

**Improving Speech-to-Text Recognition for Malaysian English Accents  
Using Accent Identification**

BY

LEN SHU YUAN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

Faculty of Information and Communication Technology

(Kampar Campus)

JAN 2022

## REPORT STATUS DECLARATION FORM

**Title:** IMPROVING SPEECH-TO-TEST RECOGNITION FOR MALAYSIAN  
ENGLISH ACCENTS USING ACCENT IDENTIFICATION

**Academic Session:** JAN 2022

I LEN SHU YUAN  
(CAPITAL LETTER)

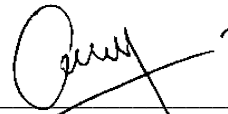
declare that I allow this Final Year Project Report to be kept in  
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

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



(Author's signature)

Verified by,



(Supervisor's signature)

**Address:**

69, Jalan BS 6/15 Taman  
Bukit Serdang 43300,  
Seri Kembangan, Selangor

*Ts. Dr. Cheng Wai Khuen*

Supervisor's name

**Date:** 22<sup>nd</sup> April 2022

22/4/2022  
**Date:** \_\_\_\_\_

<b>Universiti Tunku Abdul Rahman</b>			
Form Title : <b>Sample of Submission Sheet for FYP/Dissertation/Thesis</b>			
Form Number: <b>FM-IAD-004</b>	Rev No.: <b>0</b>	Effective Date: <b>21 JUNE 2011</b>	Page No.: <b>1 of 1</b>

**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**  
**UNIVERSITI TUNKU ABDUL RAHMAN**

Date: 22/4/2022

**SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS**

It is hereby certified that *Len Shu Yuan* (ID No: **18ACB07073** ) has completed this final year project entitled “*Improving Speech-to-Text Recognition for Malaysian English Accents Using Accent Identification*” under the supervision of Ts Dr Cheng Wai Khuen (Supervisor) from the Department of Computer Science, Faculty of Information and Communication Technology , and Dr Jasmina Khaw Yen Min (Co-Supervisor) from the Department of Computer Science, Faculty of Information and Communication Technology.

I understand that University will upload softcopy of my final year project in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,





---

(Len Shu Yuan)

\*Delete whichever not applicable

## DECLARATION OF ORIGINALITY

I declare that this report entitled “**Improving Speech-to-Text Recognition for Malaysian English Accents Using Accent Identification**” 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 : Len Shu Yuan \_\_\_\_\_

Date : 22/4/2022 \_\_\_\_\_

# ACKNOWLEDGEMENTS

In this final year project, I would like to express my sincere thanks to all the parties that have support me and help me to complete my project.

Firstly, I would like to thank and express my appreciation to my supervisor, Ts Dr Cheng Wai Khuen who has provided guidance and suggestions on the project which had help me in building this project.

In addition, I would like to thank my family and friends that encourage and support me during the period of project development and throughout the course of study.

# ABSTRACT

Automatic Speech Recognition (ASR) is the technology that helps user to use their voice as a form of input and it is used in many areas such as mobile devices, embedded systems, and other industrial areas. However, performance and accuracy of the speech recognition system is heavily influenced by the non-native accents, for example, Malaysian English. In this project, the Accent Identification (AID) techniques will be implemented to improve the performance of the ASR systems in recognizing Malaysian English accents. Kaldi toolkits is used in developing proposed ASR models (GMM-HMM and DNN-HMM). CNN based AID is implemented using Python language. The datasets used in this project are from Mini Librispeech, Speech Accent Achieve and other Malaysian English speakers. Then, CNN based AID will be developed and the results is investigated and compared. The Word Error rate is selected as the evaluation metric to compare the recognition performance and accuracy.

# TABLE OF CONTENTS

DECLARATION OF ORIGINALITY	iv
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xii
Chapter 1 Introduction	1
1.1 Problem Statement and Motivation	1
1.2 Project Scope	2
1.3 Project Objectives	2
1.4 Impact, Significance and Contribution	2
1.5 Background Information	3
1.6 Report Organization	5
Chapter 2 Literature Review	6
2.1 Overview	6
2.2 Automatic Speech Recognition	7
2.2.1 Kaldi	7
2.2.2 Gaussian Mixture Models-Hidden Markov Model Hybrid System	8
2.2.3 Deep Neural Network-Hidden Markov Model Hybrid Systems	8
2.3 Existing Accent Identification Techniques	10
2.3.1 I-Vector Based Accent Identification	10
2.3.2 Gaussian Mixture Models (GMM) Based Accent Identification	12
2.3.3 Support Vector Machines (SVM) Based Accent Identification	14
	vii

2.3.4	Convolutional Neural Networks (CNN) based Accent Identification	16
2.3.5	Comparison of 3 AID Techniques	17
Chapter 3	Methodology	18
3.1	Overview	18
3.2	Data collection and preprocessing	18
3.2.1	Speech corpus	18
3.2.2	Data preparation and preprocessing	19
3.3	Kaldi models	20
3.3.1	GMM-HMM model	21
3.3.2	DNN model	22
3.4	CNN based accent identification model	23
3.5	Tools to use	23
3.6	Implementation issues and challenges	24
3.7	Evaluation metric	25
Chapter 4	Experimental Setup and Result	26
4.1	ASR training and decoding process	26
4.1.1	Data file preparation	26
4.2	Model training	28
4.2.1	Feature extraction	28
4.2.2	Monophone model (mono)	28
4.2.3	Triphone models (tri1, tri2b, tri3b)	29
4.2.4	TDNN model (chain)	29
4.3	CNN AID model	31
4.4	Result analysis	32
Chapter 5	Conclusion	34
5.1	Project Review, discussions and conclusion	34
5.2	Novelties and contribution	34



5.3 Future work	35
Bibliography	36
Appendices	39

# LIST OF FIGURES

Figure 1.1.1: Overall accuracy by accent group.....	1
Figure 1.5.1: Challenges faced by state-of-the-art ASR systems..	4
Figure 2.1.1 : Architecture of ASR systems (Yu & Deng, 2015).....	6
Figure 2.2.1.1 : A simplified view of the different components of Kaldi.....	7
Figure 2.2.2.2: GMM-HMM mixture model.....	8
Figure 2.2.3.1: Architecture of the DNN-HMM hybrid system (Yu & Deng, 2015)...	9
Figure 2.2.3.2: An example deep neural network.....	10
Figure 2.3.1.1: Confusion matrix for the i-vector accent identification system (NE: Northern English, SE: Southern English, SC: Scottish).....	11
Figure 2.3.1.2: Visualization of the i-vector accent space.....	11
Figure 2.3.1.3: Summary of ASR results.....	12
Figure 2.3.2.1: Accent identification error rate with different number of components.....	13
Figure 2.3.2.2: Accent identification error rate with different number of utterances..	13
Figure 2.3.2.3: Accent identification confusion matrix.....	14
Figure 2.3.3.1: Performance - matching content case (SVM).....	15
Figure 2.3.4.1: Training and test accuracies categorized by accents (CNN).....	16
Figure 3.3.1: : Overview of proposed Kaldi models.....	20
Figure 3.4.1: : CNN trained with MFCC arrays.....	23
Figure 4.2.4.1: : DNN training and decoding screenshot.....	23

# LIST OF TABLES

Table 2.3.5.1: Comparison of AID Techniques (Advantages & Disadvantages).....	17
Table 3.2.1.1: Data used to train and test the acoustic model in Kaldi.....	18
Table 3.2.2.1: Files need to create manually.....	19
Table 3.5.1: OS and tools used in development.....	23
Table 3.5.2: Hardware used in development.....	24
Table 4.2.3.1: Triphone decoding results.....	29
Table 4.3.1: Dataset used in training and test CNN AID model.....	31
Table 4.4.1: WER of different acoustic models.....	32
Table 4.4.2: Confusion matrix of CNN based accent identifier.....	33

# LIST OF ABBREVIATIONS

<i>ASR</i>	Automatic Speech Recognition
<i>AID</i>	Accent Identification
<i>DNN</i>	Deep Neural Network
<i>DNN-HMM</i>	Deep Neural Network-Hidden Markov Model
<i>HMM</i>	Hidden Markov Model
<i>MFCC</i>	Mel-frequency cepstral coefficient
<i>SVM</i>	Support Vector Machines
<i>WER</i>	Word Error Rate
<i>GMM</i>	Gaussian Mixture Model
<i>GMM-HMM</i>	Gaussian Mixture Model-Hidden Markov Model
<i>ERR</i>	Error Rate Reduction
<i>MLP</i>	Multi-Layer Perceptron
<i>CMVN</i>	Cepstral Mean and Variance Normalization

# Chapter 1 Introduction

## 1.1 Problem Statement and Motivation

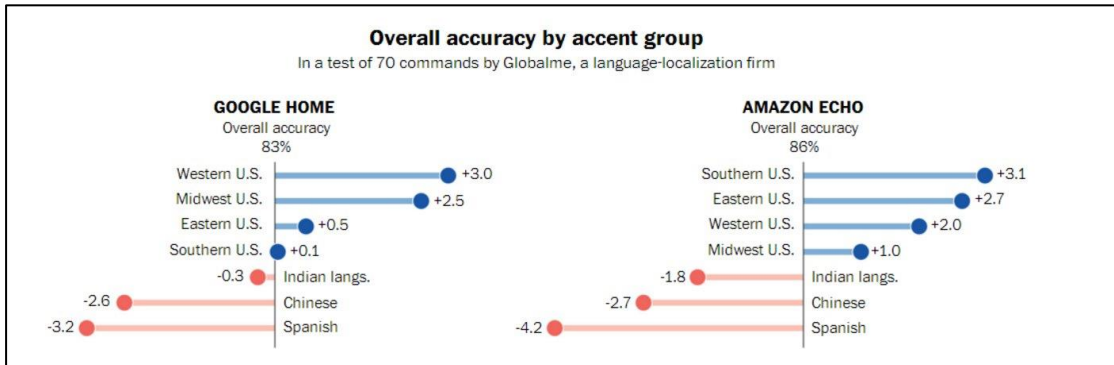


Figure 1.1.1: Overall accuracy by accent group

The Washington Post newspaper investigated the issue of smart speakers' accent imbalance and found out that people with non-American accents faced the biggest setbacks to get accurate responses from the smart speakers (Harwell, 2018). Malaysian English is one of the most spoken languages among Malaysians in business and education but still one of the minority accents that are more likely to face accuracy problem in using voice assistant applications and smart speakers, which is because American accents were used in ASR technologies from the very beginning.

