

**PRONUNCIATION MODELLING FOR
PENANG HOKKIEN**

BY
LEE CHUI CHUN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology
(Kampar Campus)

JANUARY 2020

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: PRONUNCIATION MODELLING FOR PENANG HOKKIEN

Academic Session: JANUARY 2020

I LEE CHUI CHUN
(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,



(Author's signature)



(Supervisor's signature)

Address:

BLK 19, 1-07, JLN Mesra 2,

TMN Mesra 13400,

BUTTERWORTH P.Penang

DR. JASMINA KHAW YEN MIN

Supervisor's name

Date: 23-04-2020

Date: 23-04-2020

**PRONUNCIATION MODELLING FOR
PENANG HOKKIEN**

BY

LEE CHUI CHUN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of


BACHELOR OF COMPUTER SCIENCE (HONS)

**Faculty of Information and Communication Technology
(Kampar Campus)**

JANUARY 2020

DECLARATION OF ORIGINALITY

I declare that this report entitled “**PRONUNCIATION MODELLING OF PENANG HOKKIEN**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature : _____ 

Name : _____ LEE CHUI CHUN _____

Date : _____ 23-04-2020 _____

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisors, Dr. Jasmina Khaw Yen Min who has given me this bright opportunity to engage in a Penang Hokkien Pronunciation Dictionary project. It is my first step to establish a career in this field. A million thanks to you.

Next I would like to express my sincere thanks and appreciation to the person who has help me in data collection part, recording sound. Finally, I must say thanks to my parents and my family for their love, support and continuous encouragement throughout the course.

ABSTRACT

Hokkien is a dialect of Chinese spoken in many countries such as in Singapore, Taiwan and China. For example, Minnan is spoken in Taiwan and Philippine Hokkien (lan lang oe) is spoken in Philippine Chinese group. Hokkien spoken in different countries and even within a country itself might vary in terms of pronunciation and vocabulary from one place to another where dialect occurred. There are several Hokkien dialects can be found in Malaysia such as in Johor, Penang and Kedah. In this study, Penang Hokkien (PH) is focused as it is very distinctive. PH is a dialect spoken in Penang, Perlis, and Kedah. In this thesis, an approach to generate pronunciation dictionary of PH is proposed. Besides, unique words of PH are determined which include borrow words from Malay or English. The pronunciation dictionary generated will be useful in building PH text-to-speech (TTS) system. The TTS system will be useful for those who like to learn PH. Besides, it will help to preserve the dialect and culture in it. The system is also useful in some different places that require TTS technologies such as in local animation film. Last but not least, it can be a useful tool for communicating with the local.

Keyword – Penang Hokkien, Grapheme to Phoneme (G2P)

Table of Contents

PRONUNCIATION MODELLING FOR PENANG HOKKIEN	i
DECLARATION OF ORIGINALITY	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
Table of Contents	v
LIST OF TABLE	vii
LIST OF FIGURE	vii
LIST OF ABBREVIATIONS	ix
Chapter 1: Introduction	1
1.1 Introduction	1
1.2 Problem Statement	1
Problem 1: Unidentified of unique vocabularies of Penang Hokkien.....	1
Problem 2: Unknown phoneme set of Penang Hokkien.	2
Problem 3: Inexistence of formal pronunciation dictionary of Penang Hokkien. 	2
1.3 Motivation.....	2
1.4 Project Scope	3
1.5 Objectives	3
1.6 Impact, significance and contribution	4
1.7 Background Information	4
Chapter 2: Literature Review	6
2.1 Penang Hokkien	6
2.2 Way to get Data Collection for unique words	7
2.3 Phoneme Identification	7
2.4 Grapheme to Phoneme	9
2.5 Tone Sandhi	11
2.6 Speech synthesis	13
2.7 Hidden Markov Model (HMM)	13
Chapter 3 Methodology	14
3.1 Introduction	14

3.2 Data collection through Conversation Recording	14
3.3 To determine Unique vocabularies in PH	15
3.4 To determine the phoneme set in PH	16
3.5 Pronunciation modeling for PH	17
3.6 Implementation Issues and Challenges	17
Chapter 4: Experiment and Results	19
4.1 Data collection	19
4.2 Unique vocabularies in PH	19
4.3 Determine Phoneme Set of PH	24
4.4 Generate Pronunciation modeling of PH	34
Chapter 5: Data Analysis	36
Chapter 6: Conclusion	38
REFERENCE	39

LIST OF TABLE

Table Number	Title	Page
Table 1:	Mapping of Objectives and Problem Statement	4
Table 2:	compare of Traditional technique with Standard modelling technique	8
Table 3:	advantage and disadvantage for some G2P technique	11
Table 4:	Detail of dialogue	19
Table 5:	Record of word occur in dialogue	21
Table 6:	List of unique word	24
Table 7:	Example of mandarin word convert to PH word	25
Table 8:	Show the IPA and PH grapheme	26
Table 9:	Show the original and changed word or grapheme	26
Table 10:	Show the original and simplies grapheme	27
Table 11:	Show the percentage of each sample WER and PER	37

LIST OF FIGURE

Figure Number	Title	Page
Figure 1:	Overview of a typical TTS system	13
Figure 2:	Probabilistic parameter of a HMM (Automatic Speech recognition system Voice recognition Authôt, n.d.)	13
Figure 3:	Basic architecture of Pronunciation Modelling for Penang Hokkien	14
Figure 4:	Sub flowchart of determine unique vocabularies	15
Figure 5:	Sub flowchart of determine phoneme for PH	16
Figure 6:	sub flowchart of pronunciation modelling for PH	17
Figure 7:	Praat data alignment	20
Figure 8:	Text file that save the align word	21
Figure 9:	Calculate length of total word	22
Figure 10:	Function of remove the unique word	22

Figure 11: Calculate length of total general word, total unique word and percentage unique word	23
Figure 12: Data of Penang Hokkien	25
Figure 13: Remove Chinese word	27
Figure 14: Changing of file from beginning to generate phoneme type	28
Figure 15: Simplifies and add consonant and vowels	28
Figure 16: Function to generate phonemes for non-changing general word	29
Figure 17: Save as into text file	29
Figure 18: PH phonemes for second set data (non-changing general word)	30
Figure 19: Function of changing the grapheme	30
Figure 20: Saving all changed data in a file and prepare data for generate phonemes	31
Figure 21: Changed all grapheme text file	31
Figure 22: Changed grapheme file that only have pinying use to generate phonemes	32
Figure 23: Function of simplifies grapheme and add consonant and vowel	32
Figure 24: Generate phonemes for changed pinying (grapheme)	33
Figure 25: Result for the changed pinying phonemes	33
Figure 26: Result of combined changed pinying phonemes and non-changed pinying phonemes	34
Figure 27: Read unique word file and apply simplified rule	34
Figure 28: Generate phonemes and save final result in text file	35
Figure 29: Penang Hokkien Pronunciation Dictionary	35
Figure 30: Generate 10 set of testing sample with each 100 pronunciation	36
Figure 31: Sample of PH Vocabularies with Pronunciation	36

LIST OF ABBREVIATIONS

<i>PH</i>	Penang Hokkien
<i>NLP</i>	Natural Language Processing
<i>TTS</i>	Text-To-Speech
<i>HMM</i>	Hidden Markov Model
<i>G2P</i>	Grapheme to Phoneme
<i>PPH</i>	Penang Peranakan Hokkien
<i>LSTM</i>	Long Short-Term Memory
<i>RNN</i>	Recurrent Neural Network
<i>IPA</i>	International Phonetic Alphabet

Chapter 1: Introduction

1.1 Introduction

Hokkien is a dialect of Chinese spoken in many countries such as in Singapore, Taiwan and China. Hokkien spoken in different countries and even within a country itself might vary in terms of pronunciation and vocabulary from one place to another where dialect occurred.

There are several Hokkien (PH) dialects can be found in Malaysia, and PH is one of the dialects. This dialect mostly spoken in Penang, Perlis, and Kedah. Although it used by many states but PH is not simple to learn as it does not have formal writing system, pronunciation, and also it mixes much with the borrow words. All of these cause non-native speakers hard to understand.

In this research, the unique vocabularies in PH need to be found out. Besides, the phonemes set in PH will be identified. Finally, pronunciation dictionary for PH is developed.

1.2 Problem Statement

Penang Hokkien is a dialect language that used in many states especially in Penang state but there is an issue which is it does not have its own or formal writing system and also pronunciation dictionary. As there are quite many words in PH is borrow with other language or dialect, such as the word *mata* (English: police) in PH, it is a Malay word with the meaning of *eyes*, but, in PH, it brings different meaning which is *police*. All of the borrow words do not have formal documented pronunciation. Hence, it might be challenging to develop pronunciation dictionary for PH as it does not have writing system, quasi-unknown unique vocabularies and phoneme set. Furthermore, most of the youngest are using Mandarin and English that might cause PH becoming extinct language in future.

In short, the problem statement shows below:

Problem 1: Unidentified of unique vocabularies of Penang Hokkien

The unique vocabularies of PH are identified. This would happen because the PH dialect is keep changing and combined with other language (Chen and Chen*, 2018). As mentioned in (Chen and Chen*, 2018), Nowadays, there are a lot of people do not

know how to speak with PH. Therefore, they will mix the PH with other languages. The number of native speakers who speak well in PH is gradually decrease. Because of mixture of language, it makes lot of unique vocabularies form. (Chen and Chen*, 2018)

Problem 2: Unknown phoneme set of Penang Hokkien.

The phoneme set for PH dialect is unknown as it does not have formal writing system. It is also due to it has mixed of other language as Malaysia have some different ethics with different spoken languages or dialects (PEOPLE, POPULATION AND LANGUAGES OF MALAYSIA | Facts and Details, 2019). Since there are a lot of loan words, they do not have actual phoneme set in pronunciation. As mentioned in the research of “A Study of Penang Peranakan Hokkien” (Teoh, Lim and Lee), PH is using lot of Malay words and all of these words have changed to different pronunciation, some may keep as Malay sound but some may change the tone to PH. With these changes, it causes unknown phoneme set in PH. Some might use the similar word but with different pronunciation and meaning.

Problem 3: Inexistence of formal pronunciation dictionary of Penang Hokkien.

There are a lot of dictionary of PH in online website (Speak Hokkien Campaign, 2017), (International Phonetic Alphabet (IPA), 2020) but each of the websites are using different style or phoneme set. Therefore, there is no formal pronunciation dictionary for PH. Besides, PH consists of many unique vocabularies that making the task of developing PH pronunciation dictionary becomes more challenging.

1.3 Motivation

Penang Hokkien is a traditional dialect and recently the population of using PH in the Penang area is reduced especially for young generation. According to The Star Online, it has mention that used of PH is slowly disappearing and is slowly replaced by using Mandarin and English. (Wong 2017). Currently, many of the younger generations are not able to speech PH since most of the family, their family language are English or Chinese and many of the newborn family their parent unable to speech Penang Hokkien. This situation will happen is because English had become a national language so caused most of the family to teach their child with English when they starting to learn the language. Besides, during 20th, Dr. Sun the person who started to promote and introduce the use of Mandarin for gathering all Chinese subgroups and the purpose he

doing this is want to prevent conflict (The & Lim 2014) this causes the Chinese language (Mandarin) gradually replaced the origin dialect such as Hokkien until today Mandarin become a subject that must be learned in primary school and become a language used in normal life for communication with other. So, this becomes the reason causes missing of formal pronunciation dictionary for PH. Therefore, to let the youngest generation learns to speak PH correctly, pronunciation modeling is needed as compared to teach them through writing, speaking is faster and easier to learn. Besides, with pronunciation modeling, the youngest who want to learn become much more easier.

1.4 Project Scope

The goal of this project is to develop pronunciation dictionary for Penang Hokkien (PH). It not only can show the general PH vocabularies, it also can show the unique vocabularies of PH such as words that borrow from Malay. Thus collecting material of PH is required. Next, the project will focus on letting the computer understand and recognize the unique vocabularies, the phonemes belong to. The Natural Language Processing (NLP) have the ability to let the computer to understand human language nor matter in text-type or speech-type and also let computer ability to determine which word belongs to which phonemes. Speech synthesis technique is used for processing the input and output. To make it process, there are some algorithms need to apply such as an algorithm to recognize the phonemes.

1.5 Objectives

There are three objectives in this project.

- To identify the unique vocabularies in Penang Hokkien.
- To identify the phonemes set in Penang Hokkien.
- To develop pronunciation dictionary for Penang Hokkien.

Objectives	Problem Statements
Identify the unique vocabularies in Penang Hokkien	Unidentified of unique vocabularies of Penang Hokkien.

Identify the phonemes set in Penang Hokkien	Unknown phoneme set of Penang Hokkien
Develop pronunciation dictionary for Penang Hokkien	Inexistence of formal pronunciation dictionary of Penang Hokkien

Table 1: Mapping of Objectives and Problem Statement

1.6 Impact, significance and contribution

By having the pronunciation dictionary of Penang Hokkien (PH), the dialect becomes more complete as grapheme and phoneme that suitable for Penang Hokkien is generated. Next, the task of learning Penang Hokkien becomes much more easier. In this project, collecting the Penang Hokkien data, and identification of unique words and general words in PH will be conducted. Grapheme to phoneme conversion rules for Penang Hokkien are also crafted. Finally, PH pronunciation dictionary is developed which consists of word, pinyin, and also phonemes to represent the pronunciation.

