student work.

1.2 Problem Statement

For problem statement, in the Malaysian educational system, the Kadazan language has several objectives, especially in Sabah. Among them, the capacity of students to interact well with the subject to appreciate and practice values, positive attitudes, and patriotism, and to love the country, in addition to continuing the cultural tradition of Kadazan in line with Malaysian education (Norjietta Taisin, 2012). Thus, to help students learn the Kadazan language much better other than help from a qualified teacher is by using some system like speech recognition that will help the student to work on their Kadazan pronunciation.

1.3 Objectives

- to prepare the requirement and analysis for kadazan language speech recognition web-based system.
- to develop the web-based application for the Kadazan language speech recognition system.
- to evaluate Kadazan language speech recognition and functionality of the web-based application.

1.4 Project Scope

This project was developed to support UMS students who are studying Kadazan in Language Learning (PPIB) at Universiti Malaysia Sabah. This project would have several distinct keywords that will be used to predict the speaker's pronunciation. The distinct keywords will be in the greeting category.

1.5 Conclusion

This chapter are introduced about the system background. Including the problem statement and problem to solve, objectives to achieve and scope of the project.

CHAPTER 2

LITERATURE REVIEW

2.1 Speech Recognition

Speech recognition is the mechanism by which spoken words are identified by a computer. That involves talking to the machine and making it know what you're saying properly. This is the secret to every program concerned with expression. There are a variety of ways of achieving this, as will be discussed below, but the basic idea is to somehow remove those core characteristics from the spoken expression and then treat certain characteristics as the key to remembering the word when it is uttered again. "The purpose of speech recognition is for a computer to be able to "hear," comprehend," and "act on" spoken data. At Bell Labs, the first speech recognition devices were first tried in the early 1950s and have been taught to recognize digits. Any of the most commonly used voice recognition devices are speech recognition systems. All of the systems used by speakers, the autonomous system of speakers, the system of independent word recognition, the system of linked word recognition, and the system of random recognition. An autonomous digit recognition device for a single speaker has been developed by Davis, Biddulph, and Balashek. The first step of a speech recognition system is used to extract features from a word's voice signal, while the second stage is used to classify patterns. The extraction of features is a critical stage in the system. The system's recognition rate is determined by the meaningful data that characterizes the derived features from the voice stream.

2.2 Feature Extraction

Feature Extraction is the most important aspect of speech recognition, as distinguishing one speech from another plays an important role. A lot of research has been done in this area in the last several years. A broad variety of feature extraction methods recommended and successfully used for voice recognition tasks can be derived from the utterance. For each speech signal, feature extraction methods typically generate a multidimensional feature vector. A broad range of options is available to parametrically reflect speech signals for the recognition process, such as linear prediction cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC). The present focus of identification system research is on addressing critical issues such as adaptability, complexity, multilingual recognition, and noise resistance. According to the findings, MFCC-based techniques have been used more than any other method.

2.3 Mel-Frequency Cepstral Coefficients (MFCC)

Based on research by Nidhi Desai. *et. al.,* 2013, the most well-known and widely used feature extraction technique for speech recognition is MFCC. The MFCC attempts to replicate the human ear's nonlinear resolution of frequencies around the audio spectrum. As a result, the Mel filters distort the frequency such that it obeys the spatial relationship of the human ear's hair cell distribution. Since the frequency bands in the MFCC are located logarithmically, it approximates the reaction of the human system more accurately than any other system. The MFCC computation technique is based on short-term processing, so an MFCC vector is computed from each window. To derive the coefficients, the speech sample is taken as the input and the Hamming window is used to minimize the disruption of the signal. David and Mermelstein introduced Mel-frequency Cepstral Coefficients (MFCC), which later became the most common for SR systems. It is a good feature vector for representing human speech or any audio signal, according to Winursito A., et al., 2018. The MFCC coefficients are obtained by a six-step process.

i. Pre-emphasis

Filtering at higher frequencies of the speech signal is known as pre-emphasis, and it has been used in a variety of speech processing applications such as speech recognition. Pre-emphasis is used to balance the spectrum of spoken sounds, which often have a rapid roll-off at higher frequencies. The glottal source has a slope of roughly -12dB/octave for the sound of the voice. When compared to an acoustic energy source that vibrates from lips, the glottal source provides a +6dB/octave augmentation to the spectrum. Taking these factors into consideration, a human recording of a voice signal using a microphone will result in a -6dB/octave slant downward relative to the genuine vocal tract spectrum.

ii. Framing and windowing

Because the voice signal is concerned for stability, a stable signal can be assumed for 10-30ms. Framing's main purpose is to convert a long-period voice signal into a short-period speech signal with generally constant frequency characteristics. The word frame rate refers to the process of extracting a feature every 10 milliseconds. Windowing, on the other hand, mostly reduces aliasing. This happens when a long signal is converted to a short-time signal in the frequency domain.

iii. Fast Fourier Transform (FFT)

Fast Fourier transform (FFT) is required to convert the time domain into frequency domain from each N frame of samples. In simple terms, FFT converts the combination of the vocal tract impulse response H[n] and the glottal pulse U[n] into the time domain.

iv. Mel Filter Bank Processing

The Mel spectrum is computed by passing the Fourier transformed signal through the Mel-filter bank, commonly known as a series of band-pass filters. A Mel is a frequency measuring unit based on the human ear. FFT has a wide frequency range, and voice signals are not linear. Below 1 kHz, the Mel scale has a logarithmic frequency spacing, while beyond 1 kHz, it has a linear frequency spacing.

v. Discrete Cosine Transform

Discrete Cosine Transform is used to convert the log Mel spectrum into a time-domain (DCT). Mel Frequency Cepstral Coefficient is the name given to the output once it has been transformed. The set of coefficients is known as an acoustic vector. Each input syllable is transformed into a sequence of audio vectors as a result of this.

vi. Delta Energy and Delta Spectrum

While the slope of formant transitions, the frames and speech signal change. As a result, there is a strong desire to include elements that relate to the evolution of cepstral traits throughout time.