In the field of speech recognition, there are many methods proposed to solve speaker identification issues and speaker identification issues. The researchers focus on accent identification are relatively less compared to other issues. The main issue of the speech recognition field is the complexity of human languages. The accents of native speakers may vary considerable from region to region, not to mention that accents of non-native speaker vary even more. The motivation of this project is to solve the accuracy problems of state-of-the-art ASR systems dealing with non-native accents like Malaysian English which is a mixture of different accents from several ethnic groups. In response to this problem, I propose to implement Accent Identification (AID) techniques for ASR to adapt to accented speech input and perform experiments to find out the best model.

## **1.2 Project Scope**

The goal of this project is to train the proposed ASR models and AID model which are robust while dealing with Malaysian English accented speech. The project will focus on using Kaldi toolkit which is a state-of-the-art speech recognition toolkit. The Kaldi toolkit will be used to build the acoustic model for different proposed ASR models. The AID model will be trained using accented speech from both native and non-native speakers. Then, the performance of different models will be evaluated and analyzed.

## **1.3 Project Objectives**

The main objectives of this project are intended to accomplish:

- To implement the proposed ASR models using Kaldi
- To train the proposed AID model with CNN
- To analyse the performance of the proposed ASR models and AID model

## **1.4 Impact, Significance and Contribution**

The purpose of this project is to train and research on different ASR models and AID model to improve the performance of recognizing Malaysian English accented speech. Most of the speech recognition systems in the application or embedded system are mainly focused on British and American English. The accented ASR system built for Malaysian English can help in localization of the application or product. For example, the in-car ASR system with AID that is trained for Malaysian English accents can increase the accuracy and performance in recognizing Malaysian English speech and avoid repetition of the command or speech. It helps in improving efficiency, user satisfaction and competitiveness in Malaysian market.

## 1.5 Background Information

Speaking is the primary mode of communication among human and speech processing technology has made speech as a form of input to a system successful. Automatic Speech Recognition (ASR) techniques are invented to decode and transcribe speech signals into text. In recent years, ASR has been commercialised and implemented in many technology devices due to the rapid development of ubiquitous computing. ASR is essential in improving user experience by converting speech wave into text accurately and instantly to achieve natural Human-Computer Interaction in small devices and smart speakers that hardly come with keyboard and mouse as input devices. Applications and devices with ASR systems such as Voice Search and Personal Digital Assistance (PDA) become useful in daily life and people are adapting to the new form of input that requires no specific skills to use, which help to expand the customer base. ASR can also be integrated with other computer science areas such as Machine Translation to form new applications like Spoken Language Translation (Khadivi & Ney, 2008).

Besides, online video-sharing platforms such as YouTube implement ASR to create a text transcription for auto captions which is helpful for non-native speakers and deaf users to comprehend the speech. ASR technique is also used for Audio Search Engine to perform Information Retrieval searching multimedia contents such as audio and video (Salgado-Garza & Nolzco-Flores, 2004). There are many related technologies introduced to build a more robust ASR system such as Speaker Recognition, Gender Identification, Accent Identification, Bilingual and Multilingual ASR to fulfil different users' needs.

Nowadays, there are many state-of-the-art ASR systems that enable user to operate their devices using voice commands in different languages. However, there are some factors affecting the performance of ASR shown in Figure 1.1.1.

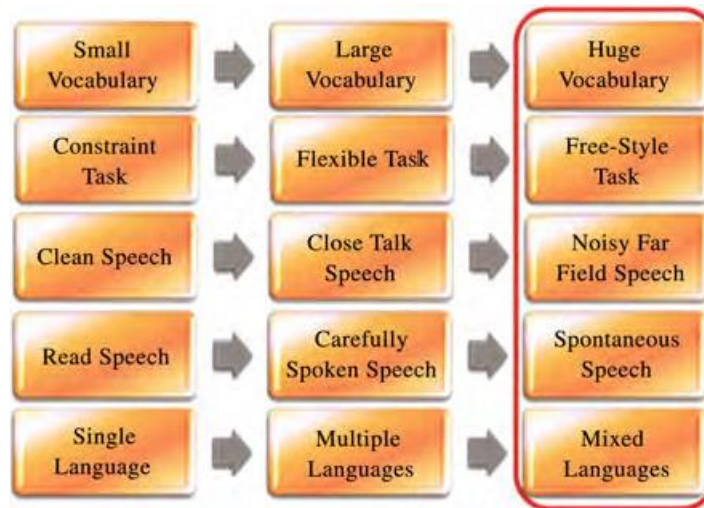


Figure 1.5.1: Challenges faced by state-of-the-art ASR systems.

The accuracy of ASR systems is not only affected by the clarity of sound, but also the cultural backgrounds and health conditions of speakers. The foreign accent problem will be the focus point in this project. The accents of the non-native speaker can deteriorate the performance and robustness of ASR systems that developed with standard English models. The pronunciation variation problem happens in accented speech due to the influence of mother tongue affecting the non-native speakers to perceive and pronounce the utterances that are not existing in their mother tongue (Liu & Fung, 2004). The one of the reasons of the accented speech problem is difficult to solve is because the accented speech resources are usually limited compared to the standard and official language accents used to develop the state-of-the-art ASR system.

English is the top spoken second language in the world and Malaysia is one of the countries using English in formal or non-formal occasions. However, only a small percentage of Malaysian are very fluent in English and able to get rid of foreign accents. The influences of local native languages such as Malay, Chinese and Tamil on Malaysian English not only affect the pronunciation and intonation, but also the grammar structures and addition of local vocabulary.



## **1.6 Report Organization**

This report consists of five chapters. Chapter one is the introduction of the project, which include the background information, problem statement and motivation, project scope and objectives.

Chapter two is the literature review on the Kaldi toolkit, GMM-HMM model, DNN-HMM model and review on different AID techniques.

Chapter three describes the methodology used in this project, which included the speech corpus used, data preparation, overview of the proposed ASR models.

Chapter four will describe the experimental setup and the results of the project.

Chapter five reports the conclusion, novelties and contributions, and future work.

# Chapter 2 Literature Review

## 2.1 Overview

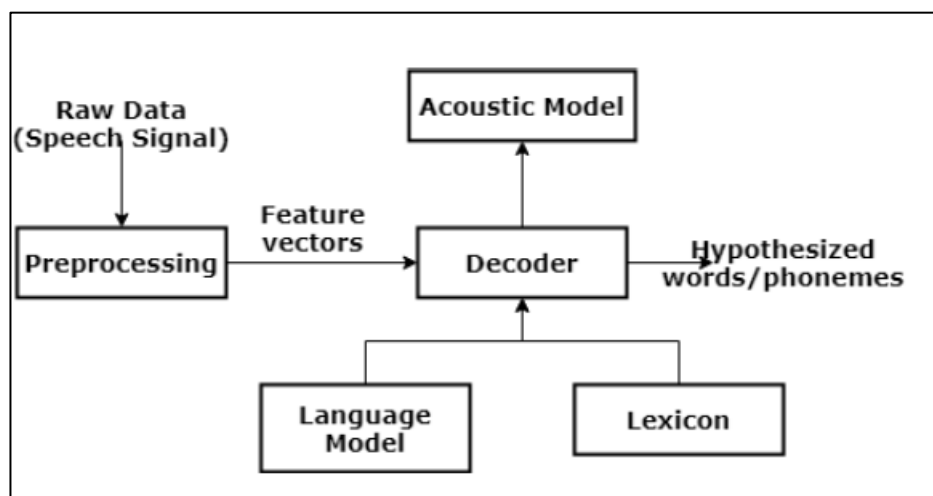


Figure 2.1.1 : Architecture of ASR systems (Yu & Deng, 2015).

ASR has been a popular research area and many researchers have done research on it and other related areas such as Language Identification, Speaker Identification and Accent Identification (AID) to integrate with ASR to improve robustness and accuracy. As shown in the previous chapter, this project concerns on accented speech problems and focuses on evaluating the performance of different ASR models and AID models on accented speech. This chapter is structured as follows: Section 2.2 describes the Kaldi toolkits and the architectures of GMM-HMM ASR and DNN-HMM ASR. Section 3 then presents the literature reviews of current AID techniques. The last section of this chapter summarizes the literature reviews and displays the comparisons of those approaches.

## 2.2 Automatic Speech Recognition

### 2.2.1 Kaldi

Kaldi is an open-source state-of-the-art automatic speech recognition (ASR) toolkit written in C++ and licensed under the Apache License v2.0. Kaldi contains the recipes for training user's own acoustic model, it also supports user to use pre-trained models to decode user's own audio data.

Figure 2.2.1.1 shows a simplified view of the different components of Kaldi. The top level of the diagram shows the external libraries depended by the toolkit, which are also freely available: one is OpenFst for the finite-state framework and the other is BLAS/LAPACK numerical algebra libraries. The second level of the diagram is the C++ library which will be called and from the Shell script for building and running a speech recognizer. The least level of the Kaldi toolkit is the Shell scripts that implements the different steps of speech recognition such as data preparation, feature extraction, training models, structuring decoding map, decoding, etc.

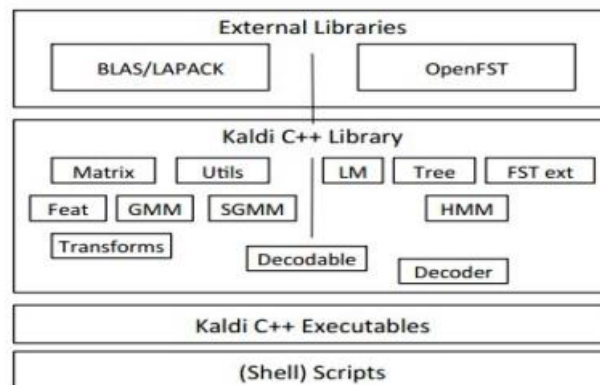


Figure 2.2.1.1 : A simplified view of the different components of Kaldi.

## 2.2.2 Gaussian Mixture Models-Hidden Markov Model Hybrid System

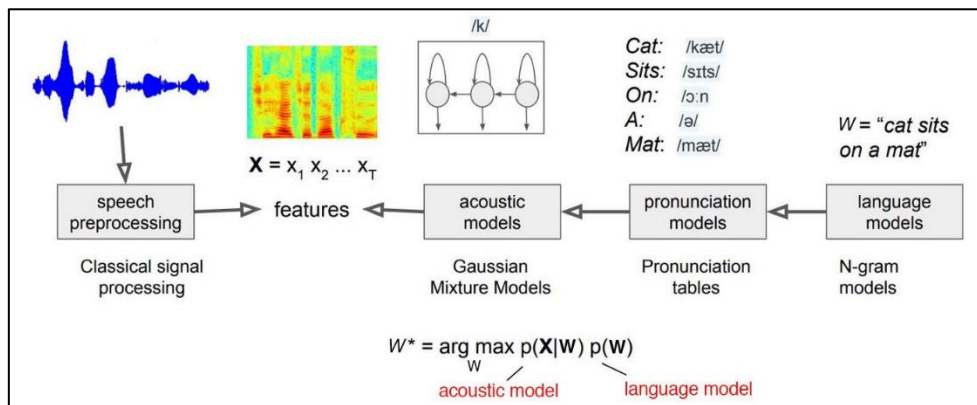


Figure 2.2.2.2: GMM-HMM mixture model

GMM-HMM model is one of the most widely used method in acoustic model training of conventional speech recognition systems. GMM is to recognize frames (features from MFCCs extraction step) into HMM states. Then the HMM combine the states into phonemes (the smallest unit of sound), phoneme then combined into words.