1.7 Background Information

Penang Hokkien (PH) is a dialect of Chinese (Mandarin) language, and this dialect language are originally comes from Fujian (福建), China. In China, it names as Hokkien. The origin of Penang Hokkien is from Southern Min dialect of Zhangzhou (Soon 2014; Hing 2017). The reason Hokkien language will bring into Malaysia this country is because there are many Hokkiens (福建人) are left from China and the reason to left is to search for a better place and also to left from war in China. Since during that time, not only people need to pay expensive tax in China, they also live in fear because war is getting nearest. Therefore to get a better life style and far away from war, they started to move out from China and this is why they will reach in Penang and this is why Hokkien will occur in Malaysia. During that time there is no national language which mean heir communication is poor, therefore, most of them speak with their own dialects such as Hokkien, Cantonese and other but not speak with Mandarin. Compare to the old generation PH, nowadays PH is changed to become modern language style as it mixed with different language dialect and not as pure as in old generation. So Penang Hokkien has become a new dialect in Malaysia and it is different

from China Hokkien, Taiwan Min-Nan since Penang Hokkien had to borrow with other language or dialect such as Malay, Mandarin, and English. Penang, Johor, Malacca, Kelantan, Terengganu, Kedah, Perlis, and Sarawak most of this state are using Hokkien to communicate. And PH used by most people in Penang, Kedah, and Perlis. So it is mainly in the Northern region of Malaysia (Teh & Lim 2014). A dialect is important to be studied first is all dialect is kind of traditional next it is important as with know well other person dialect it will make your life easier when communicate with other, third it is important as with a complete dialect it can improve the current voice recognize level and also can bring more benefic to other such like can help those person who cannot speak and blind people express out what they want.

The technology that will use in this research is speech synthesis. Speech synthesis also can know as Text-To-Speech (TTS). This research will use TTS as TTS it able to convert the text into speech which can help it pronunciation out the word. In TTS there has a process name as grapheme-to-phoneme (G2P) which can match word to it phoneme so that in the end the system know how to sound the word. Example borrow book in PH is tsioh tsheeh give the word with phoneme become /ts/ /ioh/ /tsh/ /eeh/ using this phoneme after train it able to pronouns it.

Taiwan Minnan IPA will use in this research. The main reason using Taiwan Minnan IPA is due to Taiwan Minnan are using it for converting grapheme to phoneme. Since Taiwan Minnan is almost similar to PH therefore it is a best way to use it as a base. To converting grapheme to phoneme first need to create out suitable grapheme and the converting rule, rule is the core in the G2P since it is the place that form correct phoneme.

In this research, will focus on identify the unique vocabularies in PH such like the loan word da break (stop the car), identify the phonemes set in PH such like /kha/ (leg) the phonemes is /kh/ /a/ and last using combine it to develop pronunciation dictionary for PH.

Although PH is used by many states, but there is no formal writing system for PH beside this there also do not have a formal or complete pronunciation dictionary for PH. Therefore to create a pronunciation modeling speech synthesis technology is needed.

Chapter 2: Literature Review

2.1 Penang Hokkien

According to the Lim and Teh (2014) research, it stated that Penang Hokkien is a dialect that speaks the most in Penang state. PH is the language that came from Fujian, China. Because of the assimilation with the local culture PH is identified as a variant of Minnan. This language will come to Malaysia is because during the early state in Malaysia there a lot of Hokkiens came to Penang for trading and livelihood. Besides, the earliest Chinese group live in Malaysia are Hokkiens and during that time many of Chinese people in Malaysia are spoken with Hokkien since it is the biggest dialect and also because at that time there is no national language, therefore, all people are saying own dialect such as Hokkien, Cantonese. The first group of Chinese settlers in Malaysia came to Georgetown, Penang.

PH language has changed because of several issues one of the issues is because Malaysia have different race and this causes people started to borrow the other dialect or language word one of the reason will borrow the other dialect is some word is must easier to say compare to using the original word, as keep longest it became the new PH which mixed with the Malay, English word and it is different with early Hokkien. PH borrow the word from Malay and the other local language such as the word pun in Malay the meaning of pun in Malay and PH is different. (Hing) the second reason for PH language has changed is because nowadays the use of PH is declining and this makes many of traditional PH word missing and this situation might keep going. Although PH will not totally disappear in Malaysia since there still have family take PH as family dialect, PH language will change become mixture with a lot other language or dialect as some of the words do not know how to say or mention it with PH.

In the research of (Teoh, Lim, and Lee 2017) they have classified certain loan words in Penang Peranakan Hokkien (PPH). Such as in animal area, buaya for crocodile, katak puru for the frog, kutu for lice, chacing for worm and other. Not only in the animal area, clothing, accessories also included example anting for the earring. And this of the word also used in PH however there also some word in PPH is not applied or used in PH such like mosang is not used in PH as in PH word it is named as hor lay. There also Malay word used as verbs and adjective. Such as the word laku which mean for saleable,

pantang mean for superstitious and other. There also use of imagery in Malay loan word such as haram mean as bear a grudge, kesian mean as pity, mabok mean as drunk. All of the words have to make the PH evolve to modern PH. And because of the word is belong to Malay it causes they are not listed in Hokkien dictionary beside that the pronunciation of those words is also not the same as pronunciation in Malay. It using Hokkien vowel and consonant to pronounce. Therefore, to do pronunciation modeling for PH, the loan word must be collected and provide the PH phoneme to every word.

2.2 Way to get Data Collection for unique words

There is a lot of ways to do data collection, and the data collection is an important part it must be reliable. Since the data is used to run the whole research if the data wrong it means all the result become unreliable and is error. There are many ways to collect material one of the way is to collect the data through the reliable book, article, web, lexicon and also a dictionary (Kabir, Syed Muhammad, 2016). One of the research the researcher get the material from The Ministry of Education launched the online 臺灣閩南語常用詞辭典 (Taiwan Minnan dictionary), beside it also get the material from Mandarin-Taiwanese dictionary. and there also research get the material from LDC Iraqi-Arabic Morphological Lexicon. Beside of this the material also can get by record the conversation of two or more participant. In this research, will get the material from the 臺灣閩南語常用詞辭典 (Taiwan Minnan dictionary) as a base this is because there is no official Penang Hokkien dictionary, writing system and phoneme. To make sure the data collect is correct, four participant will be attend to record their conversation in the conversation there will mixture of PH general word and also the unique word. And last the 馬來西亞北部的詞 (Penang Hokkien Accent) will be use as guided.

2.3 Phoneme Identification

According to the research system and method for pronunciation modeling (LJOLJE et al. 2010). To model speech having different dialects, pronunciation modeling is a way that can make it. There are two standard way of pronunciation modeling in art is human linguists manually creating pronunciation dictionaries and automatic approach creates acoustic clusters. Compare to traditional pronunciation

modeling techniques are not very able to address dialectal variation because it is easily recognized and it is slow and expensive.

According to the research Dau, Ren, et al. (2005), a bi-lingual large vocabulary Speech recognition experiment based on the idea of modeling pronunciation variations are described. The goal of this research is to convert Taiwanese and Mandarin speech into a Chinese character. In this research will develop one-pass, three-layer recognizer and the performance of recognizer will be determined by three different pronunciation models. The three models are bi-lingual acoustic model, integrated pronunciation model, and tree-structure based searching net.

Knowledge-based and data-driven approaches are combined with an integrated method and are used during the pronunciation model. In this paper, the experiment result showed that if using three different pronunciation models it can improve the character error rate. The best performance for test mandarin and test Taiwanese is 16.2% and 15.0%.

Based on Bellegarda (2016) research, state that pronunciation modeling is the process that provides a suitable phoneme to each word in a given vocabulary. Therefore the important thing to speech-to-text or text-to-speech is having a good pronunciation modeling. This is because phonemic expansion is the need for the selection of the proper Text-to-speech unit from which to generate the desired waveform.

To create a set of phonemes there is two way can be used which is the linguists manually create each entry and the other way is automatically derive pronunciations from the word orthography. Linguists manually create is a way that often subjects for inconsistencies inconsistencies and inherently dependent on the language considered. And for automatically derive pronunciation it is more on the processing of unique vocabulary word assign with the phoneme. (Bellegarda 2006)

Traditional pronunciation modelling technique	human linguists manually creating pronunciation dictionaries and automatic approach creates acoustic cluster
Expensive and slow	Less expensive and more faster
Not able to address dialectal variation	Able to address dialectal variation

Table 2: compare of Traditional technique with Standard modelling technique

Before 2010 there also research state that there got two way used to create phonemes which is linguists manually create each entry and automatically derive pronunciation from the word orthography. However there will have same problem in using in PH as PH is mixed lot with other dialect this make hard to know all the PH dialect therefore to manually create each entry it may a challenge.

2.4 Grapheme to Phoneme

Grapheme to Phoneme conversion (G2P) is a process of converting word to phonemes, by using the G2P algorithm (Jurafsky & Martin 2009). Such like predicting google to /g u g @ l/ (Rao, Peng, et al) when using G2P. The phoneme is the sound that distinguishes one word from another. Such as in PH 跑 (run) is /tsau/ it has 2 phoneme which is /ts/ /au/ Grapheme is known as the spelling choice to represent the phonemes or grapheme is the letter that represents a phoneme. In joint-sequence models G2P it divided into 3 parts which is Aligning: aligning au -> u, Training: learning au -> u conversions and Decoding: finding the most suitable pronunciation given the model. (Bisani and Ney 2008).

According to (Rao, Peng, and et al 2015) research. The G2P conversion is using based on Long Short-Term Memory (LSTM) recurrent neural network (RNN). The contextually-aware decision is able to make by using LSTM as LSTM are able to take more than a few graphemes before it came out any phoneme. RNNs is suited for sequence modeling tasks as it can process the current input by using cyclic connections.

Therefore it is suitable for phoneme recognition also handwriting recognition. But RNNs consist of some weakness which is having trouble to the vanishing gradient and exploding gradient problem. Therefore LSTM RNNs is used to solve the problem. To implement LSTM based G2P, configured LSTMs by a number of graphemes and phonemes is set to be equal to the input layer of size and the output layer of size. Output layer with softmax activation and cross-entropy loss function is set up for unidirectional LSTM together with 1024 memory unit. In (Rao, Peng, et al 2015) research, they using phoneme error rate (PER) and word error rate (WER) to evaluate the performance. The resource is using the publicly available CMU pronunciation dictionary, 2,670 words are used to find the stopping criteria during training, and 12,000 words used as testing

and so it is directly comparable. In conclusion, to perform G2P conversion, LSTM-based architecture, is suggested to be used.

In Yeong and Tan (2011) research, there also have used some of the G2P methods. In this paper it not only apply grapheme information, but it also combines with syllable information and word sequence information to identify the language for word. Firstly the experiment is started by using syllable structure information. Using the syllable and sequence of the syllable in a word it can predict the language of an unknown word. Before the word merged into syllables the words are segmented to a sequence of graphemes. To know the presence of the unknown word is belong to Malay or English, chain rule applies to calculate the probabilities. And because of limited training data, n-gram will be applied to calculate. Next the language identification by using grapheme. The only thing that needs to change is replacing the syllable sequence to grapheme sequence. And last combine the three methods by using interpolation to do language identification. In conclusion, it shows that the interpolation result is better to compare to a single approach.

In research (Tan and Malancon) it also used the G2P to generate the pronunciation. It mentioned that to identify the pronunciation of unknown words it is better to use G2P as by apply G2P rules it can predict out the pronunciation belong to those words. In this research, it using the G2P for generating the pronunciation dictionary for Malay automatic speech recognition system (ASR). Since the G2P is a rule-based system, therefore, it is flexible for adding, removing and handling. To detect the pronunciation of a word, morphological and syllable tool can be applied and this tool is contained in G2P. In this research eight G2P conversion rule applies the example of some rule like general replacement rule, every grapheme is designed to Malay phoneme by default. Duplicate grapheme rule, the same graphemes are changed to single phonemes. Because of the Malay text also will have chance appear of an English word. Therefore it needed to use other approaches to generate it. First, need to get know that the word belongs to English, therefore, need to compare with the English dictionary. Next, the nearest English phonemes to Malay phonemes will be mapped. In this research, it shows that using G2P to generate the pronunciation dictionary it can product must better compare to using the manually verified.

	Advantage	Disadvantage
LSTM RNN	LSTM make Contextually -aware decision able to make RNN can process current input by using cyclic connections	RNN have trouble to vanishing gradient and exploding gradient problem
Combine of grapheme, syllable information and word sequence information	Can predict language of unknown word	More suitable for short test sentences

Table 3: advantage and disadvantage for some G2P technique

2.5 Tone Sandhi

Tone sandhi is change occurring in tonal language, so every single words will assigned with tones. It normally used to simplify the tone, like from bidirectional to one direction. Tone sandhi used in many language such as Taiwan Minnan, China Hokkien, Teochew, Mandarin Chinese. Some of them will has complex system, example the syllable will change into a different tone, and the every tone changed is depend on their final consonant which mean almost all the word are keep changing the tone. In Taiwan Hokkien there will have seven tone, in this seven tone there have two tone name as checked syllables and the other five tone do not stop which will always keep changing. In the research Jane S. Tsay (2007) they using the tone sandhi into the G2P process, tone sandhi will acts as a core during G2P process since it is conversion rule, therefore the rule must be correct so that the grapheme can generate.

