



# Exploiting resources from closely-related languages for automatic speech recognition in low-resource languages from Malaysia

Sarah Flora Samson Juan

## ► To cite this version:

Sarah Flora Samson Juan. Exploiting resources from closely-related languages for automatic speech recognition in low-resource languages from Malaysia. Computation and Language [cs.CL]. Université Grenoble Alpes, 2015. English. NNT : 2015GREAM061 . tel-01314120

**HAL Id: tel-01314120**

**<https://theses.hal.science/tel-01314120>**

Submitted on 10 May 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

## THÈSE

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE**

Spécialité : **Informatique**

Arrêté ministériel :

Présentée par

**Sarah Flora SAMSON JUAN**

Thèse dirigée par **Laurent BESACIER**

préparée au sein **Laboratoire d'Informatique de Grenoble**  
et de **École Doctorale de Mathématiques, Sciences et Technologies de l'Information, Informatique**

# Exploiting Resources from Closely-Related Languages for Automatic Speech Recognition in Low-Resource Languages from Malaysia

Thèse soutenue publiquement le **9, Juillet 2015** devant le jury composé de :

**M. George QUÉNOT**

Directeur de Recherche CNRS, LIG, Grenoble, Président

**M. Yannick ESTÈVE**

Professeur à l'Université du Maine, Le Mans, Rapporteur

**M. Denis JOUVET**

Directeur de Recherche INRIA, LORIA, Nancy, Rapporteur

**M. Eric CASTELLI**

Professeur à l'Institut Polytechnique de Hanoi, Chargé de Recherche CNRS, Institut MICA, Hanoi, Examineur

**M. François PELLEGRINO**

Directeur de Recherche CNRS, Université Lyon 2, Examineur

**M. Laurent BESACIER**

Professeur à l'Université Joseph Fourier, Grenoble, Directeur de thèse





*“Without language, one cannot talk to people and understand them; one cannot share their aspirations, grasp their history, appreciate their poetry, or savour their songs”*

Nelson Mandela, *Long Walk to Freedom*

# *Abstract*

Languages in Malaysia are dying at an alarming rate. As of today, 15 languages are in danger while two languages are extinct. One of the methods to save languages is by documenting languages, but it is a tedious task when perform manually.

Automatic Speech Recognition (ASR) system can be a tool to help speed up the process of documenting speech from the native speakers. However, building ASR systems for a target language requires a large amount of training data as current state-of-the-art techniques are based on empirical approach. Hence, there are many challenges in building ASR for languages that have limited data available.

The main aim of this thesis is to investigate the effects of using data from closely-related languages to build ASR for low-resource languages in Malaysia. Past studies have shown that cross-lingual and multilingual methods could improve performance of low-resource ASR. In this thesis, we try to answer several questions concerning these approaches: How do we know which language is beneficial for our low-resource language? How does the relationship between source and target languages influence the performance of a speech recognizer? Is pooling language data an optimal approach for multilingual strategy?

Our case study is Iban, an under-resourced language spoken in Borneo island. We study the effects of using data from Malay, a local dominant language which is close to Iban, for developing Iban ASR under different resource constraints. We have proposed several approaches to adapt Malay data to obtain pronunciation and acoustic models for Iban speech.

While building a pronunciation dictionary from scratch is time consuming, as one needs to properly define the sound units of each word in a vocabulary, we developed a semi-supervised approach to quickly build a pronunciation dictionary for Iban. It was based on bootstrapping techniques for improving Malay data to match Iban pronunciations.

To increase the performance of low-resource acoustic models we explored two acoustic modelling techniques, the Subspace Gaussian Mixture Models (SGMM) and Deep Neural Networks (DNN). We performed cross-lingual strategies using both frameworks for adapting out-of-language data to Iban speech. Results show that using Malay data is beneficial for increasing the performance of Iban ASR. We also tested SGMM and DNN to improve low-resource non-native ASR. We proposed a fine merging strategy for obtaining an optimal multi-accent SGMM. In addition, we developed an accent-specific DNN using native speech data. After applying both methods, we obtained significant improvements in ASR accuracy. From our study, we observe that using SGMM and DNN for cross-lingual strategy is effective when training data is very limited.

# Résumé

Les langues en Malaisie meurent à un rythme alarmant. À l'heure actuelle, 15 langues sont en danger alors que deux langues se sont éteintes récemment. Une des méthodes pour sauvegarder les langues est de les documenter, mais c'est une tâche fastidieuse lorsque celle-ci est effectuée manuellement.

Un système de reconnaissance automatique de la parole (RAP) serait utile pour accélérer le processus de documentation de ressources orales. Cependant, la construction des systèmes de RAP pour une langue cible nécessite une grande quantité de données d'apprentissage comme le suggèrent les techniques actuelles de l'état de l'art, fondées sur des approches empiriques. Par conséquent, il existe de nombreux défis à relever pour construire des systèmes de transcription pour les langues qui possèdent des quantités de données limitées.

L'objectif principal de cette thèse est d'étudier les effets de l'utilisation de données de langues étroitement liées, pour construire un système de RAP pour les langues à faibles ressources en Malaisie. Des études antérieures ont montré que les méthodes inter-langues et multilingues pourraient améliorer les performances des systèmes de RAP à faibles ressources. Dans cette thèse, nous essayons de répondre à plusieurs questions concernant ces approches: comment savons-nous si une langue est utile ou non dans un processus d'apprentissage trans-langue ? Comment la relation entre la langue source et la langue cible influence les performances de la reconnaissance de la parole ? La simple mise en commun (pooling) des données d'une langue est-elle une approche optimale ?

Notre cas d'étude est l'iban, une langue peu dotée de l'île de Bornéo. Nous étudions les effets de l'utilisation des données du malais, une langue locale dominante qui est proche de l'iban, pour développer un système de RAP pour l'iban, sous différentes contraintes de ressources. Nous proposons plusieurs approches pour adapter les données du malais afin obtenir des modèles de prononciation et des modèles acoustiques pour l'iban.

Comme la construction d'un dictionnaire de prononciation à partir de zéro nécessite des ressources humaines importantes, nous avons développé une approche semi-supervisée pour construire rapidement un dictionnaire de prononciation pour l'iban. Celui-ci est fondé sur des techniques d'amorçage, pour améliorer la correspondance entre les données du malais et de l'iban.

Pour augmenter la performance des modèles acoustiques à faibles ressources, nous avons exploré deux techniques de modélisation : les modèles de mélanges gaussiens à sous-espaces (SGMM) et les réseaux de neurones profonds (DNN). Nous avons proposé, dans ce cadre, des méthodes de transfert translingue pour la modélisation acoustique permettant de tirer profit d'une grande quantité de langues "proches" de la langue cible d'intérêt. Les résultats montrent que l'utilisation de données du malais est bénéfique pour augmenter les performances des systèmes de RAP de l'iban. Par ailleurs, nous avons également adapté les modèles SGMM et DNN au cas spécifique de la transcription automatique de la parole non native (très présente en Malaisie). Nous avons proposé une approche fine de fusion pour obtenir un SGMM multi-accent optimal. En outre, nous avons développé un modèle DNN spécifique pour la parole accentuée. Les deux approches permettent des améliorations significatives de la précision du système de RAP. De notre étude, nous observons que les modèles SGMM et, de façon plus surprenante, les modèles DNN sont très performants sur des jeux de données d'apprentissage en quantité limités.

# *Acknowledgements*

First and foremost, I would not have been able to complete this study without the guidance of my thesis advisor, Laurent Besacier. His patience, support and motivation helped me in all time of research and writing this thesis.

My sincere thanks also goes to my thesis committee: Yannick Estève, Denis Jouvét, George Quenot, Eric Castelli and François Pellegrino, for their insightful comments and encouragement.

My sincere gratitude to Benjamin Lecouteux for his technical advice and Solange Rossato for her advice on linguistics. I would also like to thank Tan Tien Ping from Universiti Sains Malaysia for his help in my work during his sabbatical leave in LIG.

I thank RTM Sarawak for providing news data and those who have helped me in preparing the Iban speech transcription data: Suhaila Sae, Jennifer Wilfred Busu, Jonathan Sidi, Wima Nur Syahada, Ameer Joan, Doris Francis, Rosita Jubang, Imor Langgu and Anna Durin. I would also like to thank the staff from Universiti Malaysia Sarawak (Unimas) for their assistance in the data collection workshop.

I would also like to thank the Ministry of Higher Education of Malaysia and Unimas for their support in pursuing my Phd study in France.

Many thanks to my labmates (past and present): Luong Ngoc Quang, David Blachon, Frédéric Aman, Nadia Derbas, Marwen Azizi, Pathathai Na Lumpoon, Uyanga Sukhbaatar, Pedro Chahuara, Le Ngoc Tien, Mateusz Budnik, Andrew Tan and Elodie Gauthier. My Phd journey would not have been a colourful one if it was not for our adventures inside and outside the lab. I will definitely miss our coffee breaks and lunches together.

Last but not least I would like to thank Dennis Wong, my families and friends for their endless support and patience. Especially to my parents, I thank them for teaching me to always speak my mind, be strong and follow my dreams.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>Résumé</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>vi</b>
<b>Contents</b>	<b>vii</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Abbreviations</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
<b>I State-of-the-art and databases</b>	<b>7</b>
<b>2 Main Concepts in Automatic Speech Recognition</b>	<b>9</b>
2.1 Introduction . . . . .	9
2.2 Design of an automatic speech recognition system . . . . .	9
2.2.1 Acoustic signal analyzer . . . . .	10
2.2.2 Decoder . . . . .	11
2.3 Acoustic Modelling . . . . .	13
2.3.1 Hidden Markov Model . . . . .	13
2.3.2 Gaussian Mixture Models . . . . .	15
2.3.3 Subspace Gaussian Mixture Models . . . . .	16
2.3.4 Deep Neural Networks . . . . .	17
2.3.5 Acoustic modelling units . . . . .	19
2.3.6 Speaker adaptation . . . . .	21
2.4 Pronunciation Modelling . . . . .	21
2.4.1 An example of a G2P toolkit: Phonetisaurus . . . . .	23
2.5 Language Modelling . . . . .	24
2.5.1 $N$ -gram model . . . . .	24
2.5.2 Recurrent Neural Network model . . . . .	26
2.6 Evaluation . . . . .	27



2.6.1	G2P performance . . . . .	28
2.6.2	Language model performance . . . . .	28
2.6.3	ASR Performance . . . . .	28
2.7	Speech Recognition Toolkits . . . . .	29
2.7.1	Sphinx . . . . .	29
2.7.2	RASR . . . . .	30
2.7.3	Kaldi . . . . .	30
2.8	Summary . . . . .	31
<b>3</b>	<b>Challenges and recent approaches for ASR in under-resourced languages</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	The challenges . . . . .	34
3.2.1	Engaging the community . . . . .	34
3.2.2	Social and cultural influence . . . . .	35
3.2.3	Ethics . . . . .	35
3.2.4	Data collection procedures . . . . .	36
3.2.5	Low-resource models . . . . .	36
3.3	Brief background on research initiatives . . . . .	37
3.4	Most recent techniques in cross-lingual acoustic modelling . . . . .	39
3.4.1	Subspace Gaussian Mixture Model . . . . .	39
3.4.1.1	Cross-lingual SGMM for low-resource ASR . . . . .	39
3.4.1.2	Using out-of-language data for improving ASR for under-resourced language: a case study of Afrikaans . . . . .	40
3.4.2	Deep Neural Networks . . . . .	42
3.4.2.1	Shared-hidden-layer Multilingual DNN for ASR . . . . .	42
3.4.2.2	Cross-lingual and Multilingual DNN experiments using GlobalPhone corpus . . . . .	44
3.5	Summary . . . . .	47
<b>4</b>	<b>Languages from Malaysia</b>	<b>49</b>
4.1	Introduction . . . . .	49
4.2	Malay language . . . . .	50
4.2.1	Phonology . . . . .	50
4.2.1.1	Vowel phonemes . . . . .	50
4.2.1.2	Consonant phonemes . . . . .	52
4.2.1.3	Diphthongs . . . . .	54
4.2.2	Writing system . . . . .	54
4.3	Iban language . . . . .	56
4.3.1	Phonology . . . . .	57
4.3.1.1	Vowel phonemes . . . . .	57
4.3.1.2	Consonant phonemes . . . . .	57
4.3.1.3	Vowel clusters . . . . .	57
4.3.2	Writing system . . . . .	58
4.4	Malay and Iban relationship . . . . .	59
4.5	Data for ASR tasks . . . . .	62
4.5.1	MASS . . . . .	62