## 2.2.3 Deep Neural Network-Hidden Markov Model Hybrid Systems

According to Najafian and Russell (2020), the DNN-HMM ASR model achieved a WER reduction of 47% relative to the GMM-HMM model, proving the DNN-HMM model has inherent ability to accommodate accented speech compared to GMM-HMM model (Najafian & Russell, 2020). Hinton et al. (2012) achieved significant performance improvements in state-of-the-art ASR systems by replacing GMMs with DNNs (Hinton, et al., 2012).

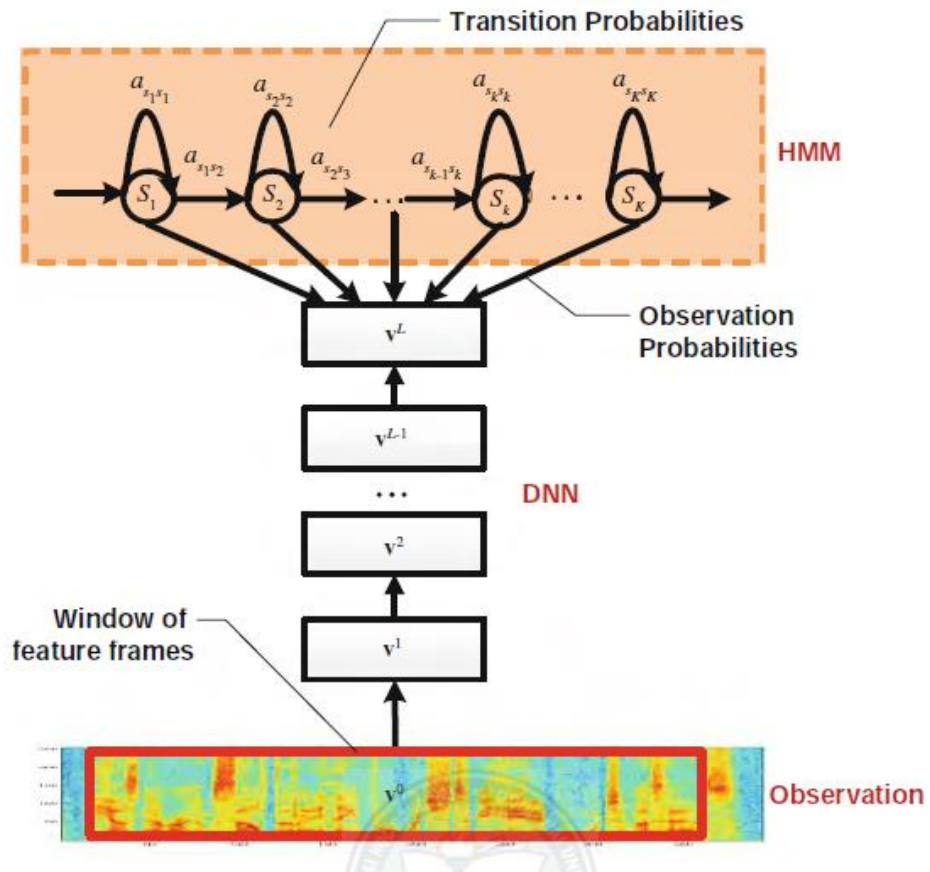


Figure 2.2.3.1: Architecture of the DNN-HMM hybrid system (Yu & Deng, 2015)

According to the architecture of DNN-HMM hybrid system shown in Figure 2.2.3.1, in which the HMM component is used to model the sequential property of the speech signal and the observation probabilities are estimated through DNN in the acoustic observations (Yu & Deng, 2015).

## Deep Neural Network (DNN)

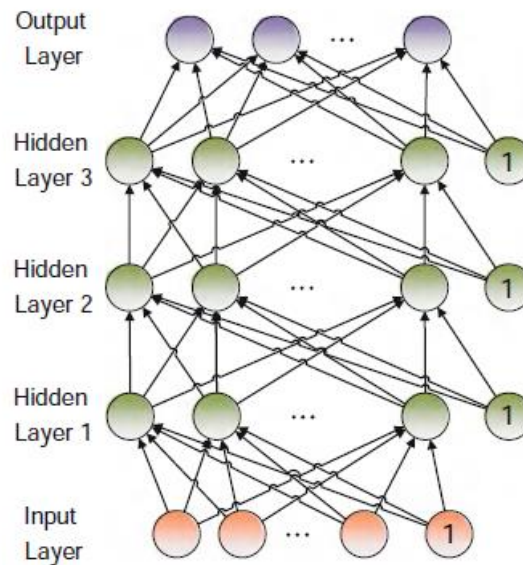


Figure 2.2.3.2: An example deep neural network

Deep Neural Network (DNN) technology is now commonly adopted in the modern speech recognition area. It is an artificial neural network with multilayer perceptron (MLP) with many hidden layers. The hidden layers store the important information such as input's importance and break down the function into specific transformations of the data (Gad, 2018).

### 2.3 Existing Accent Identification Techniques

Accent Identification (AID) is one of the approaches to resolve accented speech accuracy problems. There are many solutions proposed by other researchers and some of the solutions are found meaningful to this project. The goal of this literature review is to compare these existing AID techniques and find out the best model.

#### 2.3.1 I-Vector Based Accent Identification

I-vector based recognition is a popular technique for state-of-the-art speaker recognition and currently there are researchers interested in studying its use in AID.

Najafian and Russell (2020) presented the study of the relationship between AID and ASR accuracy using i-vector based AID. The authors used i-vector based AID for accent-specific acoustical model selection, analysing the accent properties and augmenting the acoustic features to be input to the ASR system. Besides, a multi-class SVM classifier is trained to classify the i-vectors into the different accent classes.

Accent code	Accent group	Acc.	brm	eyk	lan	lvp	ncl	nwa	ilo	sse	ean	crn	roi	uls	shl	gla
brm	NE	80%	16	0	0	0	0	1	0	1	1	1	0	0	0	0
eyk		84%	1	21	2	0	1	0	0	0	0	0	0	0	0	0
lan		76%	1	0	16	0	1	1	1	0	1	0	0	0	0	0
lvp		85%	0	0	1	17	0	2	0	0	0	0	0	0	0	0
ncl		65%	0	0	2	1	13	0	0	0	1	0	0	0	2	1
nwa	SE	52%	1	4	1	0	1	11	0	0	0	2	0	0	1	0
ilo		57%	2	1	3	0	0	0	12	0	2	0	0	0	0	1
sse		69%	0	2	0	0	0	1	2	11	0	0	0	0	0	0
ean		84%	1	1	0	0	0	0	1	0	16	0	0	0	0	0
crn		55%	0	1	0	0	1	1	3	1	1	11	0	0	1	0
roi	IR	78%	0	0	0	0	0	0	0	0	0	0	15	4	0	0
uls		90%	0	0	0	0	0	0	0	0	0	0	2	18	0	0
shl	SC	100%	0	0	0	0	0	0	0	0	0	0	0	0	22	0
gla		95%	0	0	0	0	1	0	0	0	0	0	0	0	0	19

Figure 2.3.1.1: Confusion matrix for the i-vector accent identification system (NE: Northern English, SE: Southern English, SC: Scottish)

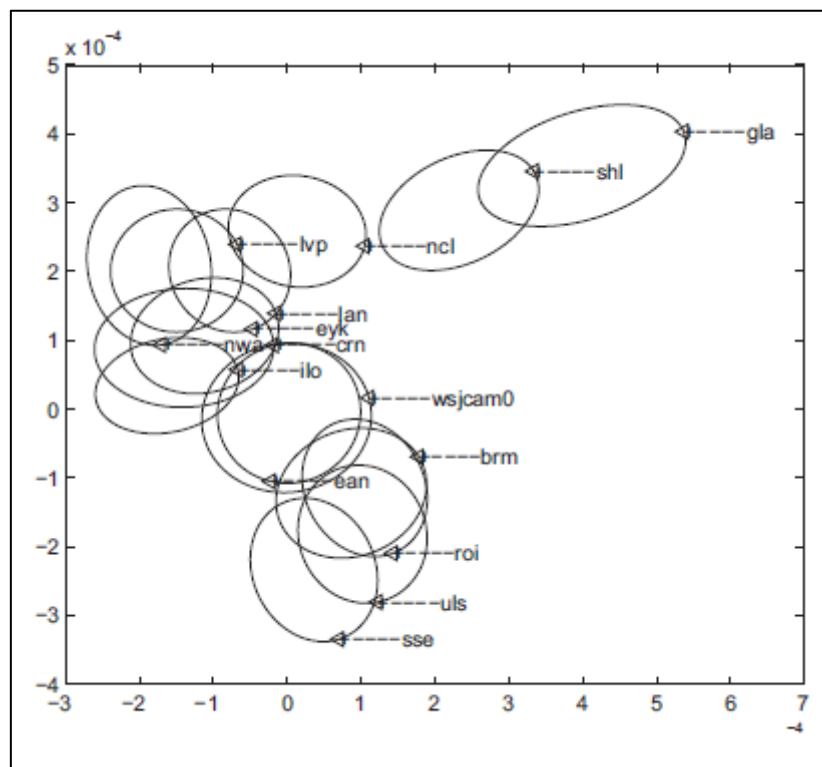


Figure 2.3.1.2: Visualization of the i-vector accent space

According to the confusion matrix in Figure 2.3.1.1, the range of accuracy of i-vector accent identification is 52% to 100%. The authors claimed that the overlap of i-vector accent spaces of the accents shown in Figure 2.3.1.2 causing low accuracy problem.