In the research (Liang, Yang, et al. 2004) research, they describes about the Taiwanese Text-to-speech system for language learning. In the research, tone sandhi has been used as one of the manual transcription, the reason used is the mandarin and Taiwanese is tonal language. There are 3stage in this research which is collect new data from internet, the next stage is select out a set to cover all Chinese characters and minimize number of sentences. The last stage is compare the automatic transcription with manual. In this paper, the researcher using the traditional way which is consist

seven lexical tones. Every tones have their different pitch and name which is High-Level, Mid-Level, Low-Falling, High-Falling, Mid-Rising, High-Stop, and Mid-Stop. The name is started from tones 1 to tones 7. In this paper the result show the performance in tone sandhi is 65.43% in Expert1 and 62.43% in expert2. In conclusion, the researcher successfully constructed Taiwanese TTS system, and most of the mandarin task can successfully change into Taiwanese.

In the research (Lunn, Lau, et al. 2007), it clearly show that the purpose of taking this research is to solve the problem happen in Taiwan Minnan tone sandhi system. In this research, the researcher will using the Taiwan Minnan text as the source, and translate it into Chinese word. And next, will access to the Chinese Electronic Dictionary (CED) to get the Part-of-speech (POS) information. The purpose of taking POS information is to use together with the tone sandhi rule to mark every syllable with own post-sandhi tone maker. And finally, the Romanized word will take as an input into the system and last generate out the outputs the tone makers. To complete this part, implement the Taiwan Minnan tone sandhi processing system is needed. In this paper result it get a high accurate with is 97.39% in training data and 88.98% in testing data. In this paper, the tone sandhi also using the traditional way which have total seven tone. This seven tone is name as following sequence 1: im-*pi*ⁿ (high flat), 2: siang (high to low), 3: im-khi (low), 4: im-jip (middle short), 5: iang-*pi*ⁿ (low rising), 7: iang-khi (middle flat), 8: iang-jip (high short). In this research also mention that in the world level, basic tone are normally pronounced in the last syllable. Which mean every word in the last syllable is not follow the tone sandhi but the other will using tone sandhi. As an example in research: tâi (platform), tâi-gí (Taiwanese language). The word that pronounced as basic tones is the word that have underline, and without underline will pronounced as sandhi tones. In this research also state that sandhi tone will have several way to manifest itself. This several way is name as Normal sandhi, Following sandhi, Neutral sandhi, Double sandhi, Pre-á sandhi, Triplicate sandhi, and the last is Rising sandhi. By using all the tones sandhi and POS create out a suitable tone sandhi rule for the system. As conclusion in this paper, the research is successful.

Base on the two research above, it show that tone sandhi is very suitable for Taiwanese Minnan, since Penang Hokkien is almost same with Taiwanese Minnan.

2.6 Speech synthesis

Speech synthesis is the computer-generated simulation of human speech. (What is speech synthesis? - Definition from WhatIs.com, 2005) Text-to-speech (TTS) system can provide the function to convert the text to speech. And it is divided into two part which is front-end and back-end in this paper will focus on front-end. In front-end, first is need tokenization the text which converts the text into the equivalent of written-out words and next will passed to do phonetic transcriptions to every single word. This process is known as grapheme-to-phoneme (G2P).

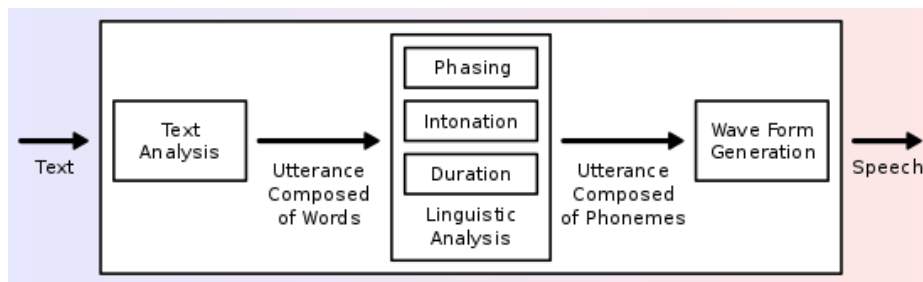
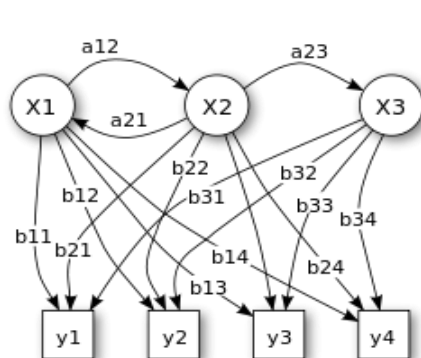


Figure 1: Overview of a typical TTS system

2.7 Hidden Markov Model (HMM)

HMM known as a probabilistic method. To determine the best possible sequence, the joint probability of a sequence of hidden state is needed, to joint probability of a set of hidden states, HMM is allowed to do. In the research Ungian lunn (楊允言) (2009) there are using the HMM probabilistic model to train the training data to get the most adequate Mandarin word. After selected the word Maximal Entropy Markov Model is used to the classifier. With this, the result gets a high accuracy rate of 91.5%.



X –states

y –possible observation

a –state transition probabilities

b –output probabilities

Figure 2: Probabilistic parameter of a HMM (Automatic Speech recognition system | Voice recognition | Authôt, n.d.)

Chapter 3 Methodology

3.1 Introduction

In this research, unique vocabularies in Penang Hokkien (PH) will identified, and the grapheme that suitable to use will be listed and the grapheme-to-phoneme (G2P) conversion rules will crafted. Figure 3 shows the overview flow for PH pronunciation modeling.

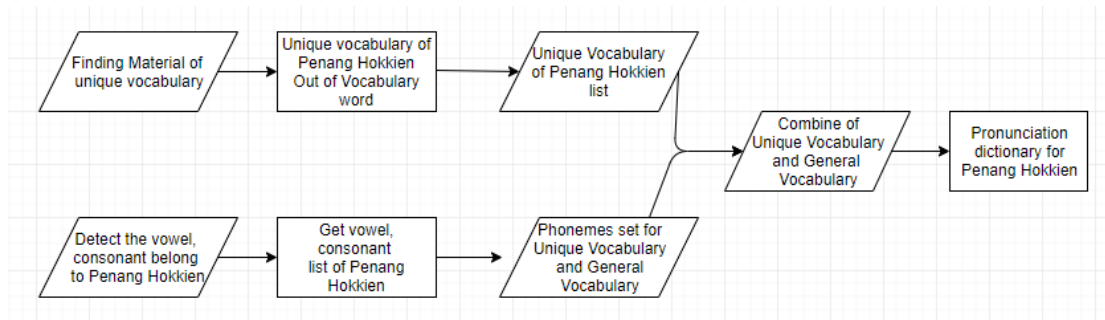


Figure 3: Basic architecture of Pronunciation Modelling for Penang Hokkien

According to the Figure 3, the whole research will begin with finding out the material consists of a unique vocabulary of PH. And next is get the unique vocabulary of PH the last step for this stage is listed out the unique vocabulary of PH. For the other stage, firstly determine out the vowel, consonant of PH. After finish, will get a list for Penang Hokkien vowel and consonant. Finally for this stage will set phonemes for unique vocabulary and general vocabulary of PH. After completion of this two-stage, will combine unique vocabularies and general vocabularies together with the phonemes. Finally, the pronunciation dictionary for PH is created out.

3.2 Data collection through Conversation Recording

Penang Hokkien does not have writing system. Therefore, to collect PH text, the most suitable way is to conduct conversation recording between 2 or more people. Then, to find out the unique words in PH, alignment between text in Mandarin, English and Malay is conducted. The unique words are then extracted out and included in pronunciation dictionary of PH.

3.3 To determine Unique vocabularies in PH

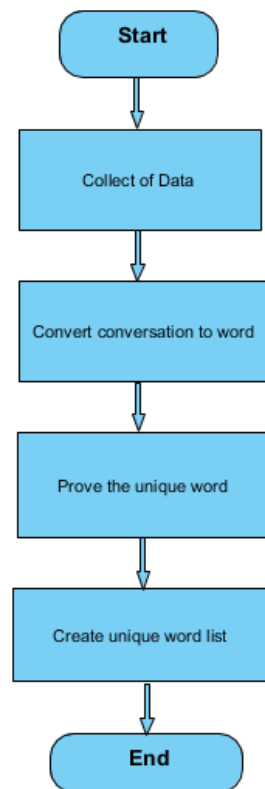


Figure 4: Sub flowchart of determine unique vocabularies

Figure 4 shows the sub flowchart of the stage one which is to identify the unique vocabularies in PH. At the beginning of this process, it is to collect of the data where conversation recording in PH is conducted. After finish recording, the conversation in PH is transcribe into Mandarin. Task in this task it consist of general word and also the unique word. So to prove the unique word is borrow word so align method is taken to compare. With this all the general word and unique word will be separate out. Last create a list that consist of all unique word in the dialogue.

3.4 To determine the phoneme set in PH

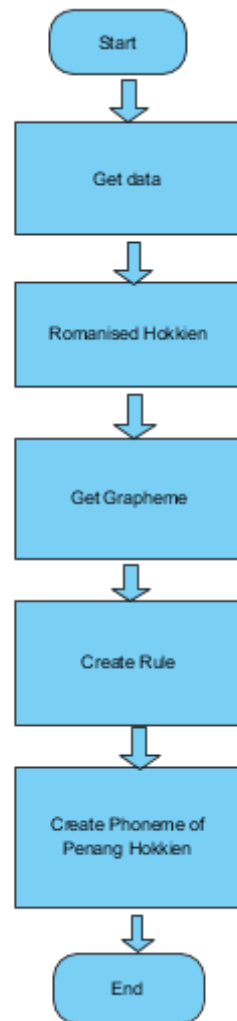


Figure 5: Sub flowchart of determine phoneme for PH

Figure 5 shows the sub flowchart of the stage two which is to identify the phonemes set in PH. At the beginning of this process, it is to find the suitable grapheme for PH. Since PH is almost similar to Taiwanese Minnan, grapheme of Taiwanese Minnan will be used. To find out the possible grapheme of PH, a website (Speak Hokkien Campaign, 2017) is used as reference. The next process is to Romanize the Hokkien which means it is to romanize the Mandarin words into Romanization word. It is similar to pinyin for the Romanized process where we use the information from Taiwanese Minnan and also the website about PH (Speak Hokkien Campaign, 2017).

After that, phoneme to phoneme conversion rules are crafted to convert Taiwanese Minnan pronunciation to PH pronunciation.

3.5 Pronunciation modeling for PH

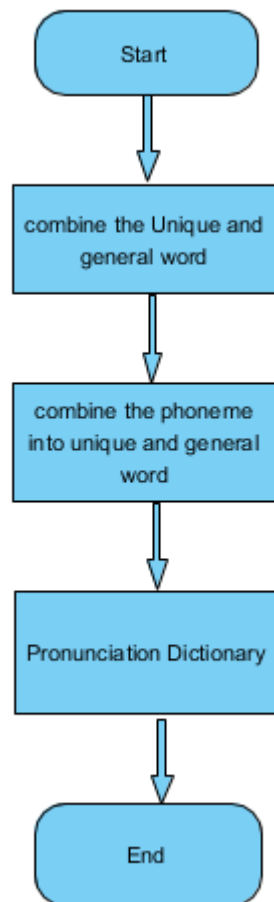


Figure 6: sub flowchart of pronunciation modelling for PH

Figure 6 shows the sub flowchart of final stage. First, all possible vocabularies of PH are listed which include the general vocabularies and unique vocabularies. Next, the pronunciation of those vocabularies will be generated where pronunciation dictionary of PH is developed.

3.6 Implementation Issues and Challenges

One of the difficult issues and challenges in this project is to collect the Penang Hokkien data. In current stage, there is lack of Penang Hokkien data set, the reason is Penang Hokkien does not have official written word, almost all the Penang Hokkien people are writing by using Chinese word but speak with Penang Hokkien. Therefore, some of the words and pronunciation are totally different. Besides, there are also some

Penang Hokkien words that totally cannot be written in Chinese characters. So, to solve this problem, first is to use the Taiwan Minnan as a base since most of the general words in Taiwan Minnan are the same as in PH. There is a complete Taiwan Minnan pronunciation dictionary (教育部臺灣閩南語常用詞辭典, 2011) that can be found online. Next, to solve the unique vocabularies that do not appear in Taiwan Minnan and Chinese words, conversation recording among Penang Hokkien native speakers is conducted and also extracting out from the PH official website. The next challenge is that there is almost no proper list of graphemes and phonemes for Penang Hokkien. Hence, graphemes to be used in Penang Hokkien need to be identified. Again, Taiwan Minnan has almost similar vocabularies and pronunciations with Penang Hokkien. Therefore, graphemes used in Taiwan Minnan will be applied to Penang Hokkien with some modification. Next, to identify the phoneme set in PH, phoneme-to-phoneme conversion rules will be crafted from Taiwan Minnan to Penang Hokkien.

Chapter 4: Experiment and Results

4.1 Data collection

In this research the data collect is get by recording the conversation between people by using Penang Hokkien. To get the real conversation that will talk in Penang so in this research, a family with 4 person is been invited to record their conversation in their daily life. The conversation is around 16 minute and 31 seconds. And this conversation can be categories as short daily conversation. In this conversation, it include of some food, daily equipment, transport, and also emotion word. The language use in this conversation is Penang Hokkien which include of the unique word. Such as kopi (coffee), mata (polish), alamak (oh my god) all the word had occur in the conversation. Beside collect the data by using record method, in this research also get the unique word from the PH website, then will combine together become a set of unique word data.