4.5.2	TED-LIUM . . . . .	63
4.5.3	Iban . . . . .	64
4.5.4	Non-native English . . . . .	64
4.6	Summary . . . . .	65
<b>II Data collection methodology and Iban automatic speech recognition development</b>		<b>67</b>
<b>5</b>	<b>Rapid development of Iban resources</b>	<b>69</b>
5.1	Introduction . . . . .	69
5.2	Quick collection of text and speech data . . . . .	69
5.2.1	Initial available data . . . . .	69
5.2.2	Transcribing speech corpus through a collaborative workshop . . . . .	70
5.2.3	Collecting large amount of text data . . . . .	70
5.3	Semi-supervised approach for building the Iban pronunciation dictionary . . . . .	71
5.3.1	Bootstrapping grapheme-to-phoneme (G2P) conversion . . . . .	72
5.3.2	Deeper analysis on our Iban lexicon . . . . .	72
5.3.3	Measuring pronunciation distance of Malay and Iban . . . . .	73
5.3.3.1	Obtaining a Malay G2P . . . . .	73
5.3.3.2	Evaluating using Levenshtein distance . . . . .	74
5.3.4	Obtaining Iban G2P training data via post-editing Malay G2P output . . . . .	74
5.3.5	Phonetization of the whole Iban lexicon . . . . .	76
5.4	Summary . . . . .	77
<b>6</b>	<b>First Iban ASR system developed</b>	<b>79</b>
6.1	Introduction . . . . .	79
6.2	Setup for experimentation . . . . .	79
6.2.1	Text and speech corpus . . . . .	79
6.2.2	Pronunciation dictionaries . . . . .	79
6.2.2.1	Analyzing phoneme sequences . . . . .	80
6.3	Baseline Iban speech recognizers . . . . .	81
6.3.1	Results and analysis . . . . .	81
6.3.1.1	Correlation between ASR and G2P performances . . . . .	82
6.3.1.2	System combination . . . . .	82
6.3.1.3	Confusion pairs . . . . .	83
6.4	Experimenting with cross-lingual and grapheme-based phonetizers for Iban ASR . . . . .	85
6.5	Solving text normalization issues . . . . .	86
6.5.1	Obtaining target words and candidates . . . . .	88
6.5.2	Creating and publishing the online survey . . . . .	89
6.5.3	Impact on ASR results . . . . .	90
6.6	Summary . . . . .	91

<b>III Cross-lingual acoustic modelling for bootstrapping ASR in very low-resource settings</b>	<b>93</b>
<b>7 Using resources from a closely-related language to develop ASR for a very under-resourced language</b>	<b>95</b>
7.1 Introduction . . . . .	95
7.2 Related work . . . . .	95
7.2.1 Motivation . . . . .	97
7.3 Cross-lingual approaches for building Iban acoustic models . . . . .	98
7.3.1 Training ASR (SGMM, DNN) on Iban data only . . . . .	98
7.3.2 Out-of-language resources for Iban ASR . . . . .	99
7.3.2.1 Closely-related language (Malay) data . . . . .	99
7.3.2.2 Non closely-related language (English) data . . . . .	100
7.3.3 Cross-lingual SGMM using monolingual and multilingual data . . . . .	100
7.3.4 Language-specific top layer for DNN . . . . .	101
7.4 Towards zero-shot ASR using a closely-related language . . . . .	103
7.5 Summary . . . . .	104
<b>8 Fine merging of native and non-native speech for low-resource accented ASR</b>	<b>105</b>
8.1 Introduction . . . . .	105
8.2 Related work . . . . .	105
8.3 Experimental Setup . . . . .	107
8.3.1 Data . . . . .	107
8.3.2 Baseline systems . . . . .	108
8.4 Language weighting for multi-accent Subspace Gaussian Mixture Models . . . . .	109
8.4.1 Proposed Method . . . . .	109
8.4.2 Results . . . . .	110
8.5 Accent-specific top layer for DNN . . . . .	111
8.5.1 Proposed Method . . . . .	111
8.5.2 Results . . . . .	112
8.6 Summary . . . . .	113
<b>9 Conclusions and Future Directions</b>	<b>115</b>
9.1 Conclusions . . . . .	115
9.2 Future Directions . . . . .	118
<b>A Personal Bibliography</b>	<b>121</b>
<b>B Reference to International Phonetic Alphabets</b>	<b>127</b>
<b>C Converting Iban words using P2P system</b>	<b>129</b>
<b>D Online Survey System for Data Normalization</b>	<b>131</b>
<b>Bibliography</b>	<b>133</b>

# List of Figures

2.1	A schematic representation of the components of a speech recognition system and the types of data for building them. . . . .	10
2.2	A three-state left-to-right HMM model. . . . .	14
2.3	SGMM parameters for a context-dependent HMM. Each state is defined by a low-dimensional vector $v_{jm}$ . The generic model UBM, $\mathbf{M}_i$ and $\mathbf{w}_i$ are globally shared. . . . .	16
2.4	An HMM/DNN architecture with $L$ hidden layers. The final layer outputs probability distribution of each HMM state. . . . .	18
2.5	A phonetic decision tree with yes/no linguistic question in each node. . .	20
3.1	An example of how Malaysians use more than one language in conversation. This post went viral in social media in 2013. . . . .	35
3.2	Architecture of the shared-hidden-layer multilingual DNN. Reprinted from Huang et al. [2013]. . . . .	43
4.1	Language Map of Malaysia and Brunei with language distribution. Reprinted from Ethnologue with permission. . . . .	51
4.2	Family tree of the Austronesian languages - focusing on Malay and Iban .	52
5.1	General process of our bootstrapping approach . . . . .	71
5.2	Steps to develop an Iban phonetizer through bootstrapping approach . . .	75
6.1	Iban ASR vs G2P based on WER (%) results . . . . .	83
6.2	Example of target word and [candidates] . . . . .	89
6.3	Bar graph showing total amount of words based on number of candidates	89
7.1	Minimum, maximum and average results for 1h and 7h Iban ASR using cross-lingual SGMM with different number of substates applied. Test data (1h) . . . . .	100
7.2	Process of obtaining language-specific DNN for Iban (right) using hidden layers from DNN trained on out-of-language data . . . . .	102
8.1	An Illustration of UBM merging through language weighting . . . . .	109
8.2	Min, max and average performance ( WER (%) ) of multi-accent SGMM based on language weighting strategy for non-native ASR (4h test). Note: non-native ( $L_1$ ) native ( $L_2$ ) and $\alpha = 0.1, ..., 0.9$ . . . . .	110
B.1	The International Phonetic Alphabets chart [IPA] . . . . .	128
C.1	Phoneme error rates (PER%) for 500 pure Iban words (500I) . . . . .	129

---

D.1	The welcome page of the survey system . . . . .	131
D.2	Questions in the survey. User will need to scroll down to see more questions.	132
D.3	User is allowed to continue the task later or click submit once they have completed. . . . .	132

# List of Tables

3.1	A summary of results by Lu et al. [2014] for cross-lingual SGMM experiments using German as the target language. . . . .	41
3.2	Results (phoneme accuracy - PA) for HMM/GMM, KL-HMM, Tandem and SGMM extracted from Imseng et al. [2014] for cross-lingual ASR through exploiting out-of-language data to improve Afrikaans ASR . . . .	42
3.3	DNN results obtained by Huang et al. [2013] in their experiment for cross-language knowledge transfer using shared-hidden-layers (SHL-MDNN) for ASR with limited conditions. Target languages : American English (ENU) and Mandarin Chinese (CHN) . . . . .	42
3.4	Tandem and Hybrid WER results on German [Swietojanski et al., 2013] .	45
3.5	DNN results based on WER for German ASR with dropout and multilingual effects [Miao and Metze, 2013] . . . . .	45
3.6	Summary of WER (%) results by Vu et al. [2014] for multilingual DNN using modified phone sets and Kullback-Leibler divergence method. Note that MUL* indicates that the multilingual datasets vary for Portuguese and Czech, Hausa, Vietnamese experiments. . . . .	46
4.1	Malay consonant phonemes . . . . .	53
4.2	List of phonemes and alphabets used in Malay . . . . .	55
4.3	Iban consonant phonemes . . . . .	58
4.4	Iban vowel clusters and examples . . . . .	58
4.5	List of phonemes and alphabets used in Iban . . . . .	59
4.6	Iban and Malay common and different phonemes . . . . .	60
4.7	Malay-Iban examples with their pronunciations . . . . .	60
4.8	Level of cognacy for eight languages calculated using normalized Levenshtein Distance method. As presented in [Ng et al., 2009] . . . . .	62
4.9	Size of the Malay data set for different speaker origins (ethnicities) in hours, number of sentences and number of speakers . . . . .	63
4.10	TED-LIUM corpus information . . . . .	64
4.11	Amount of Iban manually transcribed speech . . . . .	64
4.12	Overview of the non-native (Malaysian) English corpus . . . . .	65
5.1	Number of identical (same surface form) words found in Iban and Malay lexicons - size of Iban lexicon (36K) . . . . .	72
5.2	Malay G2P and Iban G2P performances for Iban phonetization task . . .	76
5.3	Performance of Malay and Iban phonetizers based on phoneme error rate (PER) and word error rate (WER) (%) . . . . .	77

6.1	Comparison results between two pronunciation dictionaries (total words 36K) . . . . .	80
6.2	Word statistics in Table 6.1 based on three languages . . . . .	81
6.3	Iban ASR performances based on word error rate (WER%) for different approaches . . . . .	82
6.4	System combination and WERs . . . . .	83
6.5	Top ten confusion pairs from Hybrid, Malay, Iban systems and system combination . . . . .	84
6.6	Iban ASRs performances (WER%) after using five different pronunciation dictionaries. . . . .	86
6.7	Old and new ASR results for speaker adapted (fmMLR) based on WER (%) . . . . .	90
6.8	Words that were replaced with options chosen by respondent . . . . .	90
7.1	Monolingual ASR results (WER%) on our Iban (1h) test set - no speaker adaptation at this stage. . . . .	99
7.2	Results of cross-lingual SGMM (WER %) in 1h and 7h systems on 1h test data. . . . .	100
7.3	WERs of cross-lingual DNNs - with speaker adaptation . . . . .	102
7.4	Performance of Iban ASR with Supervised and Unsupervised transcripts - Training on 7h of Iban speech . . . . .	103
8.1	Statistics of the non-native speech data for ASR. . . . .	107
8.2	Word error rates (WER %) of ASR with non-native (2h) and native (118h) acoustic models on non-native evaluation data (4h test). . . . .	108
8.3	A summary of results from the SGMM experiments on non-native ASR (4h test). Different UBMs were employed for building SGMM with 2h of non-native training data. . . . .	111
8.4	WERs of accent-specific DNN on non-native ASR task (4h test). . . . .	112
C.1	G2P and P2P performances for an Iban phonetization task . . . . .	130

# Abbreviations

<b>ASR</b>	<b>A</b> utomatic <b>S</b> peech <b>R</b> ecognition
<b>G2P</b>	<b>G</b> rapheme-to( <b>2</b> )- <b>P</b> honeme
<b>HMM</b>	<b>H</b> idden <b>M</b> arkov <b>M</b> odel
<b>GMM</b>	<b>G</b> aussian <b>M</b> ixture <b>M</b> odel
<b>LDA</b>	<b>L</b> inear <b>D</b> iscriminant <b>A</b> nalysis
<b>MLLT</b>	<b>M</b> aximum <b>L</b> ikelihood <b>L</b> inear <b>T</b> ransform
<b>fMLLR</b>	feature-space <b>M</b> aximum <b>L</b> ikelihood <b>L</b> inear <b>R</b> egression
<b>SGMM</b>	<b>S</b> ubspace <b>G</b> aussian <b>M</b> ixture <b>M</b> odel
<b>UBM</b>	<b>U</b> niversal <b>B</b> ackground <b>M</b> odel
<b>DNN</b>	<b>D</b> eep <b>N</b> eural <b>N</b> etwork
<b>RBM</b>	<b>R</b> estricted <b>B</b> oltzmann <b>M</b> achines
<b>DBN</b>	<b>D</b> eep <b>B</b> elief <b>N</b> etwork
<b>EM</b>	<b>E</b> xpected <b>M</b> aximization
<b>IPA</b>	<b>I</b> nternational <b>P</b> honetic <b>A</b> lphabets
<b>OOV</b>	<b>O</b> ut- <b>O</b> f- <b>V</b> ocabulary
<b>WER</b>	<b>W</b> ord <b>E</b> rror <b>R</b> ate
<b>PER</b>	<b>P</b> honeme <b>E</b> rror <b>R</b> ate





# Chapter 1

## Introduction

Malaysia is a multicultural and a multilingual country with a record of 140 languages recorded so far. But recent studies shows that two languages have gone extinct and 15 more are in a dying stage. The two extinct languages, Lelak and Seru, were found in Sarawak, one of the Malaysian states in Borneo [Lewis et al., 2014]. In this particular state, there are still 47 living languages spoken but most of them are unwritten. In fact one of them, Punan Batu, is regarded as a dying language in Malaysia and it is estimated that only 30 known speakers are left today [Wurm, 2008].