Code	System	WER (%)	AWR(%)			Regional accents	Broad accents	length (hours)
			Avg	target	off target			
<b>Baselines</b>								
G_0	Baseline GMM-HMM trained on WSJCAM0 only	12.9	-	-	-	-	-	-
D_0	Baseline DNN-HMM trained on WSJCAM0 only	6.9	-	-	-	-	-	-
D_U_iV	D_0 with acoustic vectors augmented with i-vectors	6.2	9.4	-	-	-	-	-
<b>GMM-HMM with unsupervised model selection</b>								
G_U_MSel	Accent-specific GMM-HMM selected using AID	7.4	-7.5	-	-	-	-	-
<b>"Oracle" DNN-HMM systems - test speaker "true" accent known</b>								
D_S_A5	Accent-specific DNN-GMM for "true" accent	5.1	25.8	-	-	-	-	-
D_S_BAG	BAG-specific DNN-HMM for "true" BAG	4.9	28.0	-	-	-	-	-
<b>Accent-independent 'multi-accent' DNN-HMM systems</b>								
D_U_MA2.25	Multi-accent DNN-HMM (2.25hrs data, 14 accents)	4.9	28.3	-	-	14	4	2.3
D_U_MA8.96	Multi-accent DNN-HMM (8.965hrs data, 14 accents)	4.4	35.9	-	-	14	4	9.0
<b>Effect of 'accent diversity' of training set on DNN-HMM performance on accented speech</b>								
D_U_AD2	Low-diversity DNN-HMM (2.25hrs data, 2 accents)	5.1	25.4	42.5	14.2	2	1	2.3
D_U_AD4	Med.-diversity DNN-HMM (2.25hrs data, 4 accents)	5.0	26.9	16.8	22.0	4	3	2.3
D_U_AD8	Med.-diversity DNN-HMM (2.25hrs data, 8 accents)	4.9	28.3	18.2	29.0	8	4	2.3
D_U_AD14	High-diversity DNN-HMM (2.25hrs data, all accents)	4.9	28.3	21.9	-	14	4	2.3
<b>Effect of selecting training data from different BAGs on DNN-HMM performance on accented speech</b>								
D_U_BAG(SC)	DNN-HMM (2.25hrs data from SC BAG)	4.7	31.7	28.7	24.4	2	SC	2.3
D_U_BAG(IR)	DNN-HMM (2.25hrs data from IR BAG)	4.9	28.3	23.2	21.6	2	IR	2.3
D_U_BAG(SE)	DNN-HMM (2.25hrs data from SE BAG)	5.6	17.7	-4.3	8.7	4	SE	2.3
D_U_BAG(NE)	DNN-HMM (2.25hrs data from NE BAG)	5.3	22.3	19.3	12.4	6	NE	2.3

Figure 2.3.1.3: Summary of ASR results

Figure 2.3.1.3 shows the summary of results of different ASR models and training data selections including the baseline systems, GMM-HMM with AID, accent-specific DNN-HMM systems with separate accent data and grouped accent data, multiple-accent DNN-HMM systems with 2.25 hours and 8.965 hours of data, levels of accent diversity on multiple-accent DNN-HMM systems and multiple-accent HNN-HMM systems with different accent group. From the results, all the different types of ASR systems with i-vector AID are performing better than the baseline systems with a range of reduction of word error rate (WER).

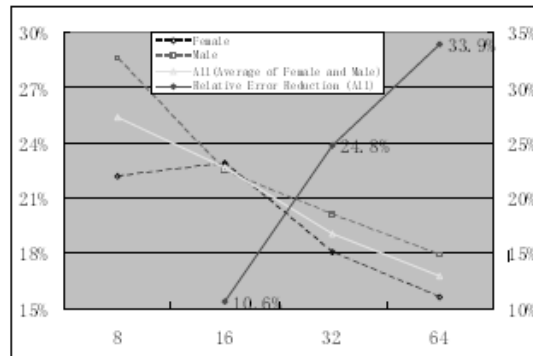
The strength of i-vector based AID is it support the visualization of i-vector accent space to show the relationship between different accents. However, a large training corpus is required to achieve best performance.

## 2.3.2 Gaussian Mixture Models (GMM) Based Accent Identification

Chen et al. (2001) proposes to apply GMM based AID method to solve the Mandarin accented speech problem. The corpus used by the researchers contained 16

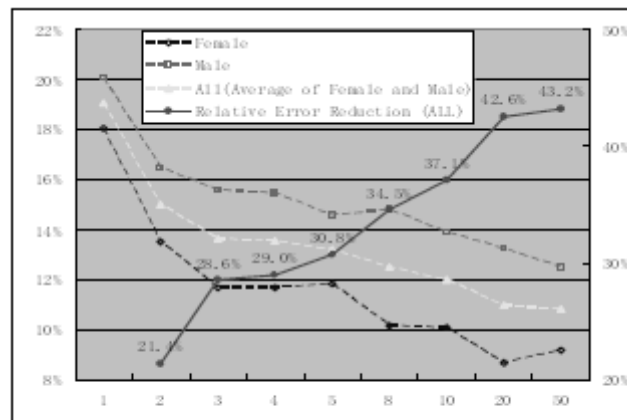


hours of accented speech data including 4 accents in China with 1440 speakers. The parameters used in GMM algorithms are estimated by the expectation maximization (EM) algorithm. The experiments were carried out to examine the effect of number of components on GMM and number of utterances per speaker on the identification accuracy and performance. The results are shown in Figure 2.3.2.1, 2.3.2.2, 2.3.2.3.



**Fig. 1:** Accent identification error rate with different number of components. The horizontal axis is the number of components in GMMs. The left vertical axis is the identification error rate; the right vertical axis is the relative error reduction of "All".

Figure 2.3.2.1: Accent identification error rate with different number of components.



**Fig. 2:** Accent identification error rate with different number of utterances. The horizontal axis is the number of utterances for averaging. The left vertical axis is the identification error rate; the right vertical axis is the relative error reduction of "All".

Figure 2.3.2.2: Accent identification error rate with different number of utterances (GMM)

Recognized As	Testing Utterances From			
	BJ	SH	GD	TW
BJ	<b>0.775</b>	0.081	0.037	0.001
SH	0.120	<b>0.812</b>	0.076	0.014
GD	0.105	0.105	<b>0.886</b>	0.000
TW	0.000	0.002	0.001	<b>0.985</b>

Table 5. Accent identification confusion matrix.

Figure 2.3.2.3: Accent identification confusion matrix (GMM)

The results show that as the number of components and utterances increase, the relative error reduction improved significantly. The strength of this approach is GMM method can avoid building models for phoneme or phoneme-class. Besides, GMM training is unsupervised and required no transcription (Chen, et al., 2001). However, the authors did not apply the AID with ASR to prove the GMM based AID can improve ASR performance. Besides, GMM algorithm is not suitable for high dimensional input and can fail to work if the dimensionality is too high.

### 2.3.3 Support Vector Machines (SVM) Based Accent Identification

Pedersen and Diederich (2007) presented an analysis of an accent classification system using time-based segments of MFCCs and with SVMs. The system training and testing were performed using a corpus of accented speech collected from 27 Arabic speakers and 13 Indian speakers, which means the system performed binary classification. The authors have compared the performances of three kernel designs (linear, polynomial, RBF) of SVMs and the best results shown in Figure 2.3.3.1 were obtained using linear SVM (Pedersen & Diederich, 2007).

Topic number	Accuracy (%)	Recall (%)	Precision (%)	Sample Duration (s)	Segment Duration (ms)
1	75	92.59	75.76	2	140
2	87.5	96.3	86.67	1	30,40,60-80, 120-150
3	97.5	100	96.43	1	130
				4	60, 80-110, 140

Figure 2.3.3.1: Performance - matching content case (SVM)

There are several strengths of SVM stated by the authors including that SVM was designed for high-dimensional input spaces which are able to work with a large number of features. Besides, the authors proved that this approach is able to achieve high accuracy with very short sample length. SVM is not suitable for large data sets, however, the authors stated that large numbers of accented speech samples are not generally available which avoided the disadvantage. However, this approach was only trained and tested for binary classification of accents, so the performance of this approach on multiple accents identification remains unknown (Pedersen & Diederich, 2007).

### 2.3.4 Convolutional Neural Networks (CNN) based Accent Identification

In Chionh, Song and Yin's paper (2018), they proposed an accent identification method based on Convolutional Neural Networks (CNNs). The researchers focused on Arabic, Korean, Italian and Japanese accent English. The audio data were preprocessed into uniform length with 30 seconds before extracting the MFCC features, then the features were inputted to CNN for training purpose. The researchers claimed that the best model was a CNN constructed with 2 convolutional/pooling layers as a 3-layer CNN did not perform much improvement over a 2-layer CNN, and yet 3-layer CNN took longer time to train.

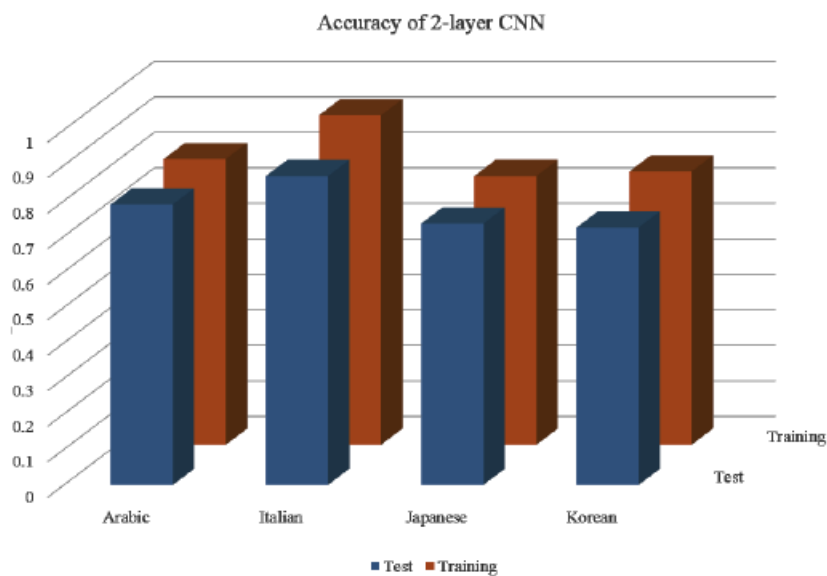


Figure 2.3.4.1: Training and test accuracies categorized by accents (CNN)

### 2.3.5 Comparison of 3 AID Techniques

Table 2.3.5.1 shows the advantages and disadvantages of different AID algorithms.