Number of participant	4
Total duration	16 min 31 sec
Category	1

Table 4: Detail of dialogue

4.2 Unique vocabularies in PH

In Penang Hokkien, it consists of general words and unique words. Those words that are belong to unique words or borrow words will be identified. After the conversation recording, it is transcribe into text using Chinese words. The transcription will then translated into Malay text. Next word-based alignment is conducted in order to find out unique words in PH. In this process, Praat is used for alignment. Figure 7 below shows the sample of alignment. The sound recorded will align with the Chinese words. By using the Chinese words will then align to each same meaning Malay words. Those words that same with or using Malay words will be place at same section with Chinese words.

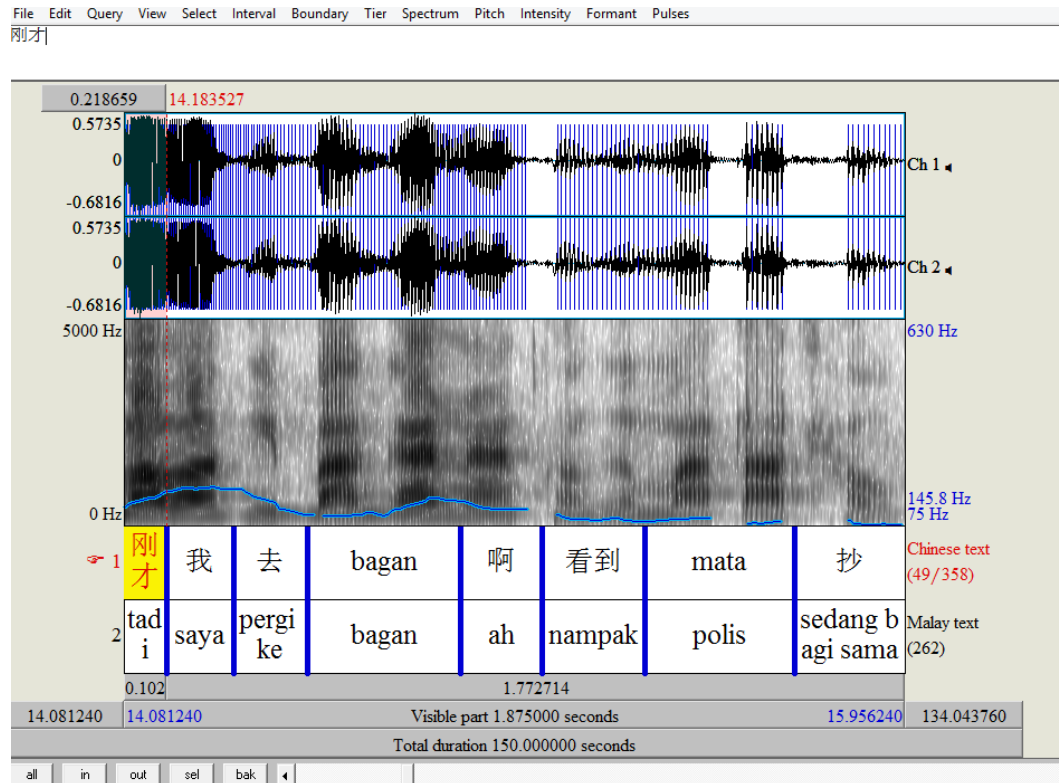


Figure 7: Praat data alignment

After the word is convert to Mandarin word with unique word (Malay or English), will save as a text file which shown as in Figure 8 diagram.

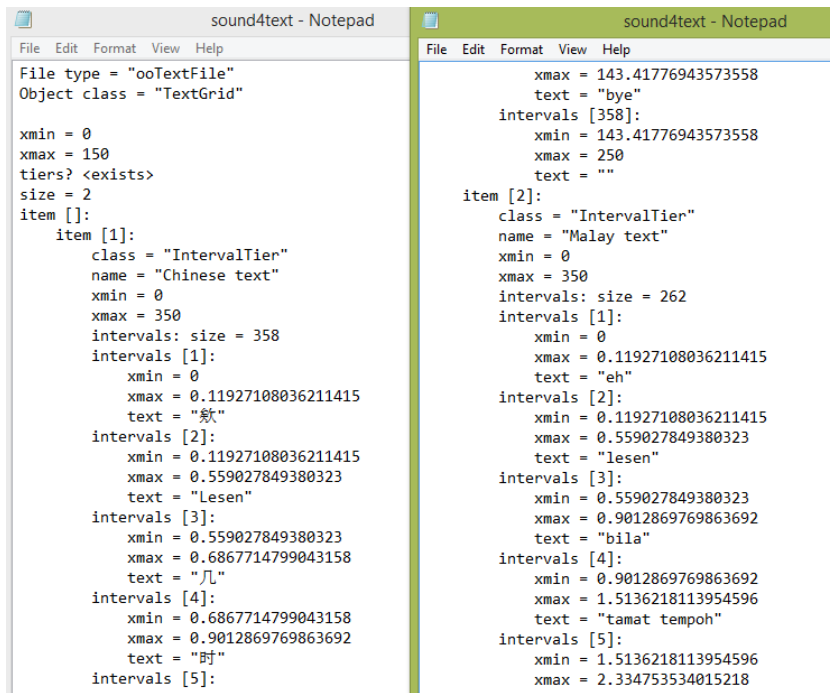


Figure 8: Text file that save the align word

All the Chinese text in text file are selected out and save in a new file, then the new file used to count the total word and total general words. In this research, jupyter notebook, python is used to do data analysis. The total words, total general words, total unique words are calculated and recorded as in Table 5 below. Percentage of unique words used is $(732/ 3390)*100 = 21.59\%$. The formula is percentage of unique word = total unique word / total word. So there is 21.59% is unique words.

Total number word	Total General word	Total Unique word	Unique word found
3390	2658	732	21.59%

Table 5: Record of word occur in dialogue


```

28 lines4 = [line4.replace('\n','')for line4 in lines4]
29 lines4 = [line4.replace(' ','')for line4 in lines4]
30 lines4 = [line4.replace(' ','')for line4 in lines4]
31 lines4 = ''.join(lines4)
32
33 line = lines1 + '\n' + lines2 + '\n' + lines3 + '\n' + lines4
34
35 with open('alldialog.txt','w',encoding='utf-8') as f:
36     f.writelines(line)
37
38 line

Out[48]: '欸回来了啊不是去买金纸nia怎么变还买batik布买来做什么哦我去做sarong么没晒你的裤我看到美籍batik就帮她买她讲要拿去做batik我今天看到这
条batik美就跟他讲哈他就tolong我先帮她买哦原来料刚才爸爸啊你了去哪里哦没看到他起来欸他去bagan做东西一下不要讲太多那我铺jioka来jio
kakham你铺下面下去啊知道了知道了铺好了那条jiokakham拿去洗衣机等一下洗啊好啊那条jioka布是哦喂等下我拿去你这tongkat拿去收起来放这边会
踢到你讲了不要乱乱放喂等下收等下收跟你讲了不是tongkat来这个讲讲这个是爬山用欸不是tongkat一直讲tongkat对了昨天下午大大雨大风我看那
间的iampan飞起来一些昨天啊是陪你爸爸讲昨晚起来时听到大雨声刚才看到盆沟水比平时高哦好菜他的iampan飞起来一些nia不然他就刺激料直接进
水进去不用怕二姐今天有回来哦喂她回alorstar哦回去二姐夫的妈妈家咯喂等下过penang去airitam你要去哦啊啊喂看先吃等下才跟你讲喂看电话不
要躺着看沙发坐着喂喂等下去我baru才搞nia等下等下喂喂喂喂的manik珠乱放没放进bottoi里面还是盒里面三姐的啥拿来做什么哦要做给
安奈班小孩鼓励他的哦三姐也是好料还做这些东西如果是我直接贴帖子就好了啊做到那么多懒惰明天我们去pasar直接吃早餐要去不要去suka你我跟你
去顺便买一些东西喂你要去买什么东西我要买背心嘛上一次跟你讲了啊你忘记买一直忘记ko啊baru我自己去买哈喂呀那样好喂你和我们一起去买喂我都
不知道你要什么size欸哈okokok料明天几点去啊明天啊明天我看就要8点哈不能太晚啊okok知道\n今天没煮啊晚上出去吃喂要去哪里吃不知道喂啊什
么都可以不知道那样去misya头那开的那个西餐店里吃哈你没去过欸要去吗呢可以pun啦我们去而已没晒三姐他们pun有去嘛她退休了开谈啊这个之前没听
你们讲过欸很久了啊还是baru而已她还喂喂ko嘛最近她和她丈夫两人一起开的嘛早上她去医院做misya晚上才去做西餐喂上一次去吃的时去吃的
时是不错哈下午看你要吃什么自己去买喂喂我要看十二点了啊我baru去买啊不等下很多人我应该是买tomyam啦不然就是买laksa喂我很久没吃laksaasam
laksa了你们要买给我喂你们喂你们要吃什么喂买哈你买sambal炒饭给我去顺便去kopi店打那个kopi热的回来啊你可以不是有买kopi粉的吗喂kopi粉
阿伯没卖了baru要喂kopi就要去外面kopi店买了喂料还要买哈爸爸喂喂你买白kopi给他哈吃的嘛pun买炒饭嘛巴拉那该喂喂你们的kopi水还有吃的来
料我还有买一点rojok你买laksa么刚才去看时laksa店还没开啊料就没买哈料你买什么tomyam面喂没有喂我买tomyam炒饭这样吃了才能饱一点
啊tomyam面等下下很快就喂料喂等下你没去airitam啊三四点时你去买那时都开料没就去转泉那边买siamlaksa喂siamlaksa上一次没听没讲
那边有开的baru才开而已喂好吃不开不久喂应该是好吃喂不然做人那么多喂喂那么等下看看先喂喂喂前面的laksa没卖我去那边买来买来吃喂喂刚才
我去楼下aunty那拿一点的lumpandan你后天要回金宝的时记得拿回去料记得放在这个衣橱里面喂喂才不能来喂喂上一次喂喂里知道那个喂喂会在橱里
面就都关着喂我回来时都有check过了喂都有看过料欸都没没开嘛喂喂知道回去时一开衣橱nia嘛突然飞出来吓到我一下重我拿全部的衣洗过跟不

```

```

In [56]: 1 print('length of sound1: ' + str(len(lines1)))
2 print('length of sound2: ' + str(len(lines2)))
3 print('length of sound3: ' + str(len(lines3)))
4 print('length of sound4: ' + str(len(lines4)))
5 print('total word: ' + str(len(line)))

length of sound1: 724
length of sound2: 1311
length of sound3: 818
length of sound4: 534
total word: 3390

```

Figure 9: Calculate length of total word

```

1 import re
2 #AllData.txt file is combine of data1,data2,data3 and data4 file
3 with open('AllData.txt','r',encoding='utf-8')as f:
4     lst=[]
5     dt = f.readline()
6     while dt != '':
7         lst.append(dt)
8         dt = f.readline()
9
10 word = ''.join(map(str,lst))
11 word_lst = word.split()
12
13 def replace_content(dict_replace, target):
14     """Based on dict, replaces key with the value on the target."""
15
16     for check, replacer in list(dict_replace.items()):
17         target = re.sub(check, replacer, target)
18
19     return target
20
21
22 # check : replacer
23 dict_replace = {
24     'air itam':'',
25     'alamak':'',
26     'alorstar':'',
27     'asam':'',
28     'aunty':'',
29     'bagan':'',
30     'baru':'',

```

Figure 10: Function of remove the unique word

```

108 wordlength = [wordlen.replace('\n','')for wordlen in wordlength]
109 wordlength = [wordlen.replace(' ','')for wordlen in wordlength]
110 wordlength = [wordlen.replace(' ','')for wordlen in wordlength]
111 wordlength = [wordlen.replace(' ','')for wordlen in wordlength]
112 wordlength = ''.join(wordlength)
113 wordlength

finish

Out[54]: *效回来了啊不是去买金金纸怎么还买买来做什么嘛拿去做什么没啦你的姨叫我看到美宿就帮她买她讲要拿去做什么农今天看到这条美就跟她讲陪她就我先帮
她买哦原来料刚才爸爸啊放你去了哪里哦没看到他起来效他去做东西一下不要讲太多帮我铺来你铺下面下去啊知道了知道了铺好了那茶拿去洗衣机等一
下洗啊好啊那布是哦嗯等下我拿去你这拿去收起来放这边会踢到你讲了不要乱乱放嗯等下收等下收跟你讲了不是来这个讲讲好菜他的飞起来一些不然
一直讲对了昨天天下很大雨大风我看那间的飞起来一些昨天啊是啥听你爸讲昨晚起来时听到大雨声刚才看到釜沟水比平时高哦好菜的飞起来一些不然
他就刺激料直接进入水进去不用睡二姐今天有回来哦没她回哦回去二姐夫的妈妈家哈哈等下过去你要去哦啊啊啊看先啦等下才跟你讲电话不要躺着
看去看啊啊等下去我才躺等下去等下吧啊啊谁的珠乱乱放没收进里面还是盒里面三姐的啥拿去做什么哦要做给安亲班小孩鼓励他的哦三姐也是
好料还做这些东西如果是我直接买贴子就好了啊做到那么多懒惰明天我们有去直接吃早餐要去不去你我跟你去顺便买一些东西哦你要去买什么东西我
要买背心啦一次跟你讲了啊你忘记买一直忘记啊我自己去买哈呀那样好啦你和我们一起去买啦我都不知道你要什么款哈哈料明天几点去啊明天啊明
天我看就要八点哈不能太晚啊知道今天没煮啊晚上出去吃啊要去哪里吃不知道啊啊什么都可以不知道啊啊那样去头那开的那个西餐店里吃哈你没去过款要
去吗呢可以啦我们去而已没啦三姐他们有去嘛她退休了开款啊这个之前没听你们讲过款很久了啊还是而已她还退休啦最近她和她丈夫两人一起开的啦
早上她去医院做晚上放工了才去做西餐哦上一次去吃的时去吃的时是不错哈下午看你要吃什么自己去买哈吧我看要十二点了啊我去买啊不等很多人我
应该是买啦不然就是买哦我很久没吃了你们要买给你们吃你们要吃什么哦买哈你买炒饭给我料去顺便去店打包那个热的回来啊你可以不是有买粉的卖粉
阿伯没卖了要喝就要去外面店买了哦料还要买给爸爸哦要哈你买白给他哈哈的勤炒炒饭哈巴拉哈哈啊啊你们的水还有吃的来料我还有买一点你没买么刚
才去看时店还没开料就没没哈哈你买什么面呀没哈我买炒饭这样吃了才能饱一点啊等等下下很快就快料啊等下你去看三三四点时你去买
哈那时都开料没就要去钟灵那边买哈上一次没听没讲那边有开的才开而已就好吃不开不久啦应该是好吃啦不然做人那么多哦哦那么等下看看先啦
啊啊前面的没卖我去那边买来吃哈看哈哈才我去楼下那拿一点的你后天要回宝宝的时记得拿回去料记得放在这个衣橱里面啊哈才不能来喂喂上一次
哈呢知道那个蟑螂会在橱里面效都关了哈我回来时都有过了啊都有看过料效都没没开嘛哪里知道回去时一开衣橱啊突然飞出来吓到我一下害我
拿全部的衣洗过啊不然会有味道啊不能橱你有擦过哦要擦一下才可以有有擦有擦嘛一次啊料我就没用了啊我去去买桶桶来放我的衣比较容易那样它