2.4 Classification

Another essential feature of a speech recognition system is classification, as the patterns are categorized into various groups. According to Michelle Cutajar. *et. al.,* 2013, there are two ways to handle the classification level. The first technique is known as the generative approach, and it involves identifying the joint probability distribution with given observations and class labels. The discriminative method, on the other hand, uses a parametric model to find the conditional distribution, with the parameters being identified using a training set of pairs of input vectors and their corresponding target output vectors. A multi-class grouping challenge is voice comprehension. As a consequence, the function vector collection obtained is categorized using three multi-class classifiers: Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Naive Bayes classifiers. Data is split into training and evaluation sets of nearly all classification systems. Each instance in the training set has a goal value that represents the class it belongs to, as well as a set of attributes. No target value is detected in the test results. The goal of the classifier is to construct a training data model that predicts the target values of the test data. Neural networks have recently risen to prominence in a variety of machine learning domains. One of them is natural language processing. For disambiguating morphological forms, two conventional classifiers are extensively used. Deep neural networks, which are artificial neural networks with direct propagation, are commonly used in acoustic models. The backpropagation method is used for training.

2.5 Related Works

Noise, speaker variability, language variability, vocabulary size, and domain are all factors that affect the accuracy of speech recognition systems. Some studies on speech recognition have recently been conducted, with various methodologies and algorithms being employed. These are Mel Frequency Cepstral Coefficients (MFCC) and a vector quantization algorithm, Discrete Wavelet Transform (DWT) & Linear Predictive Coding (LPC), Hidden Markov Model (HMM) and Linear Predictive Coding (LPC), Hidden Markov Model and Neural Networks, Fast Fourier Transform (FFT) and Fuzzy matching method, LPC and Euclidean Squared Distance, LPC and Artificial Neural Networks.

According to Hussein M. Mohammed et. Al., 2018, speech recognition is tested using three different methods of feature extraction which are MFCC and LPC. These features are used as an input to the neural network in the classification. They use 30 spoken words recorded by their author's voice using a headphone set. Those 30 words were used to train the neural network. There are two groupings of 30 words each. Every word was twice recorded. The feature extraction methods (LPC, and MFCC) are applied, and then pattern matching is done with the neural network. The neural network was built with a variety of hidden layers. In comparison to the other two approaches, the LPC approach produces fewer output data (420 neurons in the network's input layer) and (613 neurons for the MFCC method). The best recognition rate of 83.3 percent was attained with the MFCC approach, according to simulation findings.

To increase the accuracy of audio categorization or voice recognition, neural networks play a very important role. A study was conducted to see how well a convolutional neural network (CNN) could recognize command words invoice. To improve accuracy, they used CNN instead of the traditional deep neural network. In their research, they used a deep neural network (DNN), a recurrent neural network (RNN), and a CNN, and the results showed that utilizing CNN for command word speech recognition can enhance accuracy because CNN has the greatest accuracy rate compared to the other neural networks. CNN's accuracy percentage for the testing results was 92.88 percent in that study (X. Yang, et al., 2020).

A speech recognition translator from English to Indonesian utilizing HMM has been investigated and developed in a study. The words were separated into 9 groups and 87 English words were assessed. Each group has 9-10 recordings of words. The goal of grouping was to determine the average accuracy of each group. As a result of utilizing HMM, they were able to achieve an average accuracy of 70.10 percent (Winursito A., et al., 2018).

More tests and investigations have been done to help language learners improve their pronunciation, as seen in these studies and researches. One thing these studies should do is integrate their method into a web-based application so that students may use it more easily.

2.6 Conclusion

Based on the review in chapter 2, Speech recognition is the mechanism by which spoken words are identified by a computer. At Bell Labs, the first speech recognition devices were first tried in the early 1950s. MFCC is one of the famous methods to do an extraction of voice and Neural Network as classifier. The review on the existing system giving some guideline in developing this web-based speech recognition system.

CHAPTER 3

METHODOLOGY

3.1 Overview

The speech samples for this project come from people who can understand Kadazan or who are currently studying the language. The prototype contain a variety of keywords that used to decide how the speakers pronounce terms. The speech recognition technology then be incorporated into a web-based system using PHP and Python to create a working implementation that can detect the speakers' pronunciation accuracy.

There are a variety of speech recognition techniques available, including Dynamic Time Warping (DTW), Hidden Markov Model (HMM), and others. The backpropagation algorithm was used for a feed-forward neural network. Mel Frequency Cepstral Coefficients (MFCC) were used to remove speech features, resulting in a series of speech waveform feature vectors. MFCC has previously been found to be more reliable and efficient in speech recognition than other feature extraction techniques also will be used in this project.

3.2 System Development

Python is used to perform a series of speech recognition operations, including data selection, feature extraction, model creation, and analysis. The speech recognition software combined with PHP to build a web-based prototype for Kadazan speech recognition after the speech recognition processes are completed. This project use machine learning methods, which ensures that data be collected and