AID Techniques	Advantages/Strengths	Disadvantages/Weaknesses
<b>i-vector based AID</b>	<ul style="list-style-type: none"> <li>• Visualization of i-vector space indicating relationship between accents</li> </ul>	<ul style="list-style-type: none"> <li>• Requires large training corpus to achieve best results.</li> </ul>
<b>GMM based AID</b>	<ul style="list-style-type: none"> <li>• can avoid building model for phoneme or phoneme-class</li> <li>• unsupervised and required no transcription</li> </ul>	<ul style="list-style-type: none"> <li>• no results on ASR performance</li> <li>• is not suitable for high dimensional input</li> </ul>
<b>SVM based AID</b>	<ul style="list-style-type: none"> <li>• able to work with a large number of features</li> <li>• achieve high accuracy with very short sample length</li> </ul>	<ul style="list-style-type: none"> <li>• not suitable for large data sets</li> <li>• performance on multiple accents identification remain unknown</li> </ul>
<b>CNN based AID</b>	<ul style="list-style-type: none"> <li>• better result compared to GMM</li> <li>• short training time required</li> </ul>	<ul style="list-style-type: none"> <li>• perform poorly on certain accent (e.g.:Korean)</li> </ul>

Table 2.3.5.1: Comparison of AID techniques (Advantages & Disadvantages)

# Chapter 3 Methodology

## 3.1 Overview

In this research, GMM-HMM model and DNN-HMM model of ASR will be developed using Kaldi toolkits. A CNN based accent classifier will be developed using python. This chapter will describe on the data used to train and test the Kaldi models, the data preprocessing procedure and the general architecture of the Kaldi models proposed.

## 3.2 Data collection and preprocessing

### 3.2.1 Speech corpus

Due to the limited computational power, the Mini Librispeech ASR corpus is selected in training Kaldi acoustic models. It is a subset of Librispeech corpus which consists of audio book reading speech. It contains the train set (train\_clean\_5) and the test set (dev\_clean\_2) and the audio files are in flac format, which is an audio coding format for lossless compression of digital audio.

English dataset	#Speaker	#Sentences	Data Size
train_clean_5	29	1516	322M
dev_clean_2	38	1089	126M

Table 3.2.1.1: Data used to train and test the acoustic model in Kaldi.

For the accented speech part, the Speech Accent Archive (Weinberger, 2015) is used to train the accent classifier and also used in decoding of the trained Kaldi models. . It consists of over 2000 speakers from all around the world reading the same paragraph:

'Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop

these things into three red bags, and we will go meet her Wednesday at the train station.'

In this project, speech files from Malaysian speakers will be used and I have collected 20 more speech records from Malaysian Chinese and Indian speakers to ensure accents of different ethnicities are included. Besides, speech data of the American English speakers will be used to show the comparison between native and non-native speakers. 25 speech files selected for both accent classes for decoding purpose. Then, the .mp3 audio files are converted to .wav files as Kaldi requires .wav files for MFCC features extraction. The test set for American English is named as “test\_usa”, and for Malaysian English is “test\_my”.

### 3.2.2 Data preparation and preprocessing

In this project, an existing language model for Mini Librispeech is used. There are some data files required to be prepared manually in order to set up Kaldi and run with own data. The following shows the file name and the format of data in the file.

File name	Format	Usage
text	<utterance-id> <sentence>	Contains mapping between utterances and utterance IDs.
wav.scp	<recording-id> <extended-filename>	Contains location of the audio files and read by Kaldi to perform feature extraction.
utt2spk	<utterance-id> <speaker-id>	1-to-1 mapping between utterance IDs and the speaker.
lexicon.txt	<word> <phone1> <phone2> ...	Lists of real phones and silence phones of each word. Can be obtained from CMU Pronouncing Dictionary.

Table 3.2.2.1: Files need to create manually

Besides, other files are not required to be created manually as Kaldi provides some scripts (local/data\_prep.sh) to generate the files from the files mentioned in Table 3.2.2.1. The commands are listed in [https://kaldi-asr.org/doc/data\\_prep.html](https://kaldi-asr.org/doc/data_prep.html).

After the data preparation step, Kaldi will start to extract Mel Frequency Cepstral Coefficient (MFCC) feature for both train and test set. In the feature extraction process, the speech waveform is divided into segments with frame length of 25 ms and at 10 ms intervals apart which mean they are overlapping.

### 3.3 Kaldi models

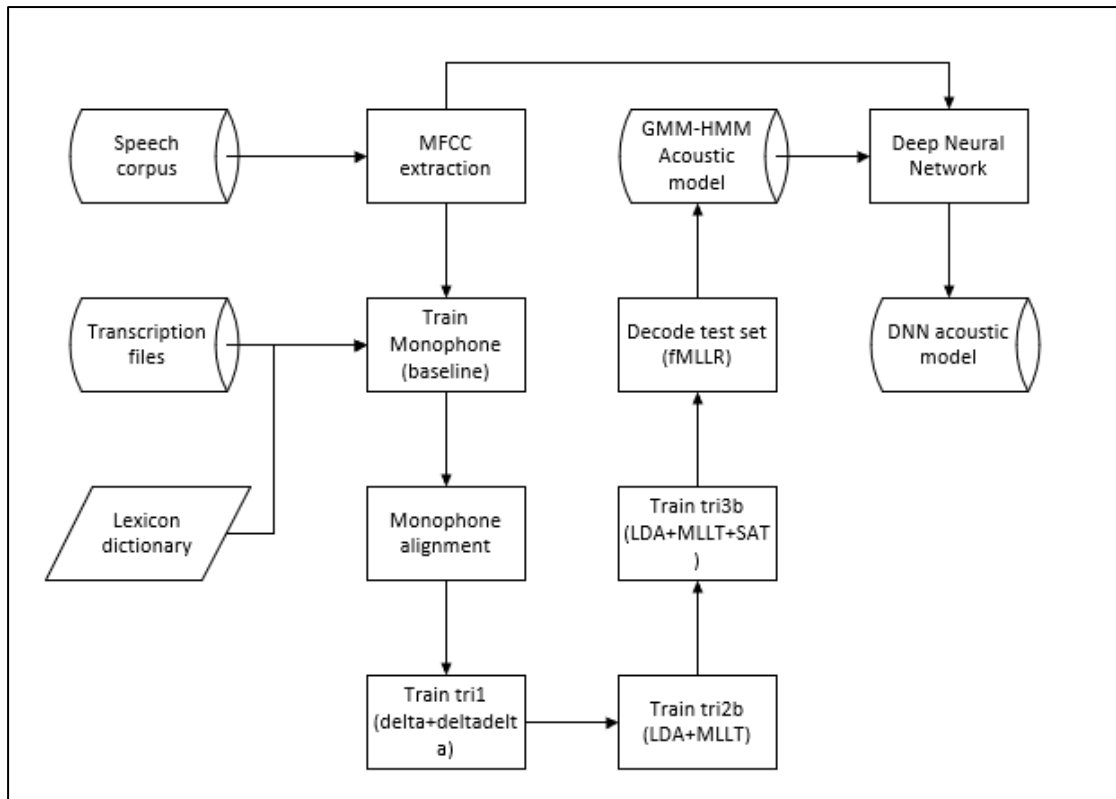


Figure 3.3.1: Overview of proposed Kaldi models.

Acoustic model training is a pipeline process as shown in Figure 3.3.1. After the data required is prepared, the GMM-HMM model will be trained then only proceed to perform DNN training.



### 3.3.1 GMM-HMM model

#### Baseline model - Monophone model

In this project, Monophone model will be the baseline model to evaluate the performance and improvement of other models. Monophone model does not include any contextual information about the previous or subsequent phone. It is used to build block for the triphone models, which does make use of contextual information (Chodroff, 2015). Then, the feature vectors will be aligned to HMM states by using utterance transcription. Viterbi training will take place by cycling through training and alignment phases to optimize the result. After that, the decoding graph, HCLG, is constructed for decoding the accented speech.

#### Triphone $\Delta+\Delta\Delta$ model (tri1)

After training the monophone model, the system uses the training alignment generated to train triphone model with MFCC + delta + delta-delta features. Triphone is a sequence of three consecutive phonemes (Jurafsky & Martin, 2008) and represent a phoneme variant in the context of left and right phonemes.

$d_t = \frac{\sum_{n=1}^N n(c_{t+n} - c_{t-n})}{s \sum_{n=1}^N n^2}$ , where  $d_t$  is a delta coefficient from frame  $t$  calculated from the static coefficients  $c_{t+n}$  to  $c_{t-n}$ , and  $n$  is usually set as 2. The acceleration coefficients are calculated in a similar way but using differentials instead of static coefficients (Raj, 2019). The alignment process is taken place again in each model and the Finite-State Transducers (FST) is created.

#### Triphone LDA+MLLT model (tri2b)

Linear discriminant analysis (LDA) is a linear transform that uses feature vectors and creates HMM states, but the feature space of all data is reduced. The Maximum Likelihood Linear Transformation (MLLT) takes the reduced feature space from the LDA and estimates the parameter of a linear transform and derives a unique transformation for each speaker. Thus, the MLLT minimizes the variation between

speakers (Chodroff, 2015). Once the tri2b model is trained, the system will start to align the utterances for the next tri3b model.

### **Triphone LDA+MLLT+SAT model (tri3b)**

SAT stands for Speaker Adaptive Training. It also normalizes speaker and noise by adapting specific data transformations for each particular speaker. This leads to more homogeneous or normalized data, allowing the model to use its parameters to estimate differences caused by phonemes rather than speakers or recording environments (Chodroff, 2015). Again, decoding graph will be created, and the test set will be decoded using the decoding graph with Feature Space Maximum Likelihood Linear Regression (fMLLR). The acoustic model is no longer trained on the original features after SAT training, but on the speaker-normalized features. For alignment, the model estimates the identity of the speaker (using the inverse of the fMLLR matrix) and then remove the identity of the speaker from the model by multiplying the inverse matrix with the feature vector (Chodroff, 2015). The tri3b will further be used for the DNN-based models.

### **3.3.2 DNN model**

The DNN model requires GPU to install and use CUDA. In this project, the Time Delay Neural Network (TDNN) architecture is expecting to get the input of 40-dimensional high-resolution MFCC features with 25 ms frames and 10 ms shift. Besides, the TDNN model requires i-vectors with 100 dimensions. With that, an i-vector extractor script is required to extract the features. Speed-perturbed data is prepared by performing data augmentation. Then the MFCC and Cepstral Mean and Variance Normalization (CMVN) of the speed-perturbed data will be compute. After that, volume-perturbation will take place and the data from both data augmentation strategies are going though same feature extraction process for training TDNN. In TDNN model a chain-type topology will be created for DNN training purpose.

### 3.4 CNN based accent identification model

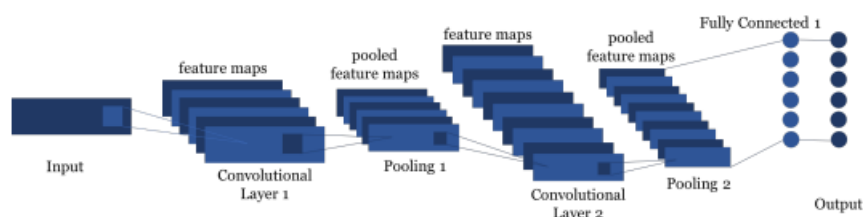


Figure 3.4.1: CNN trained with MFCC arrays

In this project, 2-dimensional CNN architecture will be used to build the accent identifier. To train and test the model, the accented speech data from Speech Accent Archive is used. The datasets chosen are English, Mandarin and Malaysian English. After the data is ready, the model will convert the audio file to .wav format and perform MFCC feature extraction. Then the features are segmented and used to train the 2D CNN model.