```

```

In [60]: 1 print('total general word: ' + str(len(wordlength)))
          total word: 3390
          total general word: 2658

In [65]: 1 totalWord = 3390
          2 totalGeneralWord = 2658
          3 totalUniqueWord = totalWord - totalGeneralWord
          4 print('total general word: ' + str(totalUniqueWord))
          5
          6 #percentage of unique word = total unique word / total word * 100
          7 PercentageUniqueWord = (totalUniqueWord / totalWord)*100
          8 print('Percentage of unique word: ' + str(PercentageUniqueWord))

          total general word: 732
          Percentage of unique word: 21.5929203539823

```

Figure 11: Calculate length of total general word, total unique word and percentage unique word

The Table 6 below shows unique vocabulary list. Which have 69 words without redundancy.

Air Itam	Alamak	Alorstar
Asam	Aunty	Bagan
Baru	Batik	Batu
Belacan	Bottoi	Bulk
Bye	Check	Cincai
Coupon	Daddy	Galo (penutup sayur)
Garing	Hana (ya la)	Iampan
Jagung	Jamban	Jambu
Jari	Jiokakham	Jioka

Karipok	Khawin	Kopi
Laksa	Lesen	Lum Pandan
Manik	Mata	Misy (nurse)
Motor	Pasar	Penang
Coupon	Kalu	Pun
Ringgit	Rojak	Roti
Sabun	Sambai	Sampah
Sampai	Sarong	Paliah (cheap thing)
Size	Siam	Sofa
Sotong	Suka	Tahan
Test	Tolong	Tomyam
Tongkat	Tapi	Nia
Meh	Ok	Ko
Pun	Kari	o

Table 6: List of unique word

4.3 Determine Phoneme Set of PH

In this process, the information on pronunciation from Taiwan Minnan Dictionary (教育部臺灣閩南語常用詞辭典, 2011) and Penang Hokkien learning website (Speak Hokkien Campaign, 2017) are extracted out. After collecting the information, Romanization process is conducted.

Since Mandarin words will make the whole process becomes more complicated, all the Mandarin words will be romanized. To Romanise all the Mandarin words, it will change to PH sound pinyin. With this, it becomes much easier. Below are the short example of Mandarin words and PH pinyin.

Mandarin word	PH pinyin
今天没有煮啊晚上出去吃啊。要去哪里吃?	kin jit bo tsu a am mee tshut khi jiah a. ia khi toh ui jiah?
等一下过 penang 去 air itam 你要去哦?	teng tsit ee kue pi neng khi a i tam lu ia khi o?

Table 7: Example of mandarin word convert to PH word

In this research, the data used is not only single word. As using only single word may hard to show the meaning out. And, to make it clearer on each pinying meaning, the Chinese word will be added inside the documentation. But during generaton of phoneme process the Chinese words will be temporally removed. Figure 12 below shows the output of Romanization process.

```

1 啊 → a
2 猶未好 → abueho
3 鴨韻 → aham
4 抑無 → ahbo
5 抑會 → ahe
6 後落 → ahloh
7 鴨卵 → ahnui
8 鴨飯 → ahpuinn
9 鴨腿麵線 → ahthuimisuan
10 抑有 → ahu
11 押韻 → ahun
12 愛啊 → aia
13 愛面 → aibin
14 愛面鬼 → aibinkui
15 愛吼 → aihau
16 愛啦 → ailah
17 愛咧 → aileh
18 愛啞 → ailim
19 愛囉 → ailo
20 愛鑼 → ailui
21 愛毋愛 → aimai
22 愛勿 → aimai1
23 愛學 → aioh
24 愛靜 → aitsenn
25 愛喔 → aiwo
26 沃飯 → akpuinn
27 沃湯 → akthng
28 沃權 → aktsang
29 沃菜 → aktshai
30 亞確 → alung

```

Figure 12: Data of Penang Hokkien

The next step is to standardize the grapheme used in PH. PH does not have own general or formal grapheme. Therefore, Taiwanese Minnan pinying will be used to generate pinying for Penang because both are them are almost similar. Table 8 below shows the graphene of Taiwan Minnan (IPA) and PH.

IPA	p	p ^h	B	m	h	T	ŋ	ts	ts ^h
PH	p	ph	b	m	h	t	ng	ts	tsh

IPA	g	dz̃	s						
PH	g	j	s	d	f	r	sh	w	y

IPA	a	ə	ɔ		I	U	e	e
PH	a	o	oo	y	i	u	e	ee

IPA								
	m	n	ŋ	ĩ	p̃	t̃	h	k̃
PH	m	n	ng	nn	p	t	h	k

Table 8: Show the IPA and PH grapheme

For the following step, it is to craft the grapheme to grapheme rules for PH. To standardize the grapheme used in PH, there are some rules crafted from the PH learning website (Speak Hokkien Campaign, 2017). One of the rule is adding loanword grapheme that not occur in Taiwan Minnan. Such as d, f, r, sh, w, y into the graphene set. The next rule is, changing certain grapheme to other grapheme as they are not pronounced same as the Taiwan Minnan. For instance, changing e become ee, example: te (茶) become tee (茶). Those grapheme that need to be modified are shown in Table 9:

Original	Changed	Original	Changed
e	ee	bian	mian
iunn	ionn	mue	muai
ue	ua	ik	ek
ing	eng	gua	wa
nng	noo		

Table 9: Show the original and changed word or grapheme

The following rules are the simplification of some grapheme in PH. The purpose of these grapheme to grapheme conversion rules is to make it easier when generating grapheme to phoneme conversion rules. Below is the grapheme that will simplifies:

Original	Simplifies	Original	Simplifies
tsh	Q	kh	K
ts	q	oo	O

ph	P	er	R
sh	S	ee	E
th	T	nn	V
ng	N		

Table 10: Show the original and simplifies grapheme

The final step in this process is to generate out the pronunciation of PH.

- Grapheme to grapheme conversion
- Simplification grapheme
- Adding loanword
- All grapheme is categories into two category Consonant and Vowel
- Category Vowel is categories to 韵头(head),韵腹(middle),韵尾(tail)
- Not all the pinying start with consonant it may also start in vowel head or middle

In this step, the vocabularies are divided into three categories. First, it is unique words which obtained from the recoding and also the PH website (Speak Hokkien Campaign, 2017). The second category is those vocabularies that do no need to apply grapheme to grapheme conversion rules. And, the third category is those vocabularies that involve grapheme to grapheme conversion. The second category will direct move to process generate phonemes. The process of generating pronunciation of PH is shown from Figure 13 to Figure 26.

```

View  Insert  Cell  Kernel  Widgets  Help
Run  Code
1 tw = pd.read_csv('NormalWord.txt', sep="\t", names=["Word", "Pinying"], encoding="utf-8")
2 tw.sort_values(by=['Pinying'], inplace=True)
3 tw
4 with open("NormalData.txt", 'w', encoding='utf-8') as f:
5     f.write(tw.to_string(header = True, index = False))
6
7 tw=tw.drop(columns='Word')
8 tw.rename(columns={'Pinying':' '}, inplace=True)
9 tw
10 with open("NormalWordPY.txt", 'w', encoding='utf-8') as f:
11     f.write(tw.to_string(header = True, index = False))

```

Figure 13: Remove Chinese word

Menu	Menu	Menu
1 啊 → a	1 Word	1 a
2 猶未好 → abueho	2 啊	2 abueho
3 鴨額 → aham	3 猶未好	3 aham
4 抑無 → ahbo	4 鴨額	4 ahbo
5 抑會 → ahe	5 抑無	5 ahe
6 後落 → ahloh	6 抑會	6 ahloh
7 鴨卵 → ahnuí	7 後落	7 ahnuí
8 鴨飯 → ahpuínn	8 鴨卵	8 ahpuínn
9 鴨腿麵線 → ahthuimisuann	9 鴨飯	9 ahthuimisuann
10 抑有 → ahu	10 鴨腿麵線	10 ahu
11 押韻 → ahun	11 抑有	11 ahun
12 愛啊 → aia	12 押韻	12 aia
13 愛面 → aibin	13 愛啊	13 aibin
14 愛面鬼 → aibinkui	14 愛面	14 aibinkui
15 愛吼 → aihau	15 愛面鬼	15 aihau
16 愛啦 → ailah	16 愛吼	16 ailah
17 愛咧 → aileh	17 愛啦	17 aileh
18 愛琳 → ailim	18 愛咧	18 ailim
19 愛囉 → ailo	19 愛琳	19 ailo
20 愛貓 → ailui	20 愛囉	20 ailui
21 愛毋愛 → aimai	21 愛貓	21 aimai
22 愛勿 → aimail	22 愛毋愛	22 aimail
23 愛學 → aioh	23 愛勿	23 aioh
24 愛靜 → aitsenn	24 愛學	24 aitsenn

Figure 14: Changing of file from beginning to generate phoneme type

Make changing of the file is to make easier for the later combination of the phoneme together with the word and pinying. The changing process is the original data will be save into table form data and last will remove the Chinese word columns and left only pinying columns and using this latest text file put in the code to generate the phonemes out.

```

1 '''simplifies consonants and vowels'''
2 dict_replace = {
3     'tsh':'Q',
4     'ts':'q',
5     'ph':'P',
6     'sh':'S',
7     'th':'T',
8     'kh':'K',
9     'oo':'O',
10    'er':'R',
11    'ee':'E',
12    'nn':'V',
13    'ng':'N'
14 }
15
16 new_content = replace_content(dict_replace, s)
17 new_content = new_content.split()
18
19 '''grapheme of Penang Hokkien'''
20 '''consonants'''
21 consonants = ['p', 'P', 'm', 'b', 't', 'T', 'n', 'l', 'k', 'K', 'N', 'g', 'q', 'Q', 's', 'j', 'h']
22 '''loanword'''
23 added_c = ['d', 'f', 'r', 'S', 'w', 'y', 'x', 'z']
24 consonants = consonants + added_c
25
26 '''vowels and loanword vowel'''
27 韻頭 = ['i', 'u']
28 韻腹 = ['a', 'e', 'i', 'o', 'u', 'm', 'N', 'o', 'Z']
29 韻尾 = ['i', 'u', 'm', 'N', 'p', 't', 'k', 'h', 'n', 'y', 'R', 'E', 'V']

```

Figure 15: Simplifies and add consonant and vowels

```

'''generate phonemes'''
def seg(new_content):
    news = new_content[:]
    res = " "

    while len(news) != 0:
        for c in consonants:
            if news.startswith(c):
                res = res + c
                news = news[len(c):]
                if len(news) != 0:
                    res = res + "|聲 "
            else:
                res = res + '|聲 ' #+ s[-1]

        if len(news) != 0:
            for v1 in 韻頭:
                if news.startswith(v1):
                    res = res + v1
                    news = news[len(v1):]
                    if len(news) != 0:
                        res = res + "|頭 "
                    else:
                        res = res + '|頭 ' #+ s[-1]

        if len(news) != 0:
            for v2 in 韻腹:
                if news.startswith(v2):
                    res = res + v2
                    news = news[len(v2):]
                    if len(news) != 0:
                        res = res + "|腹 "
                    else:
                        res = res + '|腹 ' #+ s[-1]

    return res

```

Figure 16: Function to generate phonemes for non-changing general word

```

1 data = pd.read_csv('NormalData.txt',encoding='utf-8')
2 data2 = pd.read_csv('NormalWordDict.txt', sep=",", names=["PinYing","Phonemes"], encoding="utf-8")
3 data2 = data2.drop(columns='PinYing')
4
5 df = pd.concat([data,data2], axis=1)
6 df
7 with open("NormalWordphoneme.txt", 'w', encoding='utf-8') as f:
8     f.write(df.to_string(header = True, index = False))
9
10 print('finish')

