WEB-BASED KADAZANDUSUN SPEECH RECOGNITION SYSTEM

ARRVIERRI VESPHA

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2022

NAME	**: ARRVIERRI VESPHA**
MATRIC NUMBER	**: BI18110185**
TITLE	**: WEB-BASED KADAZANDUSUN SPEECH RECOGNITIONSYSTEM**
DEGREE	**: BACHELOR OF COMPUTER SCIENCE WITH HONOURSHONOURS (NETWORK ENGINEERING)**
VIVA'S DATE	**: 21 JANUARY 2022**

CERTIFIED BY,

1. SUPERVISOR
Dr. Samry @ Mohd. Shamrie Bin Sainin **SIGNATURE**

DECLARATION

I declare that the materials in this thesis are original excerpts for quotations, excerpts, summaries, and references, which have been acknowledged.

21 JANUARY 2022

ARRVIERRI VESPHA

BI18110185

ACKNOWLEDGEMENT

I would like to acknowledge and give my thanks to my supervisor (Mohd Shamrie Sainin) who made this work possible. His guidance and advice had carried me through all phases of this project.

I would also like to thank you my examiners and panel (Ervin Gubin Moung & Norazlina Binti Khamis) that gave me a lot of insights and feedback to give room for this project to keep improving.

A special thanks to my family for their continuous support and understanding when undertaking my research and writing my project.

Besides, I am also thankful to my friends that always been there to help and support me in any way possible.

Lastly, I want to send my regards and blessing to all of personnel who supported me for completing my project.

ARRVIERRI VESPHA

21 JANUARY 2022

ABSTRACT

This project investigates a model for the classification and develops the web-based system of speech recognition for the Kadazandusun language. It is specifically designed for Universiti Malaysia Sabah (UMS) students and lecturers that are studying or teaching the Kadazandusun language. Nowadays, only the elderly of the Kadazandusun ethnic are fluent in their language and the majority of the ethnic's youth cannot speak it fluently yet still understand their language, while some cannot understand the language spoken. This caused the existential crisis of the Kadazandusun language to arise. Furthermore, there are not many speech recognition systems that were developed for the language itself. There is only a little information of research regarding Kadazandusun speech recognition. The main purposes of this project are to investigate, recognize, and evaluate the Kadazandusun language speech recognition based on Mel-frequency Cepstral Coefficient (MFCC), Feed Forward Neural Network, and Principle Component Analysis. A website application is developed which integrates the model for speech recognition system. Based on the investigation (training and testing), the MFCC and Neural Network produce 89.22, 86.82, 87.54, 85.92, and 85.61 mean for classification accuracies using 11, 12, 13, 14, and 15 of MFCC coefficients respectively. Coefficient 11 was chosen as MFCC coefficient therefore it can be used as basic speech recognition for the Kadazandusun. Future works include collecting more data to enhance the usability of admin page, improving the pre-processing method, and hyperparameter tuning, as well as adding new feature to this system.

Keywords – classification, Kadazandusun, speech recognition, Mel-Frequency Cepstral Coefficient (MFCC), hyeperparameter, and Feed Forward Neural Network.

ABSTRAK

Projek ini menyiasat model untuk klasifikasi dan membangunkan sistem pengecaman pertuturan berasaskan web untuk bahasa Kadazandusun. Ia direka khusus untuk pelajar dan pensyarah Universiti Malaysia Sabah (UMS) yang sedang belajar atau mengajar bahasa Kadazandusun. Pada masa kini, hanya warga tua etnik Kadazandusun yang fasih berbahasa mereka dan majoriti belia etnik itu tidak boleh bercakap dengan lancar namun masih memahami bahasa mereka, manakala ada yang tidak memahami bahasa yang dituturkan. Ini menyebabkan krisis eksistensial bahasa Kadazandusun timbul. Tambahan pula, tidak banyak sistem pengecaman pertuturan yang dibangunkan untuk bahasa itu sendiri. Hanya terdapat sedikit maklumat penyelidikan berkenaan pengecaman pertuturan Kadazandusun. Tujuan utama projek ini adalah untuk menyiasat, mengenali dan menilai pengecaman pertuturan bahasa Kadazandusun berdasarkan Pekali Cepstral Frekuensi Mel (MFCC), Rangkaian Neural Feed Forward dan Analisis Komponen Prinsip. Aplikasi laman web dibangunkan yang mengintegrasikan model untuk sistem pengecaman pertuturan. Berdasarkan penyiasatan (latihan dan ujian), MFCC dan Rangkaian Neural menghasilkan 89.22, 86.82, 87.54, 85.92, dan 85.61 min untuk ketepatan pengelasan masing-masing menggunakan 11, 12, 13, 14, dan15 pekali MFCC. Pekali 11 dipilih sebagai pekali MFCC oleh itu ia boleh digunakan sebagai pengecaman pertuturan asas untuk Kadazandusun. Kerja masa hadapan termasuk mengumpul lebih banyak data untuk meningkatkan kebolehgunaan halaman pentadbir, menambah baik kaedah pra-pemprosesan dan penalaan hiperparameter, serta menambah ciribaharu pada sistem ini.