There are several organisations that help to preserve, revitalize or maintain languages such as the Tun Jugah Foundation<sup>1</sup> in Sarawak, active in documenting the native cultures and oral histories in the state. Among the common method used for documenting interviews with native speakers is by transcribing the recorded speeches to verbatim transcripts manually. A tedious task when one needs to process many hours of speech and it could result having transcribed data with many human error. Following this challenge, automatic speech recognition (ASR) system could be utilized as a tool to reduce human error and also to speed up the process of obtaining speech transcription data. Despite this, there are not many ASRs localized for Malaysian languages. In fact, research on automatic speech recognition system for Malaysian languages is still far behind compared to studies involving dominant languages in the world such as English, French, Mandarin, Spanish, etc.

---

<sup>1</sup><http://tunjugahfoundation.org.my/>

Prior to this thesis, Malay is the only Malaysian language often exploited for ASR research (see works on Malay ASR, e.g. [Tan et al., 2009], [Ong and Ahmad, 2011], [Goh and Ahmad, 2005]). The language is largely used for communication (formal and informal) thus large amount of resources could be collected for conducting natural language processing (NLP) tasks. If Malay speech and text corpora exist, would it be possible to use the data from other spoken languages in the country?

At present, state-of-the-art ASR systems are designed based on statistical methods thus require large amount of data (speech and text) to train statistical models. This requirement is difficult to satisfy if one intends to build a system for a language that has low amount of native speakers. The challenge increases if the language is very poorly documented or has no writing material at all; such language is defined as an under-resourced language [Berment, 2004]. A great deal of work would be needed just for setting up the training data and the effort could need a huge amount of money and time invested.

Challenges related to building ASR systems for under-resourced languages have gained speech community's attention over the past ten years. In addition to training under low-resource constraints, social and cultural influences, ethics, code-switching and data collection procedures are among the issues raised by Besacier et al. [2014] in their survey on ASR for under-resourced languages. As of today, researchers have developed strategies to address these issues mostly at the technical level. For example, cross-lingual acoustic modelling has significantly improved ASR recognition accuracies for low-resource languages. The idea behind this approach is to adapt training data from other sources, researchers frequently used data from other languages, to the training data of the under-resourced language. There are many ways to do this such as using phone mapping strategy or porting model parameters to the target system. There are also ideas for obtaining language independent speech recognition systems by training an acoustic model on multilingual data and using universal phone set. However, how do we know which language is beneficial for our under-resourced language if we want to employ such approaches? Does the relationship between source and target languages influence speech recognition performance?

The idea of using multilingual approach has also been applied in ASR for non-native speech. To obtain a reasonably good acoustic model for accented speech is a

challenge as it needs large amount of non-native speech data to model the acoustic probabilities. Thus, researchers have recently proposed multi-accent acoustic models based on multilingual approach mentioned above. The multi-accent models are generally trained on native (large corpus) and non-native (small corpus) data. The approach has improved the performance of non-native ASR with limited training data. Nevertheless, there are several questions concerning this method. Is pooling unbalanced corpora for training a model an optimal approach? Could we *finely* merge a large amount of native speech with a small quantity of non-native data for achieving an optimized multi-accent model?

Therefore, in this thesis we try to answer the above questions by:

- Studying the effects of using resources from closely-related languages to build an ASR for an under-resourced language in Malaysia. We will study speech recognition for Iban, a language that is largely spoken in Sarawak. The language is close to Malay in terms of phonology and orthography.
- Proposing cross-lingual and multilingual/multi-accent strategies for boosting and optimizing the performance of our low-resource systems. Recent studies have shown that Subspace Gaussian Mixture Models (SGMM) and Deep Neural Networks (DNN) can outperform conventional Gaussian Mixture Models (GMM) in terms of speech recognition performance. The shared parameters, notably the Universal Background Model (UBM) in SGMM and hidden layers in DNN frameworks are transferable. We employ both frameworks to investigate our strategies.

This thesis reveals techniques which are very important for ASR development involving under-resourced or new languages. The primary contributions of this thesis are:

- We build the first Iban speech and text corpora for ASR experiments
- We develop strategies to quickly build the Iban pronunciation dictionary for ASR
- To study cross-lingual and multilingual effects on Iban ASR, we propose methods to observe the impact of using Malay and non closely-related language data.

- We develop a zero-shot ASR system through unsupervised training on Iban data. In most cases, speech transcripts for training an ASR system are not available. We hypothesized that a Malay ASR built using Malay and Iban data could generate the Iban transcripts.
- To optimize multi-accent acoustic model in ASR for non-native speech, we introduce a language weighting strategy for merging native and non-native models that have been trained on unbalanced data. We study the shared parameters in SGMM for developing our strategy.
- We build language-specific and accent-specific DNNs for our Iban and non-native ASR systems. We study methods that could improve the performance of our low-resource DNN systems using DNNs trained on Malay or native English data. In the course of obtaining language(accent)-specific DNN we employ a strategy that handles speaker adapted DNNs.

The rest of the chapters in this thesis are divided into three parts; state-of-the-art and databases (chapters 2, 3 and 4), data collection methodology and baseline Iban ASR development (chapters 5 and 6) and cross-lingual acoustic modelling for bootstrapping ASR in low-resource settings (chapters 7 and 8).

## **Part 1: State-of-the-art and Databases**

In Chapter 2, we will give a brief introduction to the field of ASR. Here, we intend to provide the concepts and terminology related to this field that are important for understanding the experiments that we carried out. Chapter 3 discusses the challenges associated to ASR for under-resourced languages and we describe some recent works by researchers to address the problems. Also in this chapter, we highlight several important works which are the motivation behind the work of this thesis. Chapter 4 which is the last chapter in this part, introduces the languages in Malaysia but focusing on Iban, the under-resourced language, and Malay, the closely-related language. Here, we present the phonological descriptions of the two languages and provide some related examples. Moreover, we discuss about the strong relationship between the two languages in several aspects. At the end of this chapter, we present the characteristics of the databases which are used in our experiments.

## **Part 2: Data Collection Methodology and Baseline Iban ASR Development**

In the second part of this thesis, Chapter 5 presents our steps in gathering Iban speech and text corpora for building our ASR systems. We describe the resources and demonstrate our strategy to quickly build an Iban pronunciation dictionary for ASR. Chapter 6 details our approaches to build the Iban ASR using the acquired data. This chapter also includes evaluation of several Iban pronunciation dictionaries on ASR.

## **Part 3: Cross-lingual Acoustic Modelling for Bootstrapping ASR in Low-resource Settings**

The final part consists two chapters which describe our contributions for boosting performance of low-resource ASRs using cross-lingual strategies based on SGMM and DNN frameworks. In Chapter 7, we extend this approach to the recently acquired Iban ASR. We present results that show the value of using closely-related data in our cross-lingual strategy as well as development of the zero-shot ASR. Then in Chapter 8, we describe a novel approach for obtaining an optimized multi-accent acoustic model. Here, we investigate an ASR system for Malaysian English. We show a language weighting strategy to merge model parameters in existing native and non-native acoustic models.



## Part I

# State-of-the-art and databases





## Chapter 2

# Main Concepts in Automatic Speech Recognition

### 2.1 Introduction

An automatic speech recognition (ASR) system is used for converting human speech signal into readable text. The process of treating the speech signals and then match them to the right word sequence is complex and involves several mathematical concepts. This chapter describes the fundamentals of ASR and the current techniques used for building state-of-the-art systems. We present brief introductions to the methods as to declare some terminologies which will be used in the later chapters.

### 2.2 Design of an automatic speech recognition system

The aim of an ASR system is to recognize human speech and produce sequence of words that correspond to the input speech. This task is a pattern recognition problem, where the pattern to be categorized is the sequence of feature vectors extracted from input speech. The vector sequence is supposed to be assigned to the class which represents the correct word utterance that belongs to the pattern and this is selected from a set of all possible word sequences.

Figure 2.1 presents a block diagram representing the architecture of a state-of-the-art system for speech recognition. In the system, the main components are the acoustic

signal analyzer and decoder. The latter, needs data-driven models; acoustic model, pronunciation model and language model.

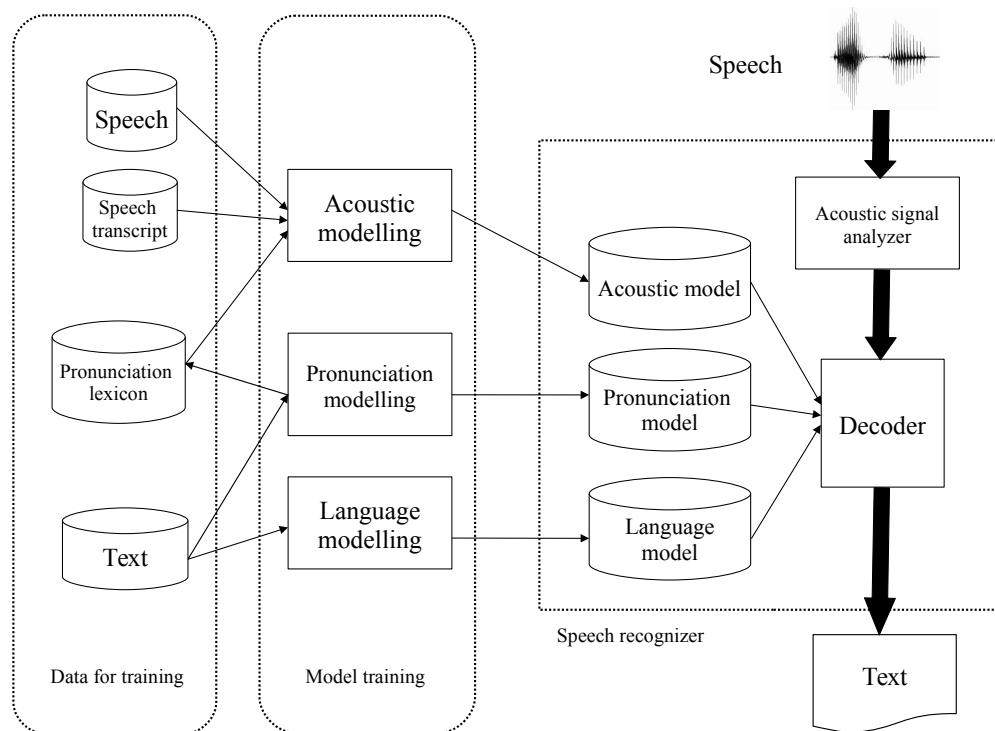


FIGURE 2.1: A schematic representation of the components of a speech recognition system and the types of data for building them.