### 3.5 Tools to use

This project will be implemented on Ubuntu 20.04 operating system. The table below shows the other tools used to develop the project:

Tools	Version
Ubuntu	20.04
Kaldi	5.5.636
CUDA	10.1
Python	3

Table 3.5.1: OS and tools used in development

Specification	Description
Processor	Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz 2.81 GHz
RAM	8.0 GB
GPU	GeForce GTX 1050 Ti

Table 3.5.2: Hardware used in development

### 3.6 Implementation issues and challenges

There are some challenges and issues I encountered when working on this project. The first challenge is the limited computational power, as my laptop specification is not enough to support the training of large speech corpus. I faced a lot of challenges and errors such as insufficient storage. Therefore, I chose a relatively small speech corpus to perform training and testing.

The next challenge is using Kaldi, a very complex toolkit with a small community and I have never learned to use it before. I am spending a lot of time familiarizing myself with Kaldi in order to deal with compiling and running errors. Besides, I am unfamiliar with programming languages such as Shell scripting that takes me a lot of time to understand the source of errors and actions needed to perform.

### 3.7 Evaluation metric

The performance of the Kaldi models is evaluated by using Word Error Rate (WER), which is the most important metric used to measure the ASR performance. The WER calculation is based on “Levenshtein distance” measurement.

*Word Error Rate = (Substitutions + Insertions + Deletions) / Number of Words Spoken*

Substitution is the number of substitutions when a word is replaced. Insertion is the number of words is added that was not said. Deletion is the number of words omitted from the transcript (Chen, 2021).

The performance of CNN based model will be calculated using the confusion matrix generated by the model.

# Chapter 4 Experimental Setup and Result

## 4.1 ASR training and decoding process

Kaldi provides a list of “recipes” for user to use their own data to train acoustic models. Each recipe contains sample Shell scripts to call different Kaldi binaries. In this project, the Librispeech recipe is chosen for the ASR training with using Mini Librispeech data corpus. After training each model, the Kaldi will perform decoding using the test set and indicate the performance of the model. After the whole ASR models are completed, the different models will be used to decode the accented speech (American English and Malaysian English).

### 4.1.1 Data file preparation

After the speech corpus is downloaded and all the .mp3 files are converted into .vac format by using the following command line:

```
for f in *.mp3; do ffmpeg -i "$f" -acodec pcm_s16le -ac 1 -ar 16000 "${f%.mp3}.wav"; done
```

The transcription files are prepared manually with the format mentioned in Chapter 3 that will be used to link the audio recordings and the recipe. Then, the other files required will be generated by Kaldi. Here are the examples of the data files:

1. text: <utterance-id> <transcripts>

```
lbi-1088-134315-0000 AS YOU KNOW AND AS I HAVE GIVEN YOU PROOF I  
HAVE THE ...
```

```
lbi-1088-134315-0001 THE APOLOGIES WHICH ARE DUE TO YOU I FEEL  
THAT ANYTHING LESS ...
```

```
lbi-1088-134315-0002 TO DISTURB A RELATIONSHIP WHICH I HAVE  
ALWAYS HOPED ...
```

2. wav.scp : <recording-id> <extended-filename>

```
lbi-1088-134315-0000 flac -c -d -s ./corpus/LibriSpeech/train-clean-5/1088/134315/1088-134315-0000.flac |
```

```
lbi-1088-134315-0001 flac -c -d -s ./corpus/LibriSpeech/train-clean-5/1088/134315/1088-134315-0001.flac |
```

```
lbi-1088-134315-0002 flac -c -d -s ./corpus/LibriSpeech/train-clean-5/1088/134315/1088-134315-0002.flac |
```

3. utt2spk : <utterance-id> <speaker-id>

```
lbi-1088-134315-0000 lbi-1088-134315
```

```
lbi-1088-134315-0001 lbi-1088-134315
```

```
lbi-1088-134315-0002 lbi-1088-134315
```

4. lexicon.txt

```
<oov> <oov>
```

```
<sil> SIL
```

```
<UNK> SIL
```

```
Please P L IY Z
```

```
call K AO L
```

```
Stella S T EH L AH
```

```
Ask AE S K
```

To avoid errors such as duplication or unsorted od data in data files, Kaldi provides the scripts *utils/validate\_data\_dir.sh* and *utils/fix\_data\_dir.sh* for user to apply after the data preparation.

## 4.2 Model training

### 4.2.1 Feature extraction

The MFCC features are extracted using Kaldi script *make\_mfcc.sh* with using the configuration as below:

```
--use-energy=false  
--sample-frequency=16000  
--frame-length=25  
--frame-shift=10
```

After the MFCC features are extracted, the next step is to use Kaldi script *compute\_cmvn\_stats.sh* to perform CMVN normalization.

### 4.2.2 Monophone model (mono)

After all the data files, language model files and MFCC features are set up. The monophone model (baseline model) can be started to train by using Kaldi script *train\_mono.sh*. The sample output of the script is shown as below:

```
steps/train_mono.sh: Initializing monophone system.  
steps/train_mono.sh: Compiling training graphs  
steps/train_mono.sh: Aligning data equally (pass 0)  
steps/train_mono.sh: Pass 1  
steps/train_mono.sh: Aligning data  
...  
steps/train_mono.sh: Done training monophone system in exp/mono
```

After the *train\_mono.sh* script finish running, the FST and training graph are generated using *mkgraph.sh* script. Then, the decoding of test set (*dev\_clean\_2*) is done based on the information generated in the training process. The best WER is chosen to be the result of the decoding process. The WER obtained for *dev\_clean\_2* is 66.08. The result of baseline model will be used as a reference to evaluate the improvement of other models.



```
%WER 66.08 [ 13308 / 20138, 973 ins, 2125 del, 10210 sub ]
exp/mono/decode/wer_7_0.5 ##dev_clean_2
```

### 4.2.3 Triphone models (tri1, tri2b, tri3b)

There are three triphone models with different features are trained in this project. The first triphone model, tri1, is trained on the alignments of monophone model and delta features by using Kaldi script *train\_deltas.sh* and mono training graph generated in the baseline model. The second triphone model, tri2b, is train by learning MLLT on top of LDA features, *tran\_lda\_mllt.sh* is used in this case. The final GMM model, tri3b, use Speak Adaption Training (SAT) on top of LDA-MLLT model (tri2b). The objective of this method is to abstract away audio differences between speakers to make the transcription process smoother. *train\_sat.sh* and *align\_fmllr.sh* scripts are used to train the tri3b model and align the SAT triphones. Each model is used to decode the test set (dev\_clena\_2) after training and alignment process. Table 4.2.3.1 shows the decoding results of the three triphone models.

Model	%WER (dev_clean_2)
tri1	48.78
tri2b	45.52
tri3b	45.91

Table 4.2.3.1: Triphone decoding results

### 4.2.4 TDNN model (chain)

The TDNN model is built on top of GMM model, therefore the training graph and information from tri3b are pass to train TDNN model. Before the training process, the i-vector features of audio are extracted using *run\_ivector\_common.sh* script from nnet3 setup. 100-dimensional i-vector features. The TDNN model creates a chain-type topology, which fasten the decode time by 3 times and training time is improving a bit

too (Kaldi, 2022). Then the TDNN model builds a new decision-tree using MFCC features, chain topology and alignment of speed-perturbed data by using *build-tree.sh*. The decoding result of *dev\_clean\_2* is 14.38 %WER. The model is also used to decode *test\_usa* and *test\_my* test set.

```

lsy@ubun: ~/kaldi/egs/2.dnntest/ss
Succeeded creating CMVN stats for test_usa_hires
fix_data_dir.sh: kept all 28 utterances.
fix_data_dir.sh: old files are kept in data/test_usa_hires/.backup
1
1
##### feature extraction completed successfully #####
#####
Thu Apr 21 22:56:08 WITA 2022
utils/mkgraph.sh: exp/chain/tree_a_sp/HCLG.fst is up to date.
steps/nnet3/decode.sh --acwt 1.0 --post-decode-acwt 10.0 --extra-left-context 0
--extra-right-context 0 --extra-left-context-initial 0 --extra-right-context-f
inal 0 --frames-per-chunk 140 --nj 5 --cmd run.pl --num-threads 4 --online-ivec
tor-dir exp/nnet3/ivectors_test_my_hires exp/chain/tree_a_sp/graph data/test_my
_hires exp/chain/tdnn1a_sp/decode_test_my
steps/nnet3/decode.sh --acwt 1.0 --post-decode-acwt 10.0 --extra-left-context 0
--extra-right-context 0 --extra-left-context-initial 0 --extra-right-context-f
inal 0 --frames-per-chunk 140 --nj 5 --cmd run.pl --num-threads 4 --online-ivec
tor-dir exp/nnet3/ivectors_test_usa_hires exp/chain/tree_a_sp/graph data/test_u
sa_hires exp/chain/tdnn1a_sp/decode_test_usa
steps/nnet3/decode.sh: feature type is raw
steps/nnet3/decode.sh: feature type is raw
steps/diagnostic/analyze_lats.sh --cmd run.pl --iter final exp/chain/tree_a_sp/
graph exp/chain/tdnn1a_sp/decode_test_my
steps/diagnostic/analyze_lats.sh: see stats in exp/chain/tdnn1a_sp/decode_test_
my/log/analyze_alignments.log
Overall, lattice depth (10,50,90-percentile)=(21,121,497) and mean=214.2
steps/diagnostic/analyze_lats.sh: see stats in exp/chain/tdnn1a_sp/decode_test_
my/log/analyze_lattice_depth_stats.log
score best paths
steps/diagnostic/analyze_lats.sh --cmd run.pl --iter final exp/chain/tree_a_sp/
graph exp/chain/tdnn1a_sp/decode_test_usa
steps/diagnostic/analyze_lats.sh: see stats in exp/chain/tdnn1a_sp/decode_test_
usa/log/analyze_alignments.log
Overall, lattice depth (10,50,90-percentile)=(13,93,536) and mean=217.7
steps/diagnostic/analyze_lats.sh: see stats in exp/chain/tdnn1a_sp/decode_test_
usa/log/analyze_lattice_depth_stats.log
score best paths
score confidence and timing with sclite
Decoding done.
score confidence and timing with sclite
Decoding done.
steps/score_kaldi.sh --cmd run.pl data/test_my_hires exp/chain/tree_a_sp exp/ch
ain/tdnn1a_sp/decode_test_my_tgsmall
steps/score_kaldi.sh: scoring with word insertion penalty=0.0,0.5,1.0
%WER 77.97 [ 269 / 345, 11 ins, 71 del, 187 sub ] [PARTIAL] exp/chain/tdnn1a_sp
/decode_test_my_tgsmall/wer_12_0.0
steps/score_kaldi.sh --cmd run.pl data/test_usa_hires exp/chain/tree_a_sp exp/c
hain/tdnn1a_sp/decode_test_usa_tgsmall
steps/score_kaldi.sh: scoring with word insertion penalty=0.0,0.5,1.0
%WER 66.72 [ 1311 / 1965, 26 ins, 321 del, 964 sub ] exp/chain/tdnn1a_sp/decode
_test_usa_tgsmall/wer_9_0.0
1
#####decode#####
lsy@ubun:~/kaldi/egs/2.dnntest/ss$

```

Figure 4.2.4.1: DNN training and decoding screenshot

### 4.3 CNN AID model

To train and test the CNN model, the English, Mandarin and Malay dataset is selected from the Speech Accent Archive. Additional 20 audio files are added into Malay dataset to form a total of 25 Malaysian English audio files. In the data preprocessing step, the MFCC features are reduced to 13 as the 13 features represent the most relevant frequency to the human voice range (Chionh, et al., 2018). The dropout rate is set to 0.25 to prevent overfitting. The dataset was split into 80% of training set and 20% of test set.