Kata kunci – klasifikasi, Kadazandusun, pengecaman pertuturan, Mel-Frequency Cepstral Coefficient (MFCC), hyperparameter, dan Rangkaian Neural Feed Forward.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Kadazan, also known as Dusun or Kadazan Dusun, is a term that refers to a group of people who together make up the largest indigenous ethnic group in the state of Sabah, Malaysia. These people are located on the island of the northeastern tip of Borneo (Britannica). It has a total of 40 ethnic altogether (KDCA). Usually, the Kadazan live in the hills around Tambunan and along the coastal plain from Kudat to Beaufort. They speak Kadazan (or Kadazandusun) which is an Austronesian language with a variety of dialects.

Following up the four languages that could be taught in school before 1996, which are Malay, Chinese, Tamil, and Iban (Dayak languages originated from Sawarak), the Malaysian Government had finally agreed to add Kadazandusun language as the fifth one that also can be taught in school at the end of 1995 (A. Reid, 1997). This leads to an opportunity for people who want to learn how to speak this language.

For some languages such as English, Japan, Spanish, and Arabic native speakers, the language itself is the main element of the identity for each of the native speakers. So, apart from learning the language either for a school-based education or self-learning (native speaker or non-native speaker), this language is also important to keep the identity of the Kadazan people.

The reason why this research project was conducted in the first place, is to find a solution on how to save the Kadazandusun language so that it would not face an extinction issue.

1.1.1 Problem background

Speech recognition is the ability of a computer or machine to recognize words and phrases in any language and convert them to a machine-readable format (Abdou, et al., 2018). The most basic way of communication among human beings is by speech or voice. Researchers have been actively trying to make computers understand any speech made by humans. With the continuously advancing of the neural network fields, a series of algorithms used for recognizing relationships in a given set of data by using the human brain operation as a model (F. Reyes, 2020), has been continuously addressing difficulties relative to speech recognition.

Despite the amount of speech recognition made throughout the years, a speech recognition system for the Kadazandusun language is yet fully explored. Kadazan-Dusun is the largest ethnic group in Sabah. Their language is one of their uniqueness but as the new generation coming, this language is not being spoken frequently in their daily life especially for the youngsters (R. Gontusan, 2017). Hence, this creates a possibility where this languagecould disappear in the future if the issues where the majority only the elderly know how to speak their language could not be solved. One way to keep Kadazandusun language from their existential crisis is by using the knowledge of machines and artificial intelligence. This project will be focusing on the Kadazandusun speech recognition system that provides an alternative language learning, not only for the Kadazandusun people but also for anyone interested in learning this language.

1.1.2 Problem Statements

Parents of KadazanDusun children born in the 1950s began to encourage their children to learn English or Malay, presumably believing that their children could learn their mother tongues at home, in the hopes of giving their children an advantage in securing jobs in both the government and private sectors (R. Gontusan, 2017). This gave a bad impact on the children and the new generation in regards to not be proficient in their mother tongue. The kadazandusun language continuously in a state of danger especially when the "Sabah Malay" slang gradually replacing the Kadazandusun language. The use of other languages contributing to the decreasing use of Kadazandusun language. As stated in Daily express, themajority of current speakers of Kadazandusun language are fifty years old and above category.

Culture and traditions go hand-in-hand with their unique language, and the languageis what glues them together which applied to all cultures and traditions in this world. That

goes without being said that, the existential of language can cause the race to vanish or to be destroyed. The same goes to Kadazandusun people which is at a high risk of having their race being forgotten by other race and by their people if this issue is not being solved givingthe fact that in an earlier 2005 the Kadazandusun language was already categorized under endangered language by Unesco (R. Gontusan, 2017).

The hope is still there for the Kadazandusun language as the Kadazandusun Language Foundation was established in 1995 to preserve, develop, and promote the language. The foundation's effort did not go to waste as the proposal that was made by the Kadazan Dusun Cultural Association (KDCA) to teach Kadazandusun language in the school at Sabah had been approved by the government. This gave results to 336 primary and 38 secondary schools in Sabah that include the Kadazandusun language for one of their subject. However, this

language will still have its uphill challenges to protect them.