### 2.2.1 Acoustic signal analyzer

The main purpose of an acoustic signal analyzer is to extract the feature vectors, given the raw speech signals. There are various feature extraction techniques such as the Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and Mel-frequency Cepstral (MFC) spectral analysis model. Among these, the latter technique is the most frequently used in speech recognition systems.

The calculation of the Mel-frequency Cepstral Coefficients (MFCC) [Davis and Mermelstein, 1980] can be carried out using the following steps:

- Use discrete Fourier Transform (DFT) to extract spectral information in each frame
- Apply a triangular mel filter bank to warp the frequencies output by the DFT onto a mel scale, a perceptual scale of pitches according to human listeners

- Smooth the spectrum using a log function
- Convert the spectrum to  $N$  cepstral coefficients using Discrete cosine transform (DCT)

The MFCC features are the amplitudes of the resulting spectrum. The energy feature from each analyzed frame can be added as  $(N + 1)^{th}$  feature. Apart from the raw MFCC features, dynamic features can be concatenated to represent temporal change between nearby frames [Furui, 1986]. Each feature is added with a delta feature and a double delta feature. The delta can be computed by estimating the difference between the frames; e.g.  $\Delta y_t = 0.5(y_{s,t+1} - y_{s,t-1})$  where  $y_{s,t}$  is the static cepstral feature vector at time  $t$ . The double delta feature vector  $\Delta^2 y_t$  can be derived using the same method in order to represent the change between frames in the delta features. Thus, the total number can end up with  $3(N + 1)$  MFCC features after appending the energy and also the delta and double delta features.

A more discriminative feature can be created by reducing the size of the feature vectors using feature reduction techniques such as linear discriminant analysis (LDA), which was first introduced by [Fisher, 1936]. It is the most widely used classification method in pattern recognition and machine learning. LDA seeks for the optimal transformation matrix as to keep most of the information useful to discriminate between the different classes in a reduced feature space. Applying this technique in speech recognition have improved recognition rates for many speech tasks.

Feature decorrelation technique based on maximum likelihood linear transform (MLLT) [Gopinath, 1998] can also be employed subsequently after LDA. The technique tries to minimize the loss in likelihood between full and diagonal covariance modelling.

### 2.2.2 Decoder

The decoder finds the most probable word sequence that matches with the incoming feature vectors extracted from the speech signals. This search can be achieved by finding the sequence of words  $W$  which gives maximum posterior probability  $P(W|X)$  given the observed feature vectors  $X$ . Using Bayes rule, the probability is calculated as follows:

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)}. \quad (2.1)$$

Therefore, the decoder finds the most likely word sequence  $\hat{W}$ :

$$\begin{aligned}
 \hat{W} &= \arg \max_W P(W|X) \\
 &= \arg \max_W \frac{P(X|W)P(W)}{P(X)} \\
 &= \arg \max_W P(X|W)P(W)
 \end{aligned} \tag{2.2}$$

where  $P(X|W)$  is the acoustic probability of the word sequence from an acoustic model and  $P(W)$  is the language probability of the word sequence from a language model. The search space for the decoder is typically constrained by a pronunciation dictionary that has a list of words for composing the word sequence.

Building the decoder for a speech recognition system requires a speech corpus along with its transcription, word lexicon and text corpus. For training an acoustic model, speech corpus and pronunciation dictionary are required. The former is obtained by recording human speech using a microphone and recordings are saved as digital audio files. Typically for a speech recognition task a resolution of 16bit and a sampling frequency of 16kHz are used. The speech transcripts can come from the text read by speakers in the data collection process, or from a transcription produced by human transcribers who know the language used in the speeches that have been collected.

The pronunciation dictionary contains words and their respective phoneme sequences. The words can be selected from texts or by pooling words available in the speech transcription data. Most importantly, the lexicon should have a high coverage of words for testing the ASR system. The orthographic representation of words is converted into sequence of phonemes. Normally, a trained linguist is required to carry out this task for obtaining phonemic transcripts. However, to convert a large vocabulary by hand requires a lot of time thus grapheme-to-phoneme tools can be used to automatically generate phoneme sequences for new entries. Besides using the pronunciation model to develop acoustic model, it is also used in the decoder for converting hypothesized phoneme sequences into words.

Besides acoustic and pronunciation models, another essential component for a decoder is the language model. An ASR system requires a model for defining the rules of combining words at the sentence level. For this,  $N$ -gram language model is widely used in most state-of-the art systems. In order to have a robust model, it must be trained on

a large text corpus. All of the components in an ASR decoder mentioned are constructed using statistical methods. Several methods are described in the next few sections.

## 2.3 Acoustic Modelling

There are many approaches for modelling acoustic units, such as the hidden Markov models (HMM), artificial neural network and template based model. For the past two decades, state-of-the-art speech recognition systems used HMM due to its robustness and cheap computational cost. Here, we describe the HMM and several methods for computing its emission probability.

### 2.3.1 Hidden Markov Model

To model the acoustic probability  $P(X|W)$ , one way is to use hidden Markov models (HMM). The models are based on Markov Chains, which is used for modelling sequence of events in time. The definition of a HMM [Rabiner, 1989] is described by the following components:

- Sequence of observations  $X = \{x_1, x_2, \dots, x_j, \dots, x_N\}$  (in case of discrete HMM, each one is taken from a vocabulary  $V = v_1, v_2, \dots, v_V$ ).
- Set of emitting states  $S = \{s_1, s_2, \dots, s_K\}$  as well as start and end states  $s_{Init}, s_{End}$  that are not connected with observations.
- State transition probability matrix  $A$ , where each element of the matrix  $a_{ij}$  is the transition probability from state  $i$  to  $j$  for  $i, j = 1, \dots, K$ .
- Probability density functions for estimating the probability of emitting an observation  $x_t$  from a state  $i$  at time  $t$ ,  $b_i(x_t) = p(x_t | s_t = i)$ .
- Initial state transition probability  $\pi = \{\pi_i = P(s_0 = i)\}$ .

There are several types of HMM such as the Ergodic model and Bakis model. The Ergodic HMM is a fully connected HMM where any states can be reached from any other state in  $N$  steps, thus there is a probability of moving between any two states. The Bakis model is a left-to-right HMM where the state transitions move from left to

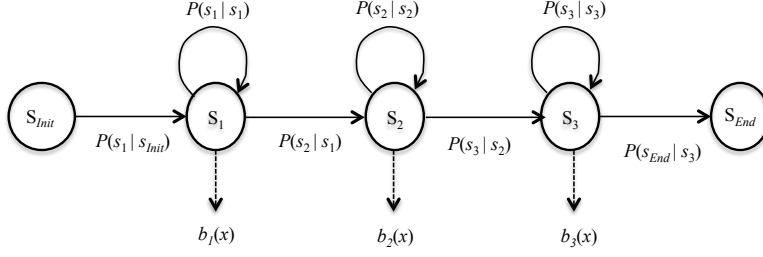


FIGURE 2.2: A three-state left-to-right HMM model.

right, as shown in Figure 2.2. For this type of HMM, there is no probability for the higher numbered state to proceed to a lower number state (e.g;  $s_3$  to  $s_1$ ). In speech applications, Bakis HMMs are generally used for modelling speech where each model represents a sub-word or phoneme (and states represent sub-phoneme information).

In the first-order HMM, there are two assumptions. The first assumption is a Markov assumption, where the probability of a given state relies only on the previous state:

$$p(s_t | s_1^{t-1}) = p(s_t | s_{t-1}). \quad (2.3)$$

The second assumption states that the probability of an output observation  $x_t$  relies only on the state  $s_t$  that emitted the observation. Thus the output independence assumption is defined as:

$$p(x_t | x_1^{t-1}, s_1^t) = p(x_t | s_t). \quad (2.4)$$

Given the above definition of HMM, there are three fundamental problems of interest:

1. Evaluation: Given the observation sequence  $X = x_1, x_2, \dots, x_T$  and an HMM  $\lambda = (A, B, \pi)$ , how do we compute the probability that the observation sequence was generated by the model  $p(X|\lambda)$ ? This problem can be efficiently solved using the Forward algorithm.
2. Decoding: Given an HMM  $\lambda = (A, B, \pi)$  and observation sequence  $X = x_1, x_2, \dots, x_T$ , how do we find the corresponding state sequence  $S$ ? This problem is related to continuous speech recognition and it is commonly solved by the Viterbi algorithm.
3. Learning: Given an HMM  $\lambda = (A, B, \pi)$ , how do we adjust the parameters  $A$  and  $B$  for maximizing the probability to observe  $X$ :  $\lambda^* = \arg\max_{\lambda} P(X|\lambda)$ ? This

problem must be solved, if we want to train an HMM for speech recognition. The Baum-Welch algorithm [Baum, 1972] (a special case of Expected-Maximization [Dempster et al., 1977] algorithm) is the standard algorithm for addressing this task.

There are several methods for computing the emission probability distribution  $B$  such as the Gaussian Mixture Models, subspace Gaussian Mixture Models and Deep Neural Networks. We will describe the three methods in the following subsections.

### 2.3.2 Gaussian Mixture Models

One of the methods for computing the emission probability is by using the Gaussian mixture model (GMM). The model estimates acoustic likelihoods with a weighted sum of  $M$  Gaussians component densities as given by the equation:

$$p(\mathbf{x}|j) = \sum_{i=1}^M w_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i), \quad (2.5)$$

where  $\mathbf{x} \in \mathbb{R}^D$  denotes the  $D$ -dimensional feature vector for HMM state  $j$ ,  $i$  is the Gaussian index,  $w_i$  are mixture weights that satisfy the constraint for a valid probability mass function:

$$\sum_{i=1}^M w_i = 1, \quad (2.6)$$

and  $\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)$  are the Gaussian component densities and each component density is a multivariate Gaussian function defined as,

$$\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)\right], \quad (2.7)$$

where  $\mu_i$  is the mean vector and  $\Sigma_i$  is the covariance matrix.

For many years, GMM has been used in state-of-the-art speech recognition systems due to its robustness and cheap computational cost. Over the past few years, the subspace Gaussian Mixture Models (SGMM) [Povey et al., 2011a] have been introduced and proven to outperform GMM in various speech recognition tasks.



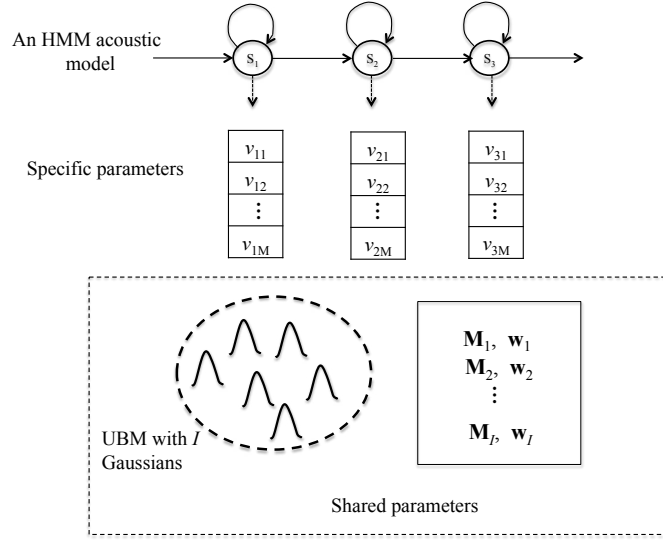


FIGURE 2.3: SGMM parameters for a context-dependent HMM. Each state is defined by a low-dimensional vector  $v_{jm}$ . The generic model UBM,  $\mathbf{M}_i$  and  $\mathbf{w}_i$  are globally shared.