```

finish

Figure 17: Save as into text file

Word	Pinyin	Phonemes
啊	a	a 腹
猶未好	abueho	a 腹 b 聲 u 頭 e 腹 h 尾 o 腹
鴨額	aham	a 腹 h 尾 a 腹 m 腹
抑無	ahbo	a 腹 h 尾 b 聲 o 腹
抑會	ahē	a 腹 h 尾 e 腹
後落	ahloh	a 腹 h 尾 l 聲 o 腹 h 尾
鴨卵	ahnui	a 腹 h 尾 n 尾 u 頭 i 腹
鴨飯	ahpuinn	a 腹 h 尾 p 聲 u 頭 i 腹 V 尾
鴨腿麵線	ahthuisuann	a 腹 h 尾 T 聲 u 頭 i 腹 m 腹 i 尾 s 聲 u 頭 a 腹 V 尾
抑有	ahu	a 腹 h 尾 u 頭
押韻	ahun	a 腹 h 尾 u 頭 n 尾
愛啊	aia	a 腹 i 腹 a 腹
愛面	aibin	a 腹 i 腹 b 聲 i 頭 n 尾
愛面鬼	aibinkui	a 腹 i 腹 b 聲 i 頭 n 尾 k 聲 u 頭 i 腹
愛吼	aiahau	a 腹 i 腹 h 尾 a 腹 u 腹
愛啦	ailah	a 腹 i 腹 l 聲 a 腹 h 尾
愛咧	aileh	a 腹 i 腹 l 聲 e 腹 h 尾
愛咁	ailim	a 腹 i 腹 l 聲 i 頭 m 腹
愛囉	ailo	a 腹 i 腹 l 聲 o 腹
愛播	ailui	a 腹 i 腹 l 聲 u 頭 i 腹
愛毋愛	aimai	a 腹 i 腹 m 腹 a 腹 i 腹
愛勿	aimail	a 腹 i 腹 m 腹 a 腹 i 腹
愛學	aioh	a 腹 i 腹 o 腹 h 尾
愛諍	aitsenn	a 腹 i 腹 q 聲 e 腹 V 尾
愛喔	aiwo	a 腹 i 腹 w 聲 o 腹
沃飯	akpuinn	a 腹 k 尾 p 聲 u 頭 i 腹 V 尾

Figure 18: PH phonemes for second set data (non-changing general word)

The next step is handle third category of vocabularies. First, it is to apply grapheme to grapheme conversion rules. After that, the vocabularies from category 2 and 3 are combined to generate the pronunciation in PH. Finally, pronunciation dictionary of PH that contains vocabularies from category 2 and 3 is developed.

```

1  '''gua,nng,mue,ue,ik'''
2  with open('ChgWrdPV.txt','r', encoding = 'utf-8')as f:
3      byte_lst = []
4
5      byte = f.readline()
6      while byte != '':
7          byte_lst.append(byte)
8          byte = f.readline()
9
10     s = ''.join(map(str,byte_lst))
11     char_lst = s.split()
12
13     def replace_content(dict_replace, target):
14         """Based on dict, replaces key with the value on the target."""
15
16         for check, replacer in list(dict_replace.items()):
17             target = re.sub(check, replacer, target)
18
19         return target
20
21
22     '''changing grapheme'''
23     dict_replace = {
24         'gua': 'wa',
25         'nng': 'nnoo',
26         'mue': 'muai',
27         'ue': 'ua',
28         'ik': 'ek'

```

Figure 19: Function of changing the grapheme

```

for key, val in res.items():
    writer.writerow([key.encode('utf-8'), val])

print("finished")

```

ished

```

Chgdt = pd.read_csv('ChgWrdData4.txt', encoding="utf-8")

#create a table like file
Chgdt4 = pd.read_csv('CWPY_all.txt', sep=",", names=['py', 'Changed Pinying'], encoding='utf-8')
Chgdt4=Chgdt4.drop(columns='py') #remove py columns

#combine data
data4 = pd.concat([Chgdt, Chgdt4], axis=1)
data4
'''duplicate the data for generate phonemes'''
ChgH = Chgdt4
ChgH

#save completed changing pinying into file
with open("ChangedPinying_all.txt", 'w', encoding='utf-8') as f:
    f.write(data4.to_string(header = True, index = False))

#save only pinying use for generate phonemes
with open("ChangedPinying_onlyPY.txt", 'w', encoding='utf-8') as f:
    f.write(ChgH.to_string(header = False, index = False))

```

Figure 20: Saving all changed data in a file and prepare data for generate phonemes

 Jupyter ChangedPinying_all.txt ✓ a day ago

File	Edit	View	Language
1	Word	Pinying	Changed Pinying
2	愛咩	aimee	aimee
3	愛情	aitseng	aitseng
4	愛情批	aitsengphe	aitsengphe
5	沃花	akhua	akhua
6	暗暝	ammee	ammee
7	暗暝工	ammeekang	ammeekang
8	暗暝頓	ammeetuinn	ammeetuinn
9	暗色	amsek	amsek
10	紅蝦	anghee	anghee
11	艇公仔冊	angkongatsheeh	angkongatsheeh
12	艇公間	angkongkeng	angkongkeng
13	紅毛兵	angmoopeng	angmoopeng
14	紅毛茶	angmootee	angmootee
15	紅毛冊	angmootsheeh	angmootsheeh
16	紅毛話	angmooua	angmooua
17	紅色	angsek	angsek
18	紅茶	angtee	angtee
19	安家	ankee	ankee
20	按脈	anmeeh	anmeeh
21	按呢	annee	annee
22	按呢款	anneekhuan	anneekhuan
23	安定	anteng	anteng
24	案情	antseng	antseng
25	按怎樣	antsuanniunn	antsuanniunn
26	阿伯	apeeh	apeeh
27	壓迫	appek	appek
28	阿丈	atiunn	ationn
29	阿叔	atsek	atsek

Figure 21: Changed all grapheme text file

jupyter ChangedPinying_onlyPY.txt

```
File Edit View Language
1 aimee
2     aitseng
3     aitsengphe
4     akhua
5     ammee
6     ammeekang
7     ammeetuinn
8     amsek
9     anghee
10    angkongatsheeh
11    angkongkeng
12    angmoopeng
13    angmootee
14    angmootsheeh
15    angmooua
16    angsek
17    angtee
18    ankee
19    anmeeh
20    annee
```

Figure 22: Changed grapheme file that only have pinying use to generate phonemes

After complete changing the grapheme will move to next step which is generate it own phoneme. The next step process is similar as previous show in the second set data (non-changing grapheme).

```
1 """Based on dict, replaces key with the value on the target."""
2
3 for check, replacer in list(dict_replace.items()):
4     target = re.sub(check, replacer, target)
5
6 return target
7
8
9
10
11 #simplifies the grapheme list
12 dict_replace = {
13     'tsh': 'Q',
14     'ts': 'q',
15     'ph': 'P',
16     'sh': 'S',
17     'th': 'T',
18     'kh': 'K',
19     'oo': 'O',
20     'er': 'R',
21     'ee': 'E',
22     'nn': 'V',
23     'ng': 'N'
24 }
25
26 new_content = replace_content(dict_replace, s)
27 new_content = new_content.split()
28 #print(new_content)
29
30 #consonant and vowels
31 consonants = ['p', 'P', 'm', 'b', 't', 'T', 'n', 'l', 'k', 'K', 'N', 'g', 'q', 'Q', 's', 'j', 'h']
32 #added_c is loanword consonant
33 added_c = ['d', 'f', 'r', 'S', 'w', 'y', 'x', 'z']
34 consonants = consonants + added_c
35
36 #vowels
37 韵头 = ['i', 'u']
38 韵腹 = ['a', 'e', 'i', 'O', 'u', 'm', 'N', 'o', 'Z']
39 韵尾 = ['i', 'u', 'm', 'N', 'p', 't', 'k', 'h', 'n', 'y', 'R', 'E', 'V']
```

Figure 23: Function of simplifies grapheme and add consonant and vowel

```

v Insert Cell Kernel Widgets Help
Run Code
#function of generate phonemes
def seg(new_content):
    news = new_content[:]
    res = ""

    while len(news) != 0:
        for c in consonants:
            if news.startswith(c):
                res = res + c
                news = news[len(c):]
            if len(news) != 0:
                res = res + "|聲"
            else:
                res = res + '|聲' #+ s[-1]

        if len(news) != 0:
            for v1 in 韻頭:
                if news.startswith(v1):
                    res = res + v1
                    news = news[len(v1):]
                if len(news) != 0:
                    res = res + "|頭"
                else:
                    res = res + '|頭' #+ s[-1]

        if len(news) != 0:
            for v2 in 韻腹:
                if news.startswith(v2):
                    res = res + v2
                    news = news[len(v2)+1:]

```

Figure 24: Generate phonemes for changed pinying (grapheme)

upyer ChangedPYphoneme.txt a day ago

Word	Pinying	Phonemes
愛咩	aimee	a 腹 i 腹 m 腹 E 尾
愛情	aitseng	a 腹 i 腹 q 聲 e 腹 N 腹
愛情批	aitsengphe	a 腹 i 腹 q 聲 e 腹 N 腹 P 聲 e 腹
沃花	akhua	a 腹 k 聲 u 頭 a 腹
暗暎	amnee	a 腹 m 腹 m 尾 E 尾
暗暎工	amneekang	a 腹 m 腹 m 尾 E 尾 k 聲 a 腹 N 腹
暗暎頓	ammeetuinn	a 腹 m 腹 m 尾 E 尾 t 聲 u 頭 i 腹 V 尾
暗色	amsek	a 腹 m 腹 s 聲 e 腹 k 尾
紅蝦	anghee	a 腹 N 腹 h 尾 E 尾
姪公仔冊	angkongatsheeh	a 腹 N 腹 k 尾 o 腹 N 尾 a 腹 Q 聲 E 尾
姪公間	angkongkeng	a 腹 N 腹 k 尾 o 腹 N 尾 k 尾 e 腹 N 腹
紅毛兵	angmoopeng	a 腹 N 腹 m 尾 o 腹 p 尾 e 腹 N 腹
紅毛茶	angmootee	a 腹 N 腹 m 尾 o 腹 t 尾 E 尾
紅毛冊	angmootsheeh	a 腹 N 腹 m 尾 o 腹 Q 聲 E 尾 h 聲
紅毛話	angmoooua	a 腹 N 腹 m 尾 o 腹 u 腹 a 腹
紅色	angsek	a 腹 N 腹 s 聲 e 腹 k 尾
紅茶	angtee	a 腹 N 腹 t 尾 E 尾
安家	ankee	a 腹 n 尾 k 聲 E 尾
按脈	anmeeh	a 腹 n 尾 m 聲 E 尾 h 聲
按呢	annee	a 腹 V 尾 E 尾
按呢款	anneekhuan	a 腹 V 尾 E 尾 k 聲 u 頭 a 腹 n 尾
安定	anteng	a 腹 n 尾 t 聲 e 腹 N 腹
案情	antseng	a 腹 n 尾 q 聲 e 腹 N 腹
按怎樣	antsuannionn	a 腹 n 尾 q 聲 u 頭 a 腹 V 尾 i 頭 o 腹 V 尾
阿伯	apeeh	a 腹 p 尾 E 尾 h 聲
壓迫	appek	a 腹 p 尾 p 聲 e 腹 k 尾
阿丈	ationn	a 腹 t 尾 i 頭 o 腹 V 尾
阿叔	atsek	a 腹 q 聲 e 腹 k 尾
後屋話	auhuaa	a 腹 u 腹 h 聲 u 頭 a 腹 u 腹 a 腹

Figure 25: Result for the changed pinying phonemes

```

啊→a→a|腹
猶未好→abueho→a|腹b|聲u|頭e|腹h|尾o|腹
鴨領→aham→a|腹h|尾a|腹m|腹
抑無→ahbo→a|腹h|尾b|聲o|腹
抑會→ahe→a|腹h|尾e|腹
後落→ahloh→a|腹h|尾l|聲o|腹h|尾
鴨卵→ahnui→a|腹h|尾n|尾u|頭i|腹
鴨飯→ahpuinn→a|腹h|尾p|聲u|頭i|腹v|尾
鴨腿麵線→ahthuisuann→a|腹h|尾T|聲u|頭i|腹m|腹i|尾s|聲u|頭a|腹v|尾
抑有→ahu→a|腹h|尾u|頭
押韻→ahun→a|腹h|尾u|頭n|尾
愛啊→aia→a|腹i|腹a|腹
愛面→aibin→a|腹i|腹b|聲i|頭n|尾
愛面鬼→aibinkui→a|腹i|腹b|聲i|頭n|尾k|聲u|頭i|腹
愛吼→aihau→a|腹i|腹h|尾a|腹u|腹
愛啦→ailah→a|腹i|腹l|聲a|腹h|尾
愛咧→aileh→a|腹i|腹l|聲e|腹h|尾
愛啾→ailim→a|腹i|腹l|聲i|頭m|腹
愛囉→ailo→a|腹i|腹l|聲o|腹
愛鑼→ailui→a|腹i|腹l|聲u|頭i|腹

```

Figure 26: Result of combined changed pinying phonemes and non-changed pinying phonemes