In this project, the desired goal is to help preserve the language using technology. This project is not big but it could help in bringing the idea of saving the endangered language by making good use of the modern solutions. It is undeniable that there is a very limited system that was developed for Kadazandusun language speech recognition.

It is important to solve this issue as the world is moving fast day by day, younger generation replacing the older generation which will create the possibility for another shift of language. The Kadazandusun language can easily be overshadowed by other popular languages such as Korean, Japanese, Spanish, and more.

On the other hand, resolving this problem can help for a better understanding of thisproject as it will involve the use of Mel-frequency Cepstral Coefficient (MFCC) and neural network such as Convolutional Neural Network (CNN) to develop the system using machine learning. Machine learning is the alternative to preserve the Kadazandusun language. This project can be leverage in any research that also attacking the same issue as only a little information of research regarding Kadazandusun speech recognition can be obtained. Comparison can be made with the findings of this project with another project by comparingthe accuracy obtained to maximize the output.

1.1.3 Hypothesis

1. Applying Mel Frequency Cepstral Coefficients (MFCC) for feature extraction, and using the Convolutional Neural Network (CNN) for audio classification can increase the accuracy ofthe Kadazandusun speech recognition system.

1.1.4 Project Objectives

i. To investigate the feature extractions for Kadazandusun language speech data using Mel-frequency Cepstral Coefficient (MFCC) and Feed Forward Neural Network to helpusers on their

Kadazandusun pronunciation.

ii. To develop the prototype of a web-based Kadazandusun language speech recognition with MFCC and Feed Forward Neural Network.

iii. To evaluate the Kadazandusun speech recognition.

Table 1.1 shows how the objectives of this project can help to solve the problem stated earlierin the problem statement.

Project Objective (No.)	The highlight of The Problem Statement
i	Kadazandusun's language is in its existential crisis. Usingmachine learning as an alternative to preserving the language.
ii	There is a very limited system that was developed for Kadazandusun language speech recognition.
iii	Little information of research regarding Kadazandusun speechrecognition.

Table 1.1: Relationship between Objective and Problem Statement

1.1.5 Project Scope

i. The purpose of this project is to develop a system mainly for educational purposes that targeting UMS students who are taking the Kadazandusun language.

ii. This project also can give an advantage for UMS lecturers who are teachingKadazandusun language *(*PPIB courses at UMS).

iii. This project will be conducted using Python and integrate with PHP.

iv. A website will be developed to make easy access to the system.

CHAPTER 2

LITERATURE REVIEW

2.0 Speech recognition

Speech recognition is any machine that has the ability of devices to give respond to the commands given through speech. Speech recognition can enter into various fields such as medical, entertainment, education, engineering, and even psychology, where it will enable hands-free control to various devices or equipment. It works best for learning and as a language preservation method in this modern world too. This technology works based on the voice as the research object and allowing the machine to automatically identify and understand the human language through speech signal processing and pattern recognition (Zhang et al., 2019). Any voice that goes through the microphone of the machine will be converted into an electrical signal on the identification of the system input. The speech recognition system creates a voice model based on human voice characteristics, analyses the input voice signal, and extracting the necessary features, and then creates the required speech recognition template or classification of the input.

In the process of recognizing the audio or speech, a computer will be used to make a comparison of the audio template that was stored in the computer with the new input took by the microphone of the device, which will be according to the model of the speech recognition. A model of speech recognition could be obtained by using any feature extraction and combine with any of the Neural Networks. One of the popular feature extraction that is used for audio extraction is

Mel-frequency Cepstral Coefficient (MFCC). As for the classificationof the audio, Convolutional Neural Network (CNN) is one of the most common algorithms usedfor audio classification.

2.1 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-frequency Cepstral Coefficients (MFCC) were introduced by David and Mermelstein whichlater become the most popular for SR systems. It is an excellent feature vector on which it can be applied to represent human voice or any audio signal (Winursito A., et al., 2018). Thereare five steps to get the MFCC coefficients.