### 2.3.3 Subspace Gaussian Mixture Models

The GMM and SGMM acoustic models are similar since each emission probability of each HMM state is modelled with a Gaussian mixture model. However, in the SGMM approach, the Gaussian means and mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections. For SGMM, the state probabilities are defined following the equations below:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^I w_{jmi} \mathcal{N}(\mathbf{x}|\mu_{jmi}, \Sigma_i) \quad (2.8)$$

$$\mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \quad (2.9)$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^I \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}} \quad (2.10)$$

Each state  $j$  can be modelled with a mixture of substates ( $m$  is the substate) and associated to a vector  $\mathbf{v}_{jm} \in \mathbb{R}^S$  ( $S$  is the phonetic subspace dimension) which derives the means,  $\mu_{jmi}$  and mixture weights,  $w_{jmi}$ , while  $I$  is the number of Gaussians for each state. The SGMM acoustic modelling for a context-dependent HMM is illustrated in Figure 2.3. The parameters of the model can be divided into state-specific ( $\mathbf{v}_{jm}$ ) and globally shared parameters. The latter consists of the phonetic subspace  $\mathbf{M}_i$ , weight

projections  $\mathbf{w}_i^T$  and covariance matrices  $\Sigma_i$ , which are common across all states. The SGMM system is initialized by a Universal Background Model (UBM) with  $I$  Gaussians, which is trained on all speech data.

The shared parameters and UBM are language independent [Burget et al., 2010] thus enables us to transfer them across systems or train on multilingual data simultaneously. Thus, the method is suitable for studying cross-lingual / multilingual effects at acoustic model level. Recently, several studies have shown the benefits of using SGMM to help ASR for under-resourced language and with low-resource setting. The studies will be discussed in Chapter 3.

### 2.3.4 Deep Neural Networks

Two decades ago, artificial neural networks (ANN) were introduced for computing the emission probabilities of HMM. Researchers then succeeded in predicting HMM states by using a single layer ANN that has non-linear hidden units. However, this structure was not efficient and it was not able to compete GMM. Furthermore, the hardware and algorithms back then were unable to train more hidden layers given a large amount of speech data. Now, graphical processing units (GPUs) are available to provide high computational power and better algorithms have emerged in machine learning. Therefore, Deep Neural Networks (DNNs) with many layers can be trained efficiently and it has been shown that the method outperforms GMM in many speech recognition tasks (see [Hinton et al., 2012a] for ASR results performed by DNN on well-known tasks).

Deep Neural Network for ASR is a feedforward neural network with hidden layers. Figure 2.4 depicts the DNN architecture used in current HMM systems. Mathematically, each output of the  $l$ -th layer of a DNN can be defined as

$$a_l = \sigma(b_l + \sum_i w_{il} z_i), \quad 1 \leq l < L \quad (2.11)$$

where  $w_{il}$  is the connection weight between unit  $a_l$  and  $z_i$ , where the latter is the output of the  $(l - 1)$ -th layer, while  $b_l$  is the bias. The hidden output unit  $a_l$  is a sigmoid function defined as  $\sigma(a) = (1 + \exp(-a))^{-1}$ . The last ( $L$ -th) layer of the DNN uses a softmax function to obtain the posterior probability of each HMM state  $j$  (clustered,

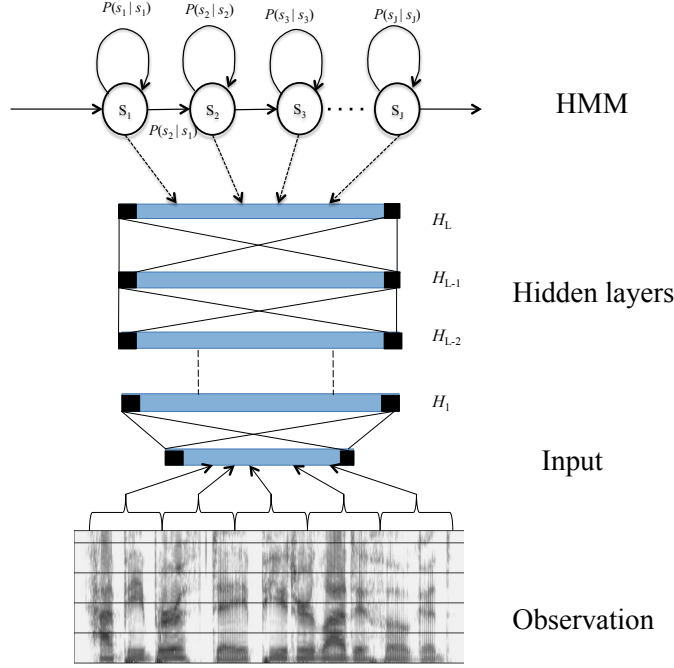


FIGURE 2.4: An HMM/DNN architecture with  $L$  hidden layers. The final layer outputs probability distribution of each HMM state.

context-dependent) given the acoustic observation  $\mathbf{x}_t$  at time  $t$ :

$$p(j|\mathbf{x}_t) = \frac{\exp(a_L)}{\sum_{j'} \exp(a_L)}. \quad (2.12)$$

Note that a DNN does not explicitly model  $p(\mathbf{x}_t|j)$  as needed by an HMM but the model can be achieved by applying the Bayes rule as in the following:

$$p(\mathbf{x}_t|j) = \frac{p(j|\mathbf{x}_t)p(\mathbf{x}_t)}{p(j)} \simeq \frac{p(j|\mathbf{x}_t)}{p(j)} \quad (2.13)$$

where  $p(j)$  is the prior probability of each state found in the training data. The observation probability  $p(\mathbf{x}_t)$  is a difficult distribution to estimate thus to perform decoding the term is ignored. The weights  $w_{il}$  are the trainable parameters of the DNN. During training, the weights are adjusted in order to minimize the cost function that estimates the error between the network outputs and the target outputs. The stochastic gradient descent method can be used to compute the derivatives of the cost function by optimizing the weights.

Deep Belief Network (DBN) has been proposed for addressing optimization issues when training many hidden layers directly with the gradient descent method. DBN

is a single, multi-layer generative model that is trained using Restricted Boltzmann Machines (RBM) [Hinton, 2010] and it is a product of a pretraining strategy that provides better initial weights for DNN. After pretraining, the final layer (softmax layer) is added and fine-tuning using backpropagation algorithm is carried out for obtaining a full DNN system. Recent studies have shown that using DBN to initialize a DNN with many layers leads to a better performing DNN compared to a DNN which had random initial weights ([Seide et al., 2011] [Dahl et al., 2012] [Mohamed et al., 2012]). Moreover, pretraining a DNN can be unsupervised which means that it can be trained on untranscribed data. This technique has been experimented by Swietojanski et al. [2013] where the authors pretrained the network on a large amount of untranscribed multilingual data. Furthermore, they showed that the parameters of the DBN are transferable, so a DBN can be used for initializing DNN of different systems. This means that DBNs are language independent - a similar concept with UBM in SGMM - which gives us an opportunity to try cross-lingual approach in DNN, too.

Besides DBN, the hidden layers from a fully trained network can also be used for DNN fine-tuning of other systems. Huang et al. [2013], Vu et al. [2014] and Miao and Metze [2013] showed that hidden layers trained on multilingual data are transferable and the performance of their targeted speech recognition systems increased significantly after employing this strategy. The above mentioned studies will be discussed in the next chapter.

### 2.3.5 Acoustic modelling units

As mentioned in Section 2.3.1, each state in an HMM model can represent a sub-word or a phoneme. Phonemic models are the preferred type for ASR because only a small amount of phonemes (less than 50) is normally used to define the sound units available in a language. In a three-state HMM, a phonemic model is defined by subphonemes that occur at the beginning, middle and end of the phoneme. For continuous HMMs, when the emission probabilities are computed by GMMs, each state (subphoneme) has its own mixture model. This kind of model is called a context-independent HMM or monophone model where a fixed GMM is used for a subphoneme. However, due to coarticulation effects, monophone models are not very accurate to model continuous speech. This is because phonemes in a word are produced

depending on their neighbouring phonemes. Hence, context-dependent HMMs are generally used in state-of-the-art speech recognition systems. Most commonly used context-dependent model is a triphone HMM, a model that defines phonemes in the context of the neighbouring left and right phonemes.

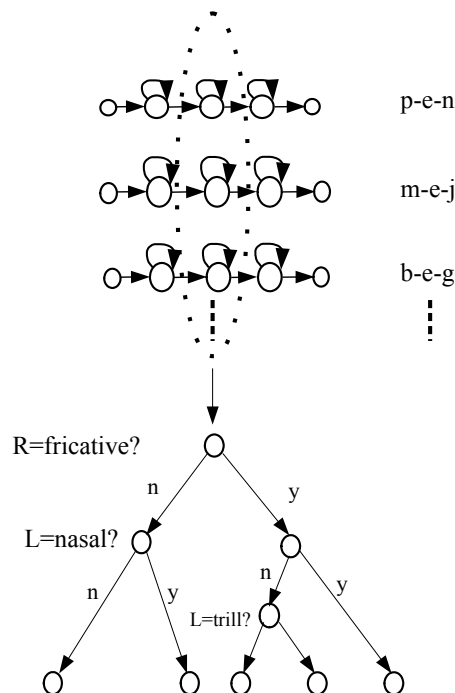


FIGURE 2.5: A phonetic decision tree with yes/no linguistic question in each node.

Assuming that a phoneme set for defining a language has 50 different phonemes, using triphones would lead to more than 100 000 triphones and this could include triphones that are not possible to occur or not available in the training corpus. Therefore, clustering context-dependent states using decision trees [Lee et al., 1990] is utilized to reduce the number of triphones that are not needed in training. A phonetic decision tree allows us to select which triphone states can be grouped together by imposing linguistic questions (nasal, fricative, etc.). Figure 2.5 depicts an example case of tying the middle states of all triphones of the phone /e/ in Malay language. This technique is commonly employed in ASR systems to robustly estimate the parameters of the GMM for every tied states.

### 2.3.6 Speaker adaptation

The motivation behind speaker adaptation techniques is that speaker-independent system performance can degrade when evaluated on speakers and environments that are less represented in the training corpus. Therefore, with speaker adaptation, the independent system is improved towards new speaker (speaker in test data) using a small amount of speaker-specific data. There are several known speaker adaptation techniques such as the vocal tract length normalization (VTLN) ([Welling et al., 1999], [Pitz and Ney, 2003]) and maximum likelihood linear regression (MLLR) [Leggetter and Woodland, 1995].

In the MLLR approach, the Gaussian means in HMM are adapted per speaker using an affine transform:

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b} \quad (2.14)$$

where  $\mathbf{A}$  and  $\mathbf{b}$  are the transformation parameters.

This standard approach has been modified to introduce a constrained (feature-space) MLLR [Gales, 1998] (cMLLR/fMLLR) approach where the same linear transform for both mean and covariance is given as follows:

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b} \quad (2.15)$$

$$\hat{\Sigma} = \mathbf{A}\Sigma\mathbf{A}^T. \quad (2.16)$$

Speaker adaptive training (SAT) can be implemented by training MLLR transforms for each training speaker [Povey et al., 2008]. Using this approach has shown better decoding results. However, the limitation of this approach is that it increases training complexity and requires storage space.

## 2.4 Pronunciation Modelling

Pronunciation of words vary due to differences in speaking style. For example, in casual speaking style, short words such as *I am*, *do not*, *I have* are concatenated when they are pronounced, as in *I'm*, *don't* and *I've*. Some words may have more than one pronunciation that depends on meaning, such as the word *sepak* in Malay that can

be pronounced as /sepak/ (kick) or /səpak/ (slap). Thus, ASR system requires a pronunciation model to model the variability found by speaking style and phonology of a language. The performance of an ASR system could degrade if the pronunciation model is not able to handle such phenomena in human speech.

A pronunciation model can be created by hand, that is, by writing out pronunciation transcripts given a list of words. Since phones or phonemes are used as acoustic modelling units in ASR, the pronunciation transcript can be directly obtained from a standard dictionary which has descriptions on how words should be pronounced.