4.4 Generate Pronunciation modeling of PH

The last process is combining of the unique vocabularies and general vocabularies in pronunciation dictionary. As the rules included the loanword grapheme, the unique vocabularies from category one will be using the same function in the code to generate their pronunciation. Finally, a pronunciation dictionary in PH is developed which includes general vocabularies and unique vocabularies found in this research.

```

1 Uw = pd.read_csv('UniqueWord.txt', sep='\t', names=['UniqueWord', 'UniqueWord Pinying'], encoding='utf-8')
2 Uw = Uw.drop(columns='UniqueWord')
3
4 with open("UniqueWord_py.txt", 'w', encoding='utf-8') as f:
5     f.write(Uw.to_string(header = False, index = False))

```

```

1 '''uniqueword'''
2 with open('UniqueWord_py.txt', 'r', encoding = 'utf-8') as f:
3     byte_lst = []
4
5     byte = f.readline()
6     while byte != '':
7         byte_lst.append(byte)
8         byte = f.readline()
9
10 s = ''.join(map(str, byte_lst))
11 char_lst = s.split()
12 #function to simplifies grapheme
13 def replace_content(dict_replace, target):
14     """Based on dict, replaces key with the value on the target."""
15
16     for check, replacer in list(dict_replace.items()):
17         target = re.sub(check, replacer, target)
18
19     return target
20
21
22 # List of simplifies grapheme
23 dict_replace = {
24     'tsh': 'Q',
25     'ts': 'q',
26     'ph': 'P',
27     'sh': 'S',
28     'th': 'T',
29     'kh': 'K',
30     'oo': 'O',
31     'en': 'R',

```

Figure 27: Read unique word file and apply simplified rule

```

        res = res + '|腹' * #+ s[-1]

    if len(news) != 0:
        for v3 in 韻尾:
            if news.startswith(v3):
                res = res + v3
                news = news[len(v3):]
                if len(news) != 0:
                    res = res + "|尾 "
            else:
                res = res + '|尾 ' * #+ s[-1]

    return res

res = {}
for i in range(0, len(char_lst)):
    phonemes = new_content[i]
    res[char_lst[i]] = seg(phonemes)

#save the result in text file
with open("UniquePyDict.txt", 'w', encoding='utf-8') as f:
    writer = csv.writer(f)
    for key, val in res.items():
        writer.writerow([key.encode('utf-8'), val])

print("finished")

```

```

'''Uwresult is data with chinese word, Uwresult2 is the file of result| generate phoneme'''
Uwresult = pd.read_csv('UniqueWord.txt', sep='\t', names=["word", "Pinying"], encoding='utf-8')
Uwresult2 = pd.read_csv('UniquePyDict.txt', sep=',', names=["Pinying", "Phonemes"], encoding='utf-8')
Uwresult2 = Uwresult2.drop(columns='Pinying')
'''combine the two data'''
UwresultDT = pd.concat([Uwresult, Uwresult2], axis=1)
UwresultDT
'''save the final result in data'''
with open("UniqueWordPhoneme.txt", 'w', encoding='utf-8') as f:
    f.write(UwresultDT.to_string(header = True, index = False))

```

Figure 28: Generate phonemes and save final result in text file

ipython PenangHokkienPhonemesDictionary.txt a day ago

Word	Pinying	Phonemes
啊	a	a 腹
猶未好	abueho	a 腹 b 聲 u 頭 e 腹 h 尾 o 腹
鴨額	aham	a 腹 h 尾 a 腹 m 腹
抑無	ahbo	a 腹 h 尾 b 聲 o 腹
抑會	ahē	a 腹 h 尾 e 腹
後落	ahloh	a 腹 h 尾 l 聲 o 腹 h 尾
鴨卵	ahnuī	a 腹 h 尾 n 尾 u 頭 i 腹
鴨飯	ahpuinn	a 腹 h 尾 p 聲 u 頭 i 腹 V 尾
鴨腿麵線	ahthuimisuann	a 腹 h 尾 T 聲 u 頭 i 腹 m 腹 i 尾 s 聲 u 頭 a 腹 V 尾
抑有	ahu	a 腹 h 尾 u 頭
押韻	ahun	a 腹 h 尾 u 頭 n 尾
愛啊	aia	a 腹 i 腹 a 腹
愛面	aibin	a 腹 i 腹 b 聲 i 頭 n 尾
愛面鬼	aibinkui	a 腹 i 腹 b 聲 i 頭 n 尾 k 聲 u 頭 i 腹
愛吼	aihau	a 腹 i 腹 h 尾 a 腹 u 腹
愛啦	ailah	a 腹 i 腹 l 聲 a 腹 h 尾
愛咧	ailēh	a 腹 i 腹 l 聲 e 腹 h 尾
愛咁	ailim	a 腹 i 腹 l 聲 i 頭 m 腹
愛囉	ailo	a 腹 i 腹 l 聲 o 腹
愛鑼	ailui	a 腹 i 腹 l 聲 u 頭 i 腹
愛毋愛	aimai	a 腹 i 腹 m 腹 a 腹 i 腹
愛勿	aimai1	a 腹 i 腹 m 腹 a 腹 i 腹
愛學	aioh	a 腹 i 腹 o 腹 h 尾
愛靜	aitseñ	a 腹 i 腹 q 聲 e 腹 V 尾
愛啞	aiwo	a 腹 i 腹 w 聲 o 腹
沃飯	akpuinn	a 腹 k 尾 p 聲 u 頭 i 腹 V 尾
沃湯	akthng	a 腹 k 尾 T 聲 N 聲
沃攏	aktsang	a 腹 k 尾 q 聲 a 腹 N 腹
沃菜	aktshai	a 腹 k 尾 o 聲 a 腹 i 腹

Figure 29: Penang Hokkien Pronunciation Dictionary

Chapter 5: Data Analysis

In this chapter, data analysis will be discussed. First, 10 set of testing samples are generated. Each set consists of 100 vocabularies. The accuracy, in term of word accuracy and phonemes accuracy are evaluated.

```
import random
'''sample'''
PH = pd.read_csv('PH_word.txt', sep='\t', names=["Word", "Pinying", "Phonemes"], encoding='utf-8')
sampled_list1 = PH.sample(n = 100)
sampled_list2 = PH.sample(n = 100)
sampled_list3 = PH.sample(n = 100)
sampled_list4 = PH.sample(n = 100)
sampled_list5 = PH.sample(n = 100)
sampled_list6 = PH.sample(n = 100)
sampled_list7 = PH.sample(n = 100)
sampled_list8 = PH.sample(n = 100)
sampled_list9 = PH.sample(n = 100)
sampled_list10 = PH.sample(n = 100)

sampled_list1

with open("sample1.txt", 'w', encoding='utf-8') as f:
    f.write(sampled_list1.to_string(header = True, index = False))

sampled_list2

with open("sample2.txt", 'w', encoding='utf-8') as f:
    f.write(sampled_list2.to_string(header = True, index = False))

sampled_list3

with open("sample3.txt", 'w', encoding='utf-8') as f:
```

Figure 30: Generate 10 set of testing sample with each 100 pronunciation

	A	B	C	D	E	F	G
78	q 聲 腹 聲 頭 尾 聲 腹 腹	做人情	tsojintseng	q o j i n q e N	0	0	
79	b 聲 腹 腹 腹 腹 尾	袂用	beiong	b e i o N	0	0	
80	k 聲 頭 腹 腹 尾 聲 腹 尾	懸低	kuankee2	k u a n k E	0	0	
81	t 聲 頭 腹 腹 聲 頭 腹	大水	tuatsui	t u a q u i	0	0	
82	m 聲 腹 腹 尾 腹	嘛好	maho	m a h o	0	0	
83	p 聲 尾 腹 聲 腹 腹 尾 腹 頭	白木耳	peehbokni	p E h b o k n i i	0	0	
84	b 聲 頭 腹 腹 腹 腹 腹 腹	舞弄獅	bulangsai	b u i a N s a i i	0	0	
85	j 聲 頭 尾 腹 聲 頭 腹	入鑪	jiplui	j i i p i u i i	0	0	
86	u 頭 腹 腹 尾 腹 腹 尾 聲 頭	晏晏起	uannuannkhi	u a V u a V K i i	0	0	
87	p 聲 腹 腹 腹 尾 腹	平和	pengho1	p e N h o	0	0	
88	h 聲 腹 腹 腹 聲 腹 尾	鹹濕	hamsap	h a m s a p	0	0	
89	t 聲 頭 腹 尾 腹 腹 腹 尾	對中	tuitiong	t u i t i o N	0	0	
90	l 聲 腹 腹 腹 聲 頭	老茨	lautshu	l a u Q u	0	0	
91	t 聲 腹 腹 腹 尾 腹 腹 尾	燈光	tengkuinn	t e N k u i V	0	0	
92	Q 聲 頭 尾 聲 尾 聲 腹	鵝家媽	tshinkeema	Q i n k E m a	0	0	
93	q 聲 頭 腹 腹 尾 聲 頭 腹 腹	轉數	tsuansiau	q u a n s i a u	0	0	
94	t 聲 頭 腹 腹 尾 聲 頭 聲 頭 腹 腹 尾	電子廠	tiantstshionn	t i a n q u Q i o V	0	0	
95	s 聲 頭 尾 聲 腹 腹	失明	sitbeng	s i t b e N	0	0	
96	t 聲 頭 腹 腹 聲 腹 尾 腹 腹 腹	大細片	tuasepeng	t u a s e p e N	0	0	
97	l 聲 頭 尾 聲 腹 尾	恁爸	linpee	l i i n p E	0	0	
98	p 聲 腹 尾	魄	phek3	p e k	0	0	
99	q 聲 頭 尾 腹 腹 尾 聲 頭 腹 腹 尾	自動門	tsutongmuinn	q u t o N m u i V	0	0	
100	s 聲 腹 腹 腹 腹 尾	霜角	sngkak	s N k a k	0	0	
101	i 頭 腹 腹 尾 聲 頭	用處	iongtshu	i o N Q u	0	0	
102			word =	100	1	0.256081946	
103			phoneme =	781			

Figure 31: Sample of PH Vocabularies with Pronunciation

The formula for word accuracy is (total correct words / total words)*100%, and the formula for phoneme accuracy is (total correct phonemes/ total phonemes)*100%.

Sample	Word	Phoneme (characters)	Word error	Phoneme error	Word accuracy(%)	Phoneme accuracy (%)
Sample1	100	756	0	0	100	100
Sample2	100	745	0	0	100	100
Sample3	100	781	1	2	99	99.74
Sample4	100	773	1	2	99	99.74
Sample5	100	772	0	0	100	100
Sample6	100	776	0	0	100	100
Sample7	100	789	2	4	98	99.49
Sample8	100	771	2	4	98	99.42
Sample9	100	781	1	2	99	99.75
Sample10	100	785	1	2	99	99.75

Table 11: Show the percentage of each sample WER and PER

From the table, it show that the highest error rate of word is 2% and highest error rate of phonemes is 0.58%. The average word accuracy achieves 99.2% while phoneme accuracy give 99.79%. Therefore, the accuracy of the pronunciation dictionary developed is high.

Chapter 6: Conclusion

In the nutshell, this project is Pronunciation modelling for Penang Hokkien (PH). The objectives in this project are to identify the unique vocabularies in PH, to identify the phonemes set in PH and to develop pronunciation dictionary for PH. The motivation to this project is let the youngest generation learns to speak and understand PH in order to prevent PH becomes extinct language in future.

The contribution of this project is to develop PH pronunciation dictionary, which may useful for speech recognition system or text-to-speech synthesis system. For the future work, by applying the PH pronunciation dictionary developed, acoustic model in PH can be trained in order to synthesize speech in PH. It will be useful for those who like to learn PH and also to help preserving the dialect and culture in it.

REFERENCE

Chai, L.T., 2011. Culture heritage tourism engineering at Penang: Complete the puzzle of “the pearl of orient”. *Systems Engineering Procedia*, 1, pp.358-364.

Teoh, B.S., Lim, B.S. and Lee, L.H., 2017. A study of Penang Peranakan Hokkien. *Journal of Modern Languages*, 15(1), pp.169-181.

Tan, T.P. and Ranaivo-Malançon, B., 2009. Malay grapheme to phoneme tool for automatic speech recognition. In *Proc. Workshop of Malaysia and Indonesia Language Engineering (MALINDO) 2009*.

Chuang, C.T., Chang, Y.C. and Hsieh, F.F., 2013. Complete and not-so-complete tonal neutralization in Penang Hokkien. *Proc. of ICPLC*, pp.54-57.

Tang, Y.L., 2016. The Difference between Taiwan Min-nan Dialect and Fujian Min-nan Dialect: Borrowed Words from Japanese in Taiwan Min-nan Dialect. *International Journal of Language and Linguistics*, 3(2), pp.82-86.

Rao, K., Peng, F., Sak, H. and Beaufays, F., 2015, April. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4225-4229). IEEE.

Yeong, Y.L. and Tan, T.P., 2011, November. Applying grapheme, word, and syllable information for language identification in code switching sentences. In *2011 International Conference on Asian Language Processing* (pp. 111-114). IEEE.

Teh, C.Y. and Lim, Y.M., 2014. An alternative architectural strategy to preserve the living heritage and identity of Penang Hokkien language in Malaysia. *Int J Hum Soc Sci*, 4, pp.242-247.

Hing, J.W., 2018. THE POLYFUNCTIONAL FOCUS PARTICLE PUN53 IN PENANG HOKKIEN: A CONTACT PERSPECTIVE. *JSEALS*, p.51.