I. Pre-emphasis:

Pre-emphasis commonly refers to filtering at higher frequencies of the speech signal and has been done in various speech processing applications such as speech recognition. The purpose of pre-emphasis is to balance the spectrum of voiced soundsthat usually steep roll-off at the higher frequency region. Normally for the sound of avoice, the glottal source has a -12dB/octave slope, approximately. However, to compare the glottal source of acoustic energy that vibrates from lips, it causes approximately +6dB/octave boost to the spectrum. Taking those into account, whena human recorded a speech signal through a microphone, it will yield an approximately

-6dB/octave slope downward compared to the true spectrum of the vocal tract. Thus,pre-emphasis will help to remove some of the glottal effects from the parameter of the vocal tracts. This function below is the most commonly used for pre-emphasis,

$$H(z) = 1 - bz{-}1$$

b → between 0.4 – 1.0 (controls the slope of the filter)

II. Framing and windowing:

Due to the stability concern of the speech signal, a stable signal can be assumed during 10-30ms. The function of framing is mainly to cut the speech with a long period to a short period speech signal to get a relatively stable frequency characteristic. The terms frame rate are the process of the feature being extracted once every 10ms.

Windowing, on the other hand, will mainly reduce the aliasing effect. This occurs when cutting the long signal to a short-time signal in the frequency domain.

III. Fast Fourier Transform (FFT):

To convert the time domain into frequency domain from each N frame of samples, fast Fourier transform (FFT) have to be performed. FFT will convert the convolutional of the vocal tract impulse response H[n] and the glottal pulse U[n] in the time domain. An equation will be supported by the statements given,

$$Y(w) = FFT[h(t) * X(t)] = H(w) * X(w)$$

IV. Mel Filter Bank Processing:

Passing the Fourier transformed signal through the Mel-filter bank or also known as a set of band-pass filters will compute a Mel spectrum. A Mel is a unit of measurement that depends on the frequency heard by the human ear. The range of frequencies in FFT is wide and the voice signal does not follow the linear scale. The Mel scale has a logarithmic frequency spacing below one kHz and a linear frequency spacing above one kHz. The approximation of Mel shown as follow:

$$F(Mel) = [2595 * \log 10 [1 + f] 700]$$

V. Discrete Cosine Transform:

This process is where the log Mel spectrum will be converted into a time-domain using Discrete Cosine Transform (DCT). After being converted, the yielded result is called as Mel Frequency Cepstrum Coefficient. An acoustic vector is the set of coefficients. Thisresulting in which each of the input utterances is being transformed into a sequence of the acoustic vector.

VI. Delta Energy and Delta Spectrum:

The frames and the voice signal change while the slope of a formant at its transitions. Due to that, there is an urge to add features that are related to the change in cepstralfeatures over time.

2.2 Convolutional Neural Network (CNN)

Machine learning has taken a strong transition in recent years as the Artificial Neural Network (ANN) has grown in popularity. Convolutional Neural Network (CNN) is a type of feed-forward neural network and on the top list of amazing ANN designs. It is a type of deep learning model which is similar to ANN. CNN was first found by Hubel and Wiesel and proposed in the 1960s. Both of them found that the uniqueness of CNN architecture could help to minimize the complexity of the feedback neural network when they were studying the neurons used for the local sensitivity and the direction selection of the cerebral cortex belongs to cats.

A convolutional neural network (CNN) is usually made up of several layers that are sorted according to their functions. CNN is a technology that mixing both the ANN and the up-to-date deep learning algorithms (Sakib et al., 2018). Furthermore, CNN has been applied to multiple systems such as image recognition and speech recognition. Researchers that come from different backgrounds and countries have shown their strong interest in this neural network as CNN has shown a promising performance especially in machine learning tasks and several computer vision.

2.3 Performance of Feature Extractions

The performance of Relative Spectral Amplitude-Perceptual Linear Prediction (RASTA-PLP), Mel-Frequency Cepstral Coefficient (MFCC), and Power Normalized Cepstral Coefficient (PNCC) were being compared in one research that compared the Speech Features using Artificial Neural Network. To improve the system performance in terms of their accuracy, the researchers extracted features from the voiced frames of the signal. An observation where the MFCC had outperforms the other two features which are PNCC and RASTA-PLP were made as to their conclusion (Rozario M., et al., 2019). The accuracy rate for their speaker identification performance considering all features of 50 speakers on a train data for MFCC, PNCC, and RASTA-PLP are 100%, 90%, and 100% respectively. For the test data, the accuracy for MFCC, PNCC, and RASTA-PLP are 98%, 88%, and 81.33% respectively.

Author	Feature	Train Function (Test data)	1 Data Frame (%)	½ Second Data (%)	1 Second Data (%)	Full Speech (%)
Rozario, etal. (2019)	MFCC	Traincgf	98.21	98.96	99.04	99
Rozario, etal. (2019)	PNCC	Traincgp	69.29	85.02	91.84	95.42