However, not all dictionaries have this information especially for dictionaries in languages that are poorly documented. Therefore, in such case a linguist is required to assist in converting the graphemes to corresponding phonemes (G2P task) using International Phonetic Alphabets (IPA)<sup>1</sup>. However, not all of the IPA symbols are computer-readable. Phoneticians came up with SAMPA, which stands for Speech Assessment Methods Phonetic Alphabet, a standard machine-readable phonetic alphabet built in 1980s ([Wells et al., 1992], [Wells, 1997]). SAMPA was designed for handling only European Union Languages therefore it was extended to X-SAMPA (Extended-SAMPA), which can be applied for any languages [Wells, 1995]. Another computer-readable phonetic alphabet is called Arpabet for describing American English phonemes. For example, the phoneme set used in the CMU pronunciation dictionary<sup>2</sup> is based on Arpabet symbols.

Currently, there are several techniques for developing grapheme-to-phoneme (G2P) systems to automatically generate pronunciation transcripts. The main problem for a G2P can be formulated as:

$$P^* = \arg \max_P p(P, G) \quad (2.17)$$

which means, given a sequence of grapheme  $G$ , find the most likely phoneme sequence  $P^*$ . Several strategies have been proposed to address this problem such as employing Baum-Welch expected maximization (EM) algorithm for producing grapheme-phoneme alignments [Kneser, 2000], joint maximum entropy N-gram model [Chen, 2003], joint N-gram model ([Bisani and Ney, 2008] [Vozila et al., 2003]) and weighted finite states transducers (WFST) for decoding [Novak et al., 2011].

<sup>1</sup>see Appendix B for an IPA chart

<sup>2</sup><http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

### 2.4.1 An example of a G2P toolkit: Phonetisaurus

An open source G2P toolkit called Phonetisaurus<sup>3</sup> has been developed by Novak et al. [2011] who employed the WFST framework in the system’s architecture. To train a G2P model, the toolkit requires a base pronunciation dictionary. The first step to be carried out is to align sequences of grapheme and phoneme in the dictionary using an EM-based multiple-to-multiple alignment algorithm proposed by Jiampojamarn et al. [2007]. Then, the sequence of aligned grapheme-phoneme pairs are obtained by concatenating the grapheme and phoneme sequences. Below is an example of a sequence of aligned grapheme-phoneme pairs (second row) for the word ‘universiti’ produced by Phonetisaurus.

```
universiti -> j u n i v @ r s i t i
u}j|u n}n i}i v}v e}@ r}r s}s i}i t}t i}i
```

The pronunciation model is then built by training an  $N$ -gram model using the set of joint sequences acquired. For this case, language modelling toolkits such as SRILM [Stolcke, 2002] or MITLM [Bo-June and Glass, 2008] can be utilized. The output model is then converted to weighted finite state acceptors (WFSAs), which are later converted to WFTs to obtain input (grapheme) and output (phoneme) labels. Pronunciations for new entries will be generated by converting the target word into a FSA and then compose the acceptor using the pronunciation model. The toolkit returns hypothesized phoneme sequence through the following process,

$$H_{best} = nBest(\pi(oProj(W \circ M))) \quad (2.18)$$

where  $H_{best}$  contains a weighted list of pronunciation outputs,  $W$  denotes the FSA obtained from the new entry and  $M$  is the WFST acquired from the pronunciation model.  $oProj$  makes sure that only output labels are projected and  $\pi$  removes unwanted symbols such as skip, ‘\_’, epsilons,  $\langle s \rangle$  and  $\langle /s \rangle$ ; while  $nBest$  is a shortest-path algorithm for obtaining the best  $N$  hypotheses.

We will use this toolkit for creating G2P systems due to its capability to train models in a reasonable time. Besides that, it is able to generate  $N$ -best list of hypotheses, which

<sup>3</sup><https://code.google.com/p/phonetisaurus>



may be interesting if we want to include pronunciation variants into the pronunciation dictionary. Furthermore, the performance of this toolkit for a G2P task has been measured and compared with other G2P modelling techniques [Hahn et al., 2012]. It has been shown that Phonetisaurus along with Sequitur G2P<sup>4</sup>, another G2P converter based on joint N-gram model [Bisani and Ney, 2008], outperformed other methods in terms of phoneme accuracy.

## 2.5 Language Modelling

Recall that in Section 2.2.2, language model is one of the components of an ASR decoder for computing the probability of a word sequence,  $P(W)$ . The language model defines the language grammar by providing rules for concatenating words to become meaningful phrases. There are two types of language models; grammar-based language models or stochastic language models. The latter is used in ASR systems for continuous speech thus we will focus on several language modelling techniques for building this type of model.

### 2.5.1 N-gram model

The classical approach used for building a stochastic language model is the N-gram method. The method was proposed for calculating language probability over a large text (training) corpus. It tries to reduce the complexity of calculating the probability of a sentence  $P(W)$  given a series of words  $W = w_1, w_2, w_3, \dots, w_n$ . The problem can be written as:

$$\begin{aligned}
 P(W) &= P(w_1, w_2, w_3, \dots, w_n) \\
 &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\dots P(w_n|w_1, \dots, w_{n-1}) \\
 &= \prod_i^n P(w_i|w_1, w_2, \dots, w_{n-1})
 \end{aligned} \tag{2.19}$$

where the chain rule is applied for decomposing the joint probability into a set of conditional probabilities. The computational cost to calculate the sentence probability in this manner is very expensive and requires many training samples for estimating it.

---

<sup>4</sup><http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html>

Thus,  $N$ -grams were introduced as a feasible solution.  $N$ -gram language modelling is based on Markov assumption where the occurrence of a word can be estimated using only several previous words that appeared before the target word. A context unit is built by an  $N$ -gram along with  $N - 1$  preceding words, called history.  $N$  can be set depending on the constraint in the available text.

For speech recognition, a  $N$ -gram language model of order 2 is commonly used. This means that the trigrams of a sentence can be found by clustering three words that appear at the same time. The trigram probability is calculated using maximum likelihood approach and it is obtained by finding the count of a current  $N$ -gram and the value is normalized by the count of  $N$ -gram that shares the same  $N - 1$  history. The trigram probability is formulated as:

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \quad (2.20)$$

where  $C$  is called the  $N$ -gram count. Therefore the language probability in Equation 2.19 can be rewritten as:

$$P(W) = P(w_1)P(w_2|w_1) \prod_{i=3}^n P(w_i|w_{i-2}, w_{i-1}). \quad (2.21)$$

This equation is much simpler to calculate if we compare to the first problem defined for the language probability. However, the setback of using this model to estimate word sequence on a training corpus is that some  $N$ -grams have very small count (or not appearing at all) thus create a data sparsity problem. This may degrade decoding performance when the model is used in ASR systems where a test data may have sentences containing the unseen  $N$ -grams. Hence, smoothing methods were proposed to reduce this data sparsity issue in language model.

Smoothing methods for  $N$ -gram model that are based on discounting approach are such as the Good-Turing [Good, 1953], Witten-Bell [Witten and Bell, 1991] and Kneser-Ney [Kneser and Ney, 1995]. The goal of these methods is to reduce seen  $N$ -grams (frequently occurring in training corpus) counts and distribute the values to the lower count or unseen  $N$ -grams. This will make sure that the sum of  $N$ -gram probabilities equal to 1. A different smoothing approach called the Jelinek-Mercer smoothing [Jelinek and Mercer, 1980], applies linear interpolation between  $N$ th order

maximum likelihood model and  $N - 1$ th order maximum likelihood model. For example, a bigram interpolated model is expressed as:

$$P_{interpolate}(w_i|w_{i-1}) = \lambda P_{ML}(w_i|w_{i-1}) + (1 - \lambda)P_{ML}(w_i) \quad (2.22)$$

where  $\lambda$  is the interpolation weight. Other approach for obtaining smoothed models is the back-off approach [Katz, 1987] which is also a common method employed. The intuition of this approach is to look for a  $(N - 1)$ -gram if there is no  $N$ -gram for a specific word sequence. This means, if we do not have an example of a particular trigram  $w_{i-2}, w_{i-1}, w_i$  to compute  $P(w_i|w_{i-2}, w_{i-1})$ , then we can use the bigram probability  $P(w_i|w_{i-1})$ . Similarly, if we cannot compute  $P(w_i|w_{i-1})$ , then we can look to the unigram  $P(w_i)$ . The back-off algorithm for trigram can be computed using the following equation [Jelinek, 2001]:

$$P_{Katz}(w_i|w_{i-2}, w_{i-1}) = \begin{cases} P_{ML}(w_i|w_{i-2}, w_{i-1}) & \text{if } r > k \\ \alpha Q_T(w_i|w_{i-2}, w_{i-1}) & \text{if } k \geq r > 0 \\ \beta(w_{i-2}, w_{i-1})P_{Katz}(w_i|w_{i-1}) & \text{if } r = 0 \end{cases} \quad (2.23)$$

where  $\alpha$  and  $\beta$  are weights,  $r$  is the number of counts in the training data,  $k$  is a threshold and  $Q_T$  is a Good-Turing function. The probability  $P_{Katz}(w_i|w_{i-1})$  is defined similarly as:

$$P_{Katz}(w_i|w_{i-1}) = \begin{cases} P_{ML}(w_i|w_{i-1}) & \text{if } r > k \\ \alpha Q_T(w_i|w_{i-1}) & \text{if } k \geq r > 0 \\ \beta(w_i)P_{ML}(w_i) & \text{if } r = 0 \end{cases} \quad (2.24)$$

### 2.5.2 Recurrent Neural Network model

An advanced technique for language modelling is the Recurrent Neural Network (RNN) introduced by Elman [1990] and later the architecture was adopted by Mikolov et al. [2010] for building language model. We will not use this technique for training our language model as it is beyond the scope of this thesis. However, we believe that this method is worth being mentioned as it is the latest technique employed in language modelling. The RNNLM equations are described as:

$$\text{input: } x(t) = w(t) + s(t - 1) \quad (2.25)$$

$$\text{hidden: } s_j(t) = f\left(\sum_i x_i(t)u_{ji}\right) \quad (2.26)$$

$$\text{output: } y_k(t) = g\left(\sum_j s_j(t)v_{kj}\right) \quad (2.27)$$

where  $x(t)$  is the input to the network (input layer). The input is computed by concatenating  $s(t)$ , which is the state of the network (hidden/context layer) and  $w(t)$ , which is a vector for current word encoded using 1-of-N coding.  $f$  is a sigmoid function, defined as  $f(z) = (1 + \exp(-z))^{-1}$  and the output unit  $g(z)$  is a softmax function:

$$g(z_m) = \frac{\exp(z_m)}{\sum_k \exp(z_k)}. \quad (2.28)$$

To train the network, backpropagation through time algorithm [Werbos, 1990] is used where the weights of the network are updated based on the error that is propagated back in time for a certain number of steps. Hence, the network has a short term memory of previous context layer. The size of the input layer  $x$  corresponds to size of the vocabulary and the hidden layer  $s$  typically has between 30-500 hidden units.

Experiments by Mikolov and colleagues ([Mikolov et al., 2010], [Mikolov et al., 2011]) have shown that RNN language model performed better than smoothed  $N$ -gram language model by providing lower perplexity value and improving ASR performance in terms of word error rate. The advantage of this approach is that it is able to capture long sequences of preceding words. However, the training time is long and there are constraints on the number of context words used for training. Several methods have been recently proposed to address the disadvantages (see informative survey and proposed improvements in [Mulder et al., 2015]).

## 2.6 Evaluation

We have described several techniques in previous sections for building the main components of an ASR system. Here, we explain how the G2P system, language model and speech recognition task are evaluated.

### 2.6.1 G2P performance

One way to evaluate the accuracy of a G2P model is by comparing hypotheses generated by the system with a reference transcript to measure its phoneme error rate (PER). The PER of a system is computed based on the following equation:

$$PER = \frac{S + I + D}{N} \times 100 \quad (2.29)$$

where  $N$  is the total number of phonemes in the reference transcript and  $S$ ,  $I$ ,  $D$  are the number of substitution errors, insertion errors and deletion errors, respectively.