Kabir, Syed Muhammad. 2016. Method of Data Collection. In *An Introductory Approach for All Disciplines, Edition: First, Chapter: 9, Publisher: Book Zone Publication, Chittagong-4203, Bangladesh*, pp.201-275

Magistry, P., 2016. Design of an Input Method for Taiwanese Hokkien using Unsupervised Word Segmentation for Language Modeling. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016)* (pp. 284-298).

LJOLJE et al, 2010 Patent Application Publication

Hassan, Hsiao. R, Lane. I, Alan W. Black, Waibel. A Pronunciation Modelling for Dialectal Arabic Speech Recognition

Liang, M.-S., J.-C. Yang, Y.-C. Chiang, and R.-Y. Lyu, "A Taiwanese Text-to-Speech System with Applications to Language Learning," *Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies*, 2004, pp. 91-95.

Iunn, U. G., Lau, K. G., Tan-Tenn , H. G., Lee, S. A., and Kao, C. Y., "Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 12, No. 4, December 2007, pp. 349-370, 2007.

Speakhokkien.org. 2017. *Speak Hokkien Campaign*. [online] Available at: <<https://www.speakhokkien.org/guan-yu-wo-men>> [Accessed 23 April 2020].

臺灣閩南語常用詞辭典. 2011. *教育部臺灣閩南語常用詞辭典*. [online] Available at: <https://twblg.dict.edu.tw/holodict_new/index.html> [Accessed 23 April 2020].

Penang Travel Tips. 2020. *International Phonetic Alphabet (IPA)*. [online] Available at: <<https://www.penang-traveltips.com/hokkien/international-phonetic-alphabet.htm>> [Accessed 23 April 2020].

Factsanddetails.com. 2019. *PEOPLE, POPULATION AND LANGUAGES OF MALAYSIA | Facts And Details*. [online] Available at: <http://factsanddetails.com/southeast-asia/Malaysia/sub5_4b/entry-3153.html> [Accessed 23 April 2020].

Authôt. n.d. *Automatic Speech Recognition System | Voice Recognition | Authôt*. [online] Available at: <<https://www.authot.com/en/2016/09/09/automatic-speech-recognition-system/>>.

WhatIs.com. 2005. *What Is Speech Synthesis? - Definition From Whatis.Com*. [online] Available at: <<https://whatis.techtarget.com/definition/speech-synthesis>>.



PRONUNCIATION MODELLING FOR PENANG HOKKIEN

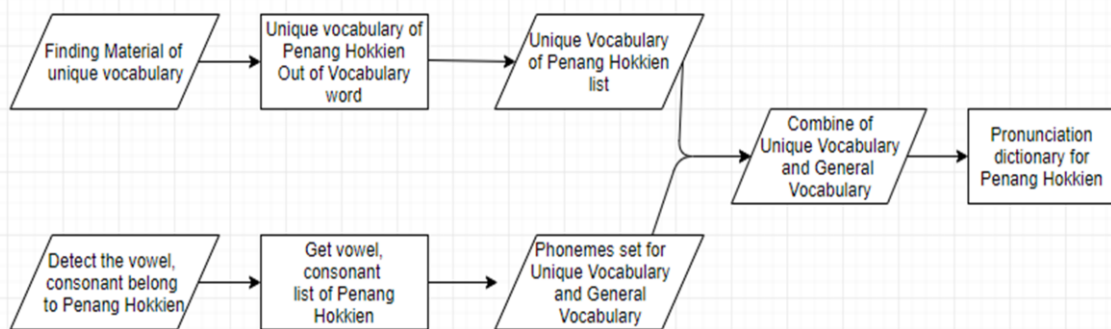
Introduction:

Hokkien is a dialect of Chinese. There are several Hokkien dialect in Malaysia, Penang Hokkien (PH) is one of the dialects. Used the most in Penang, Perlis, and Kedah. PH use by many state but it does not have formal pronunciation and also writing system, unlike Taiwan. This project is to develop pronunciation dictionary for PH. Pronunciation dictionary generated will helpful in building PH text-to-speech (TTS) system.

Objectives:

- To identify the unique vocabularies in Penang Hokkien.
- To identify the phonemes set in Penang Hokkien.
- To develop pronunciation dictionary for Penang Hokkien.

Methodology



RESULT:

Sample	Word error	Phoneme error	Word accuracy (%)	Phoneme accuracy (%)
Sample1	0	0	100	100
Sample2	0	0	100	100
Sample3	1	2	99	99.74
Sample4	1	2	99	99.74
Sample5	0	0	100	100
Sample6	0	0	100	100
Sample7	2	4	98	99.49
Sample8	2	4	98	99.42
Sample9	1	2	99	99.75
Sample10	1	2	99	99.75

Conclusion:

This project is Pronunciation modelling for Penang Hokkien. In this project the Penang Hokkien Pronunciation Dictionary is completed and there is 10 set of sample used for check accuracy. The error rate in this 10 set sample test is low. The motivation to this project is let the youngest generation learn to speech PH correctly and prevent the PH missing.

Pronunciation Modelling For Penang Hokkien

ORIGINALITY REPORT

5%

SIMILARITY INDEX

2%

INTERNET SOURCES

3%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Universiti Sains Malaysia

Student Paper

1%

2

Yin-Lai Yeong, Tien-Ping Tan. "Applying Grapheme, Word, and Syllable Information for Language Identification in Code Switching Sentences", 2011 International Conference on Asian Language Processing, 2011

Publication

1%

3

aclclp.org.tw

Internet Source

1%

4

Submitted to Mililani High School

Student Paper

<1%

5

Kanishka Rao, Fuchun Peng, Hasim Sak, Francoise Beaufays. "Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks", 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015

Publication

<1%

費儒伯;Robert Fox. "英譯歌仔戲：翻譯廖瓊枝的

6

《陳三五娘》", 國立台灣師範大學台灣語文學系; Department of Taiwan Culture, Languages, and Literature, NTNU, .

Publication

<1%

7

Submitted to Indiana University

Student Paper

<1%

8

Submitted to Universiti Teknologi Malaysia

Student Paper

<1%

9

Ren-Yuan Lyu. "A Taiwanese text-to-speech system with applications to language learning", IEEE International Conference on Advanced Learning Technologies 2004 Proceedings, 2004

Publication

<1%

10

ijarcce.com

Internet Source

<1%

11

whatis.techtarget.com

Internet Source

<1%

12

link.springer.com

Internet Source

<1%

13

Yu-Ying Chuang, Janice Fon. "On the dialectal variations of voiced sibilant /dz/ in Taiwan Min young speakers", *Lingua Sinica*, 2017

Publication

<1%

14

Submitted to Nanyang Technological University, Singapore

Student Paper

<1%

15

www.futuretdm.eu

Internet Source

<1%

16

Submitted to University of Computer Studies

Student Paper

<1%

17

Tien-Ping Tan, Sang-Seong Goh, Yen-Min Khaw. "A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System", 2012 International Conference on Asian Language Processing, 2012

Publication

<1%

18

Submitted to University of Technology, Sydney

Student Paper

<1%

19

Submitted to Nepal College of Information Technology

Student Paper

<1%

20

Submitted to Florida Atlantic University

Student Paper

<1%

21

www.semanticscholar.org

Internet Source

<1%

Exclude quotes On

Exclude matches < 8 words

Exclude bibliography On

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1 of 1



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	LEE CHUI CHUN
ID Number(s)	15ACB04748
Programme / Course	COMPUTER SCIENCE
Title of Final Year Project	PRONUNCIATION MODELLING OF PENANG HOKKIEN

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index: <u> 5 </u> % Similarity by source Internet Sources: <u> 2 </u> % Publications: <u> 3 </u> % Student Papers: <u> 3 </u> %	
Number of individual sources listed of more than 3% similarity: <u> 0 </u>	
Parameters of originality required and limits approved by UTAR are as follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i>	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor
Name: Dr. JASMINA KHAW YEN MIN

Date: 23-04-2020

Signature of Co-Supervisor
Name: _____

Date: _____

FINAL YEAR PROJECT WEEKLY REPORT

(*Project I / Project II*)

Trimester, Year: Trimester 1, Year 4	Study week no.: 1
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- None

2. WORK TO BE DONE

- Go through FYP 1
- Know about Praat

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Slow progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 2
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Learn about the Praat, how to use and how it look.
- Complete using Praat to select all the unique belong sound and picture.

2. WORK TO BE DONE

- Mapping the unique word with the Praat sound file.
- Complete the unique word part

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Slow progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 3
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- None

2. WORK TO BE DONE

- Whole FYP2

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- No progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 4
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Go through the FYP1 report
- Make changes in FYP1 on what previous done wrong
- Arranging the data
- Meet with supervisor for discussion

2. WORK TO BE DONE

- Check through the FYP1 report
- Make some changes in FYP1
- Understand back the whole FYP1 done and what the next step

3. PROBLEMS ENCOUNTERED

- Found that the data written in FYP1 not very clear
- After get discussion with supervisor found that previous doing is wrong for the Praat part
- Found that previous alignment process using in FYP1 is not good

4. SELF EVALUATION OF THE PROGRESS

- Fair



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(*Project I / Project II*)

Trimester, Year: Trimester 1, Year 4	Study week no.: 5
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Learn about how using Praat do alignment.
- Align the sound1 to respective Chinese word
- Align the Chinese word to same meaning Malay word
- Align the sound4 to respective Chinese word
- Align the Chinese word to same meaning Malay word

2. WORK TO BE DONE

- Complete the sound2 alignment
- Complete the sound3 alignment

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Fair



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 6
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Learn about how using Praat do alignment.
- Align the sound2 to respective Chinese word
- Align the Chinese word to same meaning Malay word
- Align the sound3 to respective Chinese word

2. WORK TO BE DONE

- Complete the sound3 alignment

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Slow progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 7
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Align the Chinese word to same meaning Malay word (sound3)
- Data arrangement to the completed alignment text file
- Created a text file that removed the Malay word and not useful data
- Collect all the unique word data

2. WORK TO BE DONE

- Data analysis on the count of total word, count of general word, count of unique word
- Create unique word list

3. PROBLEMS ENCOUNTERED

- Manual count of data is confusing, plan to using python to solve it

4. SELF EVALUATION OF THE PROGRESS

- Fair



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 8
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Arrange out all the needed file to use, remove all the extra data
- Using python to do data cleaning
- Finish doing of data analysis (count word, calculate percentage)
- Created unique word list

2. WORK TO BE DONE

- Research more about Penang Hokkien Phoneme, Taiwan Minnan
- Create grapheme list

3. PROBLEMS ENCOUNTERED

- Used certain time on find how to data cleaning

4. SELF EVALUATION OF THE PROGRESS

- Fair



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 9
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Get the data from Taiwan Minnan dictionary and Penang Hokkien website
- Sort out the data
- Complete grapheme list

2. WORK TO BE DONE

- Romanized the Chinese word

3. PROBLEMS ENCOUNTERED

- Slow progress

4. SELF EVALUATION OF THE PROGRESS

- Slow progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 10
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Complete Romanized the Chinese word
- Get unique word data from website, romanized it

2. WORK TO BE DONE

- Create rule

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Slow Progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 11
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Created the rule
- Classified data to three category

2. WORK TO BE DONE

- Generate Penang Hokkien Phoneme
- Coding

3. PROBLEMS ENCOUNTERED

- Finding a way to apply rule inside

4. SELF EVALUATION OF THE PROGRESS

- Fair



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 12
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- None

2. WORK TO BE DONE

- Generate Phonemes (on progress)
- FYP2 report

3. PROBLEMS ENCOUNTERED

- Create function to changing phoneme
- MCO cause slow progress
- Extend due date cause slow progress

4. SELF EVALUATION OF THE PROGRESS

- Slow progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 13
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Complete generate Penang Hokkien phonemes
- Generate phonemes for unique word

2. WORK TO BE DONE

- create Pronunciation Modelling for Penang Hokkien
- Data Analysis about accuracy

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Fair



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 14 (Extend week due to MCO)
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Complete create Pronunciation Modelling for Penang Hokkien
- Data Analysis about accuracy
- FYP 2 (Chapter 1 – 4(a bit))

2. WORK TO BE DONE

- FYP2 report

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Good



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: Trimester 1, Year 4	Study week no.: 15 (Extend week due to MCO)
Student Name & ID: Lee Chui Chun 15ACB04748	
Supervisor: Dr. Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- FYP2 report

2. WORK TO BE DONE

- None

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- Good



Supervisor's signature



Student's signature



UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	15ACB04748
Student Name	LEE CHUI CHUN
Supervisor Name	Dr. JASMINA KHAW YEN MIN

TICK (✓)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
✓	Front Cover
✓	Signed Report Status Declaration Form
✓	Title Page
✓	Signed form of the Declaration of Originality
✓	Acknowledgement
✓	Abstract
✓	Table of Contents
✓	List of Figures (if applicable)
✓	List of Tables (if applicable)
✓	List of Symbols (if applicable)
✓	List of Abbreviations (if applicable)
✓	Chapters / Content
✓	Bibliography (or References)
✓	All references in bibliography are cited in the thesis, especially in the chapter of literature review
✓	Appendices (if applicable)
✓	Poster
✓	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)

*Include this form (checklist) in the thesis (Bind together as the last page)

<p>I, the author, have checked and confirmed all the items listed in the table are included in my report.</p> <div style="text-align: center; margin-top: 20px;"> </div> <p style="text-align: center;">(Signature of Student)</p> <p>Date: 23-04-2020</p>	<p>Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction.</p> <div style="text-align: center; margin-top: 20px;"> </div> <p style="text-align: center;">(Signature of Supervisor)</p> <p>Date: 23-04-2020</p>
--	--