Dataset	Train set	Test set
English	463	116
Mandarin	52	13
Malaysia	19	6

Table 4.3.1: Dataset used in training and test CNN AID model

#### 4.4 Result analysis

Kaldi Model		WER(%)		
		dev_clean_2	test_usa	test_my
GMM-HMM	mono	66.08	79.69	86.20
	tri1 ( $\Delta+\Delta\Delta$ )	48.78	69.31	82.78
	tri2b (LDA+MLLT)	45.52	65.19	81.04
	tri3b (LDA+MLLT+SAT)	45.91	65.19	80.64
DNN-HMM	TDNN	14.38	66.72	77.97

Table 4.4.1: WER of different acoustic models

Table 4.4.1 illustrate the WER results of the acoustic models built in this project. The test set (dev\_clean\_2), American English (test\_usa) and Malaysian English (test\_my) are decoded using the trained ASR models. For dev\_clean\_2, the WER gives the best performance and has the biggest improvement from tri3b to TDNN, which can say, the DNN-HMM model has a significant effect on the improvement of GMM-HMM models. Looking into the 4 GMM models, the biggest decline of WER take place from monophone to triphone, which can say that triphone with delta features can significantly reduce the WER. Then, the WER is slightly reduced from tri1 to tri2b. However, the WER results of tri2b and tri3b do not show any improvement, which can assume that SAT speaker adaptive training does not improve the accuracy in recognizing the Mini Librispeech corpus, or the corpus size is just not enough to see an improvement.

For the accented speech part, the overall performance of recognizing the American English (test\_usa) is better than Malaysian English (test\_my). It might be due to the pronunciation variation between these two accent groups. However, the performance of decoding test\_usa dataset is relatively poor compared the result of dev\_clean\_2, and the TDNN model does not give improvement in recognizing accented speech compared to tri3b model. The result of decoding test\_my is not good and it may cause by limited number of audio files (25 recordings) and the language model used in this project is not adapted with Malaysian English pronunciation.

Actual/Predict	English	Malaysia	Mandarin	Recall	F <sub>1</sub>
English	110	0	6	94.828%	0.92
Malaysia	4	2	0	33.333%	0.5
Mandarin	8	0	5	38.462%	0.42
Precision	90.164%	100%	45.455%		

Table 4.4.2: Confusion matrix of CNN based accent identifier

The overall accuracy of the classifier is 86.67%. The accuracy results seem good because the CNN model able to identify most of the speech from native speakers correctly. The CNN model does not perform well on classifying non-native speech recordings as the recall and F<sub>1</sub> score for Malaysia and Mandarin accent groups are quite low, and it may cause by limited number of samples in the speech corpus.

# Chapter 5 Conclusion

## 5.1 Project Review, discussions and conclusion

In this project, GMM-HMM models and DNN-HMM model of speech recognition is built using Mini Librispeech speech corpus and Kaldi toolkit. GMM and DNN models with different phoneme structure and feature transformation approaches are compared with respective WER results. The triphone model, tri1 is significantly improve the decoding results compare to the monophone model. However, the tri3b model does not improve the performance of the ASR. The results shows that the DNN-HMM is the best model in decoding test set from Mini Librispeech with only 14.38 WER but the performance of recognizing accented speech is relatively poor. We can assume that the performance of DNN model is only significant in dealing with accented speech when the speech corpus size is large enough. Using the full Librispeech corpus to train the acoustic models may significantly improve the performance of the DNN model. For the CNN accent identification, the result of identifying accented speech from native speakers are much better than data from non-native speakers. The problem may be the sample size gap is too large as there are 579 samples in English dataset but only 25 samples in Malaysian English dataset.

## 5.2 Novelties and contribution

This research adopts different speech recognition models used widely in the field and compared the performance on recognizing accented speech. The method proposed is to help the organization to develop and build the speech recognition functions that adapted to Malaysian English so that they can build goodwill with local customers. The DNN model can help in speech-to-text accuracy and increase the use of voice search devices and smart home devices.

### **5.3 Future work**

There are still a lot of work could be done to improve the performance of the proposed ASR models and AID model. Due to the limited computational power and time, the performances of the ASR models and CNN AID model are not good enough. To obtain better performance, more feature processing methods could be tried to find out the best combination. Besides, the parameters in feature extraction, model training and decoding process can be adjusted and observe whether there is any result improvement. A larger Malaysian English speech corpus is needed for the future work. Furthermore, the influences of the ethnicity and native language of the speaker could be further study to create a language model adapted to Malaysian English.

# Bibliography

Chen, H., 2021. *JANUARY*. [Online] Available at: [https://www.smartaction.ai/blog/does-word-error-rate-matter/#:~:text=Word%20Error%20Rate%20\(WER\)%20is,word%20error%20rate%20of%204%25](https://www.smartaction.ai/blog/does-word-error-rate-matter/#:~:text=Word%20Error%20Rate%20(WER)%20is,word%20error%20rate%20of%204%25).

[Accessed 18 April 2022].

Chen, T., Huang, C., Wang, J. & Chang, E., 2001. *Automatic accent identification using Gaussian mixture models*. Madonna di Campiglio, Italy,, IEEE.

Chionh, K., Song, M. & Yin, Y., 2018. *Application of Convolutional Neural Networks in*, Pittsburgh Pennsylvania: Project Report Carnegie Mellon University.

Chodroff, E., 2015. *Kaldi Tutorial*. [Online] Available at: <https://www.eleanorchodroff.com/tutorial/kaldi/training-overview.html>

[Accessed 13 April 2022].

Gad, A., 2018. *Beginners Ask “How Many Hidden Layers/Neurons to Use in Artificial Neural Networks?”*. [Online]

Available at: <https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning#:~:text=In%20neural%20networks%2C%20a%20hidden,inputs%20entered%20into%20the%20network>.

[Accessed 5 April 2020].

Harwell, D., 2018. *TRhe Washington Post*. [Online]

Available at: <https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/>

[Accessed 10 September 2020].

Hinton, G. et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Magazine*, 29(6), pp. 82-97.

Jakobovski, 2021. *Free Spoken Digit Dataset (FSDD)*. [Online]

Available at: <https://github.com/Jakobovski/free-spoken-digit-dataset>

[Accessed 13 March 2021].



Jurafsky, D. & Martin, J. H., 2008. *Speech and Language Processing*. s.l.:Prentice Hall.

Kaldi, 2022. *Introduction to 'chain' models*. [Online] Available at: [https://kaldi-asr.org/doc/chain.html#chain\\_model](https://kaldi-asr.org/doc/chain.html#chain_model) [Accessed 18 April 2022].

Khadivi, S. & Ney, H., 2008. Integration of Speech Recognition and Machine Translation in Computer-Assisted Translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8), pp. 1551-1564.

Liu, Y. & Fung, P., 2004. Pronunciation Modeling for Spontaneous Mandarin Speech Recognition. *International Journal of Speech Technology*, Volume 7, pp. 155-172.

Najafian, M. & Russell, M., 2020. Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Elsevier*, Volume 122, pp. 44-55.

Pedersen, C. & Diederich, J., 2007. *Accent Classification Using Support Vector Machines*. Sharjah, IEEE.

Povey, D. et al., 2011. *The Kaldi Speech Recognition Toolkit*, Hawaii: IEEE Signal Processing Society.

Raj, D., 2019. *A note on MFCCs and delta features*. [Online] Available at: <https://desh2608.github.io/2019-07-26-delta-feats/> [Accessed 18 April 2022].

Ramon, Y., 2018. *How to start with Kaldi and Speech Recognition*. [Online] Available at: <https://towardsdatascience.com/how-to-start-with-kaldi-and-speech-recognition-a9b7670ffff6> [Accessed 5 April 2021].


Salgado-Garza, L. R. & Nolasco-Flores, J. A., 2004. *On the Use of Automatic Speech Recognition for Spoken Information Retrieval from Video Databases*. Berlin, Heidelberg, Springer.

Weinberger, S., 2015. *Speech Accent Archive*. [Sound Recording] (George Mason University).

Yu, D. & Deng, L., 2015. *Automatic Speech Recognition - A Deep Learning Approach*. 1st ed. London: Springer.

# Appendices

## Poster



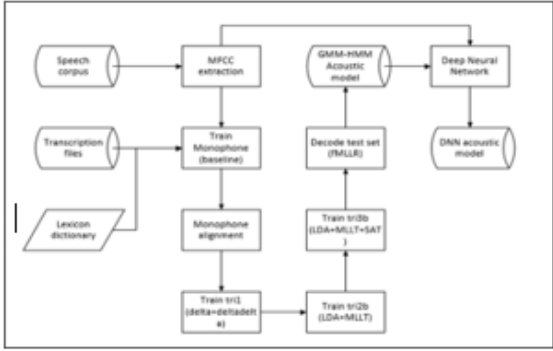
### IMPROVING SPEECH-TO-TEXT RECOGNITION FOR MALAYSIAN ENGLISH ACCENTS USING ACCENT IDENTIFICATION

Automatic Speech Recognition (ASR) is the technology that helps user to use their voice as a form of input and it is used in many areas such as mobile devices, embedded systems, and other industrial areas.