### 2.6.2 Language model performance

Perplexity is the typical evaluation metric for language models which are used in ASR system. The measure helps to know how good is a language model in predicting the next word. A language model can be considered good if the perplexity is low when the model is tested on a held-out dataset (not the same set for training the language model). The perplexity of a language model, for example, is first determined by approximating the entropy (measuring the uncertainty) of a test data which has  $n$  number of sentences,  $s_1, s_2, s_3, \dots, s_n$  using the following equations:

$$H = -\frac{1}{N} \sum_{i=1}^n \log_2 P(s_i) \quad (2.30)$$

where  $N$  is the total number of words in the test corpus. Then perplexity is defined as  $2^H$ . Although lower perplexity language models are the preferred ones for speech recognition task, perplexity and ASR word accuracy do not necessarily correlate well.

Measuring out-of-vocabulary (OOV) rate on test data using a language model is also one way to evaluate the model. The OOV rate provides the percent of words in a test set which are not covered by the vocabulary in the language model.

### 2.6.3 ASR Performance

To evaluate the performance of the ASR system on a certain speech task, the most popular metric is the word error rate or word accuracy. For obtaining word error rate

(WER) of an ASR system, the method is similar to the one mentioned in Section 2.6.1 for measuring the performance of G2P. Instead of phoneme level, ASR is often evaluated on word level. WER is calculated using the following equation:

$$WER = \frac{S + I + D}{N} \times 100 \quad (2.31)$$

where  $S$ ,  $I$  and  $D$  are respectively the number of words wrongly substituted, inserted and deleted; while  $N$  is the number of words in a reference transcript. There are other similar WER measurements such as character error rate (CER) or syllable error rate (SER) which are mostly used for languages with segmentation issues.

## 2.7 Speech Recognition Toolkits

To date, there are several open source speech recognition toolkits available for research and application development. The following briefly describes the ones that we have tested ourselves including the toolkit which we will use for implementing our proposed approaches.

### 2.7.1 Sphinx

Sphinx<sup>5</sup> is one of the earliest open source speech recognizers distributed and the first version was developed by Kai Fu Lee for his Ph.D. thesis in 1988 and later the recognizer was made available for public in 2000 through the CMUSphinx website. The latest version now is Sphinx4 written in Java. Besides the recognizer, the speech group at the Carnegie Mellon University published Sphinxtrain, a tool for building acoustic models based on HMM and GMM. The tutorial for building a speech application using this tool is quite straightforward with many guides and active forums on the Internet. The most publicized tool by Sphinx is the lightweight version called PocketSphinx, which is appealing for building speech applications on embedded systems. The drawback of Sphinx is that the tools are targeted for (low-cost) practical development and not for research thus there is no support for exploring new techniques.

---

<sup>5</sup><http://cmusphinx.sourceforge.net>.

### 2.7.2 RASR

Another open source speech recognition toolkit is RASR<sup>6</sup> (short for RWTH ASR) written in C++ by the Human Language Technology and Pattern Recognition Group at the RWTH Aachen University. The system has been developed since 2001 and details on the toolkit were published later in 2009 [Rybach et al., 2009]. The toolkit offers latest techniques for acoustic modelling, such as discriminative training using minimum phone error criterion [Sixtus and Ney, 2002] and using deep neural networks. It also supports speaker adaptation using feature space maximum likelihood linear regression [Gales, 1998] and speaker normalization through vocal tract length normalization ([Welling et al., 1999], [Pitz and Ney, 2003]) and unsupervised training through confidence scores on state level (weighted states) [Gollan and Bacchiani, 2008]. Furthermore, the toolkit has been used for building large vocabulary systems for several research projects such as TC-STAR (European English and Spanish) [Lööf et al., 2007] and GALE (Arabic, Mandarin) ([Rybach et al., 2007], [Plahl et al., 2008]). One of the key aspects of RASR in developing systems with large vocabulary is that it supports grid computing that enables parallel acoustic model training.

At the start of this thesis, we explored this toolkit for implementing several experiments to investigate Malay ASR with speakers from different region [Juan et al., 2012]<sup>7</sup>. During the course of building the system, we encountered difficulties in understanding the configurations available and the lack of information in the wiki to explain each available features did not ease our problems. Although it is a promising tool for building state-of-the-art systems and could be beneficial for many types of research, new users may be discouraged with the limited documentation about the toolkit.

### 2.7.3 Kaldi

The Kaldi<sup>8</sup> [Povey et al., 2011b] speech recognition toolkit is built for the purpose of speech recognition research by Daniel Povey and his colleagues. Its stable release has just been published in 2013. It offers a system based on weighted finite-state-transducers (WFSTs) using OpenFST and written in C++. Furthermore, the toolkit offers latest

---

<sup>6</sup><http://www-i6.informatik.rwth-aachen.de/rwth-asr/>

<sup>7</sup>the work is not reported in this thesis, but we have included an abstract of the paper in Appendix A

<sup>8</sup><http://kaldi.sourceforge.net/>

techniques for acoustic modelling which include using subspace Gaussian mixture models and deep neural networks. The toolkit is a close competitor with RASR and HTK<sup>9</sup>, each aiming to provide state-of-the-art approaches particularly for acoustic modelling. Kaldi also provides numerous recipes for trying out its features and the contributors update the recipes from time to time. Most importantly, Kaldi describes its features thoroughly on its website and has an active forum that discusses issues pertaining Kaldi implementation.

For the most part of our work we use this toolkit due to various techniques offered, mainly, using SGMM and DNN for obtaining our acoustic models. Moreover, the codes can be fairly understood and all modules used are explained in the website which enables us to modify them for testing our proposed approaches.

## 2.8 Summary

We have introduced the basic concepts of ASR and presented the current techniques employed for developing acoustic model, language model or pronunciation model. Most of the techniques mentioned in this chapter will be used in our work, particularly the ones for implementing cross-lingual approach in acoustic modelling and bootstrapping our target language G2P system. In the next chapter, we will review some of the recent works concerning ASR for under-resourced languages, which is the main subject of this thesis.

---

<sup>9</sup><http://htk.eng.cam.ac.uk>.





## Chapter 3

# Challenges and recent approaches for ASR in under-resourced languages

### 3.1 Introduction

Speech applications have assisted the human-computer interaction for many tasks, e.g. voice command, speaker identification and speech translation systems. Today, these applications are within reach and can be found in desktop computers or mobile devices. Voice command systems are used to assist human to perform physical activities especially for helping disabled persons, speaker identification systems are used for determining people's identity, and speech translation systems can help people to communicate in several languages. Many ASR systems nowadays can work on dominant languages such as English, French, Mandarin, Japanese, etc. As large amount of resources is available in these languages, robust statistical models can be trained for ASR. There are still many languages remaining unavailable in speech applications, particularly under-resourced languages. An "under-resourced language" in the context of Human Language Technology (HLT) refers to a language that has the following problems (but not limited to): insufficient resources in terms of transcribed speech data or pronunciation dictionaries, unstable writing system, low amount of vocabulary or lack of (or non-existing) electronic resources for speech and language processing, etc ([[Berment](#),

2004], [Scannell, 2007], [Maxwell and Hughes, 2006]). This chapter discusses issues related to ASR for languages in this category. In addition, we bring up a history of research initiatives and present latest research development in acoustic modelling for helping low-resource speech recognition systems.

## 3.2 The challenges

As pointed out by Besacier et al. [2014] in their survey on ASR for under-resourced languages, the main issue of building systems for these languages is the lack of resources. Most HLT systems require large amount of data (speech, text, etc) in the target language to train statistical models that are robust for speech recognition tasks. However, it is not easy to acquire large amount of data in these languages, especially for rare or endangered languages. Here, we discuss five challenges as raised by the aforementioned authors that are related to under-resourced languages. In the following section, we present some of the initiatives by researchers to address them.

### 3.2.1 Engaging the community

Walsh [2005] mentioned that involving the natives in language maintenance or revitalization work is necessary. As working on ASR for under-resourced languages can be directly linked to one of language preservation efforts, the natives could take part in building the corpora for ASR. Of course, one cannot expect to have native speakers to have adequate technical skills. Hence, bridging the gap between language experts (native speakers) and technology experts must be done. To achieve this goal is not an easy task. Translators may be required to help with the communication and technology experts may need to develop user friendly tools for the native speakers to use. Besides that, when dealing with undocumented or rare languages, linguists would be required for defining the linguistic properties. Therefore, ASR projects must involve researchers who are from various disciplines.

### 3.2.2 Social and cultural influence

Code-switching is a phenomenon that appears around the world especially in an environment where people are exposed to more than one language. A person may switch between languages in his conversation. Malaysia for example, is estimated to have 138 living languages. The local dominant languages are Malay and English, as both are the official and second language used, respectively. Children learn both languages in schools from seven years old. At home, family members may speak in their mother tongue, which is not necessarily in Malay. Like the people in Borneo, there are many languages spoken such as Iban, Kadazan, Melanau, Bidayuh, etc. When all languages (official, non-official, mother tongue) are used often for communication as they are all equally important in the society, *code-switching* can occur in daily conversations.

For example, Figure 3.1 shows a Twitter post shared by a Malaysian. The sentence, “Wei macha, you want to makan here or tapau?” means, “Hey man, (do) you want to eat here or take away?”. In this sentence, there are words from Chinese (*wei*, *tapau*), Indian (*macha*), Malay (*makan*) and English languages. A sentence such as this could cause limitations to the capability of a monolingual speech recognizer because it cannot identify foreign words that appear in between utterances or phrases.



FIGURE 3.1: An example of how Malaysians use more than one language in conversation. This post went viral in social media in 2013.

### 3.2.3 Ethics

Ethical and technical rules must be applied when dealing with some languages, particularly for endangered or rare languages. This is to avoid misinterpretation of data, invading privacy issues or not respecting intellectual property rights e.g, indigenous

knowledge<sup>1</sup>. One may need to decide which type of data to be used for building resources for the system and how it will be published (public or private).

Adda et al. [2014] have recently investigated ethical and economical issues related to crowdsourcing for speech. Crowdsourcing is a method for collecting data through Internet, either by downloading contents from websites or by proposing tasks to Internet users to complete. One system which is commonly used for crowdsourcing is the Amazon Mechanical Turk<sup>2</sup>. It is a platform for employers to find employees to perform Human Intelligence Tasks (HITs). The authors have identified several issues concerning Turkers (employees): they are paid with very low salary<sup>3</sup>, produce low quality of work and receive late payments from employers. Thus, rules and regulations must be employed to make sure a win-win situation for both parties.

### 3.2.4 Data collection procedures

Engineers need to design methodologies for collecting data that can ensure high quality and sustainable corpus. When dealing with rare or endangered languages, researchers may need to travel to rural areas for collecting speeches from native speakers. Heavy, expensive and internet dependent devices for recording speeches could cause problems in conducting data collection campaigns in such areas. Thus, researchers must be innovative in developing devices and applications that are suitable.

### 3.2.5 Low-resource models

Current state-of-the-art ASR systems are based on statistical modelling approaches, which require large amount of training data to train models. As data in under-resourced languages are limited, training models on these data could result very weak models for speech recognition. Performance of ASR system could degrade due to this constraint. Hence, researchers are challenged in developing strategies for improving low-resource models.

---

<sup>1</sup>Indigenous knowledge, also referred as traditional knowledge, is defined as knowledge that is commonly used by indigenous people that is passed down from generation to generation [Mugabe et al., 2001]

<sup>2</sup><https://www.mturk.com/mturk/welcome>

<sup>3</sup>see more analyses by Panos Ipeirotis regarding this issue at <http://www.behind-the-enemy-lines.com>

### 3.3 Brief background on research initiatives

In the early years of speech recognition research, speech recognizers could only understand digits or isolated words in single language. The first speech recognition system ever built was called Audrey, which stands for Automatic Digit Recognizer, made in 1952 by researchers of Bell Labs. Fast forward thirty years, speech recognition technology has greatly improved. Nowadays, ASR systems could understand continuous speech and systems could be built for any language. In the 90s, several systems were capable in handling multiple language data e.g. systems developed in research labs such as MIT [Glass et al., 1995], Philips [Dugast et al., 1995], LIMSI [Lamel et al., 1995], Dragon [Barnett et al., 1996], IBM [Cohen et al., 1997], BBN [Billa et al., 1997] and Cambridge [Young et al., 1997]. Dominant languages besides English, such as German, French, Japanese and Mandarin Chinese have made their way into ASR research, thanks to the availability of vast amount of data.