**OBJECTIVES**

- To implement the proposed ASR models using Kaldi
- To train the proposed AID model with CNN
- To analyse the performance of the proposed ASR models and AID model

**Methodology**



**Result**

Kaldi Model	WER(%)		
	dev_clean_2	test_usa	test_intl
mono	66.08	79.69	86.20
tri1	48.78	69.31	82.78
tri2b	45.52	65.19	81.04
tri3b	45.91	65.19	80.64
TDNN	14.38	66.72	77.97

Actual/Predict	English	Malaysia	Mandarin	Recall	F1
English	110	0	6	94.828%	0.92
Malaysia	4	2	0	33.333%	0.5
Mandarin	8	0	5	38.462%	0.42
Precision	90.164%	100%	45.455%		

Table 4.4.2: Confusion matrix of CNN based accent identifier

The results shows that the DNN-HMM is the best model in decoding test set from Mini Librispeech with only 14.38 WER but the performance of recognizing accented speech is relatively poor.

## Improving Speech-to-Text Recognition for Malaysian English Accents Using Accent Identification

### ORIGINALITY REPORT

13%

SIMILARITY INDEX

7%

INTERNET SOURCES

9%

PUBLICATIONS

4%

STUDENT PAPERS

### PRIMARY SOURCES

1	Submitted to Columbia University Student Paper	2%
2	wikitranslation.org Internet Source	1%
3	www.dongcoder.com Internet Source	1%
4	eprints.utar.edu.my Internet Source	1%
5	Maryam Najafian, Martin Russell. "Automatic accent identification as an analytical tool for accent robust automatic speech recognition", <i>Speech Communication</i> , 2020 Publication	1%
6	ia801805.us.archive.org Internet Source	<1%
7	Imad K. Tantawi, Mohammad A. M. Abushariah, Bassam H. Hammo. "A deep learning approach for automatic speech recognition of The Holy Qur'ān recitations",	<1%

International Journal of Speech Technology,  
2021  
Publication

8	"Rule Extraction from Support Vector Machines", Springer Science and Business Media LLC, 2008 Publication	<1 %
9	"New Era for Robust Speech Recognition", Springer Science and Business Media LLC, 2017 Publication	<1 %
10	<a href="http://www.students.ncl.ac.uk">www.students.ncl.ac.uk</a> Internet Source	<1 %
11	Submitted to National Institute of Technology, Kurukshetra Student Paper	<1 %
12	Submitted to University of Birmingham Student Paper	<1 %
13	<a href="http://cobecoballes-embedded.blogspot.com">cobecoballes-embedded.blogspot.com</a> Internet Source	<1 %
14	Anna Jarosz. "English Pronunciation in L2 Instruction", Springer Science and Business Media LLC, 2019 Publication	<1 %
15	<a href="http://community.intel.com">community.intel.com</a> Internet Source	<1 %

16 "Speech and Computer", Springer Science and Business Media LLC, 2016 <1 %  
Publication

---

17 S. Gereg, P. Vizslyay, J. Stas, M. Lojka. "Semi-automatic processing and annotation of meeting audio recordings", 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2019 <1 %  
Publication

---

18 John H. L. Hansen, Maryam Najafian, Rasa Lileikyte, Dwight Irvin, Beth Rous. "Speech and language processing for assessing child-adult interaction based on diarization and location", International Journal of Speech Technology, 2019 <1 %  
Publication

---

19 Submitted to Universiti Tunku Abdul Rahman <1 %  
Student Paper

---

20 "Human-Computer Interaction with Special Emphasis on Converting Brain Signals to Speech", International Journal of Innovative Technology and Exploring Engineering, 2020 <1 %  
Publication

---

21 [citeseerx.ist.psu.edu](http://citeseerx.ist.psu.edu) <1 %  
Internet Source

---

[www.rev.com](http://www.rev.com)

22	Internet Source	<1 %
23	F. Bashir, A. Khokhar, D. Schonfeld. "Automatic Object Trajectory-Based Motion Recognition Using Gaussian Mixture Models", 2005 IEEE International Conference on Multimedia and Expo, 2005 Publication	<1 %
24	Arvind Kumar, Rampravesh Kumar, Kamlesh Kishore. "Performance analysis of ASR Model for Santhali language on Kaldi and Matlab Toolkit", 2020 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2020 Publication	<1 %
25	<a href="http://pow.jpn.org">pow.jpn.org</a> Internet Source	<1 %
26	<a href="http://www.cs.uta.fi">www.cs.uta.fi</a> Internet Source	<1 %
27	<a href="http://www.nature.com">www.nature.com</a> Internet Source	<1 %
28	<a href="http://conservancy.umn.edu">conservancy.umn.edu</a> Internet Source	<1 %
29	<a href="http://mountainscholar.org">mountainscholar.org</a> Internet Source	<1 %
	<a href="http://www.ncbi.nlm.nih.gov">www.ncbi.nlm.nih.gov</a>	

30	Internet Source	<1 %
31	Fahimeh Ghasemi, Afshin Fassihi, Horacio Pérez-Sánchez, Alireza Mehri Dehnavi. "The role of different sampling methods in improving biological activity prediction using deep belief network", Journal of Computational Chemistry, 2017 Publication	<1 %
32	Too Chen, Chao Huang, E. Chang, Jingehan Wang. "Automatic accent identification using Gaussian mixture models", IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01., 2001 Publication	<1 %
33	Yi Wu, Jian Liu, Yingying Chen, Jerry Cheng. "Semi-black-box Attacks Against Speech Recognition Systems Using Adversarial Samples", 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), 2019 Publication	<1 %
34	dspace.mit.edu Internet Source	<1 %
35	hal.archives-ouvertes.fr Internet Source	<1 %



36	Alam, Md Jahangir, Vishwa Gupta, Patrick Kenny, and Pierre Dumouchel. "Speech recognition in reverberant and noisy environments employing multiple feature extractors and i-vector speaker adaptation", EURASIP Journal on Advances in Signal Processing, 2015. Publication	<1 %
37	Borsky, Michal, Petr Mizera, Petr Pollak, and Jan Nouza. "Dithering techniques in automatic recognition of speech corrupted by MP3 compression: Analysis, solutions and experiments", Speech Communication, 2017. Publication	<1 %
38	Uday Kamath, John Liu, James Whitaker. "Deep Learning for NLP and Speech Recognition", Springer Science and Business Media LLC, 2019 Publication	<1 %
39	Submitted to University Tun Hussein Onn Malaysia Student Paper	<1 %
40	Yingchun Guo, Kunpeng Zhao, Xiaoke Hao, Ming Yu. "Deep Regression Neural Network for End-to-End Person Re-identification", IEEE Access, 2019 Publication	<1 %

41	Yoshioka, T., and M.J.F. Gales. "Environmentally robust ASR front-end for deep neural network acoustic models", Computer Speech & Language, 2015. Publication	<1 %
42	riunet.upv.es Internet Source	<1 %
43	Andrew Sewell. "The Hong Kong English accent continuum: insights from implicational scaling", Asian Englishes, 2022 Publication	<1 %

---

Exclude quotes    Off                      Exclude matches    Off

Exclude bibliography    On

The screenshot shows the Turnitin Feedback Studio interface. The main document area displays an abstract with several highlighted segments. A sidebar on the right, titled "Match Overview", shows a total match percentage of 13% and lists eight individual matches with their respective percentages and source types.

**Match Overview**

1	Submitted to Columbia... Student Paper	2%
2	wikitranslation.org Internet Source	1%
3	www.dongcoder.com Internet Source	1%
4	eprints.utar.edu.my Internet Source	1%
5	Maryam Najafian, Marti... Publication	1%
6	ia801805.us.archive.org Internet Source	<1%
7	Imad K. Tantawi, Moha... Publication	<1%
8	"Rule Extraction from S... Publication	<1%

Page: 1 of 35    Word Count: 6048    Text-Only Report    High Resolution

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

<b>Trimester, Year: Trimester 1, Year 4</b>	<b>Study week no.: 6</b>
<b>Student Name &amp; ID: Len Shu Yuan 1807073</b>	
<b>Supervisor: Ts Dr Cheng Wai Khuen</b>	
<b>Project Title: Improving Speech-to-Text Recognition for Malaysian English Accents Using Accent Identification</b>	

## 1. WORK DONE

Trained GMM Kaldi model

## 2. WORK TO BE DONE

Solve CUDA installation problem  
DNN model training

## 3. PROBLEMS ENCOUNTERED

Unable to use GPU in VMware, unable to continue to train DNN model.

## 4. SELF EVALUATION OF THE PROGRESS

Need to speed up on the progress.



\_\_\_\_\_  
Supervisor's signature



\_\_\_\_\_  
Student's signature

## FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

<b>Trimester, Year: Trimester 1, Year 4</b>	<b>Study week no.: 12</b>
<b>Student Name &amp; ID: Len Shu Yuan 1807073</b>	
<b>Supervisor: Ts Dr Cheng Wai Khuen</b>	
<b>Project Title: Improving Speech-to-Text Recognition for Malaysian English Accents Using Accent Identification</b>	

<b>1. WORK DONE</b> Complete building Kaldi GMM and DNN model and decoding the accented speech.
<b>2. WORK TO BE DONE</b> FYP2 report and improve CNN AID performance
<b>3. PROBLEMS ENCOUNTERED</b> High error rate in decoding accented speech.
<b>4. SELF EVALUATION OF THE PROGRESS</b> Need to figure how to improve the performance



Supervisor's signature



Student's signature

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



**FACULTY OF Information and Communication Technology**

<b>Full Name(s) of Candidate(s)</b>	Len Shu Yuan
<b>ID Number(s)</b>	18ACB07073
<b>Programme / Course</b>	Bachelor of Computer Science (HONOURS)
<b>Title of Final Year Project</b>	Improving Speech-to-Text Recognition for Malaysian English Accents Using Accent Identification

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
<b>Overall similarity index: <u>13</u> %</b> <b>Similarity by source</b> Internet Sources: <u>7</u> % Publications: <u>9</u> % Student Papers: <u>4</u> %	OK
<b>Number of individual sources listed of more than 3% similarity: <u>0</u></b>	
<b>Parameters of originality required and limits approved by UTAR are as follows:</b> (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:     Ts. Dr. Cheng Wai Khuen    

Date:     22/4/2022    

\_\_\_\_\_  
Signature of Co-Supervisor

Name: \_\_\_\_\_

Date: \_\_\_\_\_

## UNIVERSITI TUNKU ABDUL RAHMAN

### FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

#### CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	18ACB07073
Student Name	Len Shu Yuan
Supervisor Name	Ts Dr Cheng Wai Khuen

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 Plastic Cover (for hardcopy)
✓	Title Page
✓	Signed Report Status Declaration Form
✓	Signed FYP Thesis Submission Form
✓	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)
✓	Weekly Log
✓	Poster
✓	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
✓	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

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

I, the author, have checked and confirmed all the items listed in the table are included in my report.



\_\_\_\_\_  
(Signature of Student)

Date: 22/4/2022