Consequently, research on language independent speech recognizers has become popular as researchers are now interested to find ways to use resources from various languages to build a generic system which could deal any language. Schultz and colleagues for example, embarked many works related to this research topic. They have shown methods to bootstrap ASR systems for new languages by using multilingual acoustic models that were trained on multiple language data and shared phoneme set (see [Schultz and Waibel, 1997], [Schultz and Waibel, 1998], [Schultz and Waibel, 2001]). Cross-lingual strategies have emerged for improving performance of systems for under-resourced languages such as, Tamil [Çetin et al., 2007], Afrikaans [Imseng et al., 2014], Vietnamese [Le and Besacier, 2009], Polish [Löf et al., 2009], Basque [Barroso et al., 2011], Indonesia [Ferdiansyah and Purwarianti, 2012] and Mo Piu ([Caelen-Haumont and Sam, 2008], [Caelen-Haumont et al., 2011], [Caelen-Haumont, 2012]), an endangered language in Vietnam. The idea behind cross-lingual approaches is that parameters from out-of-language data are shared by transferring them to systems with target language data. Recently, development of Subspace Gaussian Mixture Models (SGMM) [Povey et al., 2010] and Deep Neural Networks (DNN) ([Hinton et al., 2012a], [Dahl et al., 2012]) for ASR has given new ideas to conduct cross-lingual strategies. Some of the model parameters in SGMM and DNN can be ported across systems and

no universal phoneme sets required. We will review several works related to these two techniques in the following section (see Section 3.4).

Code-switching problems in ASR have led to several solutions, for example, incorporating language identification (LID) into a multilingual system by combining acoustic model score and language information from Mandarin-English training data [Weiner et al., 2012]. Another similar approach has been made for Northern Sotho-English data [Mabokela et al., 2014] where they used phonotactic information for LID. Besides that, there are other approaches such as mapping phonetic representation of the foreign words to target graphemes ([Molapo et al., 2014] - for Shona and English) for improving systems that have multilingual content.

An interest in building a large multilingual corpus was also realized to support independent speech recognizers such as the GlobalPhone corpus ([Schultz, 2002], [Schultz et al., 2013]). As of today, the corpus has up to 20 languages that covers languages from Asia, Middle East, Africa, Europe and Americas. The availability of this corpus made many research ideas relating to multilingual and cross-lingual techniques successfully explored. Data collection solutions such as RLAT<sup>4</sup> appeared, for suggesting methods to achieve high quality corpus and open source mobile applications ([Hughes et al., 2010], [de Vries, 2011], [Bird et al., 2014b], [Bird et al., 2014a]) for more effective data acquisition process that can benefit research that deals with native speakers who live in far places. Open source tools for ASR development such as Sphinx<sup>5</sup>, Julius<sup>6</sup>, Kaldi<sup>7</sup>, HTK<sup>8</sup>, RASR<sup>9</sup> and YAST<sup>10</sup> are available for enthusiasts to create ASR systems for any language using state-of-the-art techniques.

Guidelines for good practices to crowdsourcing have been suggested by Adda et al. [2014] to ensure that data are ethically produced by Turkers and fair payment for the Turkers. Besides that, academic publishing platforms for speech community to share knowledge focusing on under-resourced languages have been established since less than a decade ago. Among them are: Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU); workshops in Language Resources

<sup>4</sup>Rapid Language Adaptation Toolkit, <http://csl.ira.uka.de/rLAT-dev>

<sup>5</sup><http://cmusphinx.sourceforge.net>.

<sup>6</sup>[http://julius.sourceforge.jp/en\\_index.php](http://julius.sourceforge.jp/en_index.php).

<sup>7</sup><http://kaldi.sourceforge.net/>

<sup>8</sup><http://htk.eng.cam.ac.uk>.

<sup>9</sup><http://www-i6.informatik.rwth-aachen.de/rwth-asr/>

<sup>10</sup><http://pi.imag.fr/xwiki/bin/download/PUBLICATIONS/WebHome/YAST.zip>

and Evaluation Conference (LREC) or Computational Linguistic (COLING) conference such as, Workshop on Indian Language Data: Resources and Evaluation; Workshop on Language Resources and Technologies for Turkic Languages; Workshop on Parsing in Indian Languages; Workshop on South and Southeast Asian Natural Language Processing; conferences such as International Conference on Asian Languages (IALP) and Pacific Asia Conference on Language, Information and Computing (PACLIC).

### 3.4 Most recent techniques in cross-lingual acoustic modelling

In this section, we will discuss recent contributions on low-resourced ASR focusing on sharing acoustic parameters across language data. Two techniques are involved in this review, the Subspace Gaussian Mixture Model and Deep Neural Networks for ASR. Technical details on the two have been presented in Chapter 2 of this thesis.

#### 3.4.1 Subspace Gaussian Mixture Model

##### 3.4.1.1 Cross-lingual SGMM for low-resource ASR

Lu et al. [2014] demonstrated their work in porting SGMM shared parameters that are trained with multilingual data to systems that have limited training data. Previous cross-lingual techniques requires universal phone units (e.g. [Schultz and Waibel, 1997], [Schultz and Waibel, 1998], [Lin et al., 2009], [Wang et al., 2002]) for dealing with various spoken language data. In SGMM on the other hand, shared parameters can be transferred across SGMM systems and do not depend on the Hidden Markov Model (HMM) states.

In Lu et al. [2014]’s work, the authors conducted a study on cross-lingual SGMMs using the GlobalPhone corpus [Schultz, 2002]. The ASR experiments involved German (GE) as the target language, Spanish (SP), Portuguese (PT) and Swedish languages as source languages. Each of the data sets contain 15 to 23 hours of speech. Three GE data with different amount of speech were used, 1 hour, 5 hours and 15 hours of data. Using the Kaldi speech recognition toolkit [Povey et al., 2011b], monolingual GE systems



were built based on GMM and SGMM acoustic modelling, utilizing the three training data. The baseline systems showed that monolingual SGMM gives better performance in terms of word error rate (WER) compared to GMM and the WER reduced significantly as larger training data sets were used.

For preparing cross-lingual SGMM, UBMs for the source languages were obtained either in monolingual or multilingual approach. Then, the shared parameters were transferred to the GE system for SGMM modelling. This cross-lingual approach resulted significantly better WER results compared to the results obtained from the monolingual system. In particular, the GE system with only 1 hour training data has up to 17% relative improvement when UBM from the source languages were used to initialize the SGMM.

They also investigated the effects of cross-lingual SGMM with regularization, maximum a posteriori (MAP) adaptation and speaker subspace. Regularization of maximum likelihood (ML) estimation of SGMMs was used for reducing overfitting when the training data is small [Lu et al., 2011]. MAP was applied in order to reduce the possible mismatch between the shared parameters from the source language data and the target language system (phones, recording conditions). The method involved an adaption of the phonetic subspace  $\mathbf{M}_i$  using MAP, which has been experimented in their previous work [Lu et al., 2012]. The final method considered speaker adaptive training using speaker subspace. Small target language data like 1h (8 speakers) was not sufficient to train the speaker subspace  $\mathbf{N}_i$  on a speaker basis. Thus, Lu et al. estimated a multilingual  $\mathbf{N}_i$  by merging the subspace across the source language system and then, obtained the target language system on the acquired speaker subspace. Table 3.1 shows the ASR results from the experiments for German language.

#### **3.4.1.2 Using out-of-language data for improving ASR for under-resourced language: a case study of Afrikaans**

Imseng et al. [2014] developed an Afrikaans ASR by employing multilingual data to build posterior features and SGMMs. The authors conducted experiments using Dutch data (80 hours) for improving the monolingual system. Dutch was chosen as source language as it has been found to be suitable for developing Afrikaans ASR [Heeringa and de Wet, 2008]. Their initial experiment showed that the HMM/GMM system

System	1 hour		5 hour		15 hour	
	Dev	Eval	Dev	Eval	Dev	Eval
GMM baseline	23.2	34.1	18.5	28.0	15.4	24.8
SGMM baseline	20.4	31.4	14.9	24.9	13.0	22.1
+ speaker subspace	-	-	14.6	24.7	12.4	21.5
Cross-lingual						
w/SP, $S = 20$	18.8	32.4	15.4	26.5	-	-
w/PO, $S = 20$	17.9	30.9	14.6	25.2	-	-
w/SW, $S = 20$	18.0	31.0	14.6	25.4	-	-
w/Mul + $\ell$ , $S = 40$	15.5	26.9	12.7	22.1	12.0	21.6
+ speaker subspace	<b>15.3</b>	<b>26.7</b>	-	-	-	-

TABLE 3.1: A summary of results by Lu et al. [2014] for cross-lingual SGMM experiments using German as the target language.

for Afrikaans ASR is outperformed by SGMM with 4.3% phoneme accuracy (PA) improvement when using three hours of training data. Moreover, SGMM system also surpass the results of posterior feature based systems like Kullback-Leibler divergence based HMM (KL-HMM) and Tandem systems, which are also better than HMM/GMM.

To demonstrate cross-lingual effect, the Dutch data was employed for training the multilayer perceptron (MLP) for the feature based models and the globally shared parameters for SGMM. Then, the target language data was used to train the state specific parameters, HMM distributions and MLLR adaptation. This time, the setup for Afrikaans data was based on three level of data sparsity: 6 mins, 1 h and 3 h. Table 3.2 presents some of the ASR results that were obtained in the study. The SGMM experiment result yield 7.3% PA improvement from the monolingual HMM-GMM system for cross-lingual data for SGMM, given 3 h target language data. Furthermore, the result outperforms monolingual SGMM by 3% PA improvement. Using Tandem and KL-HMM acoustic models that were trained on Dutch data also improve HMM-GMM system but both results are lower than SGMM result. By comparing results from the three systems based on different size or target language data, the cross-lingual SGMM gradually improves when there are more data for training the state specific parameters. However, KL-HMM system works better than SGMM for 6 mins of target data. As stated in the paper, the SGMM performance could be influenced by the dimensionality of the substates used in the model.

A multilingual version of MLP and SGMM were obtained to further improve the current cross-lingual ASR results. Imseng et al. performed merging method to combine the output of MLPs from Dutch and Afrikaans and trained the shared parameters for

SGMM in multilingual fashion. For SGMM, the shared parameters were obtained by training the data of both languages (3h Afrikkans and 80h Dutch). Then, the state specific parameters were estimated on the 3h target language data. By employing this approach, slight improvements can be seen for all systems when compared to the first cross-lingual experiment results. For the SGMM experiment, using target and source language data to build the shared parameters resulted a better SGMM model where the system outperforms the monolingual SGMM and HMM/GMM systems.

System	Monolingual	Cross-lingual (Dutch)			Multilingual
	3 h (PA %)	6 mins (PA %)	1 h (PA %)	3 h (PA %)	3h (PA %)
HMM/GMM	61.2	38.6	55.3	61.2	-
KL-HMM	60.6	53.1	<b>61.5</b>	67.3	<b>68.8</b>
Tandem	64.7	41.0	61.3	68.2	68.4
SGMM	<b>65.5</b>	40.2	60.4	<b>68.5</b>	68.6

TABLE 3.2: Results (phoneme accuracy - PA) for HMM/GMM, KL-HMM, Tandem and SGMM extracted from [Imseng et al. \[2014\]](#) for cross-lingual ASR through exploiting out-of-language data to improve Afrikaans ASR

### 3.4.2 Deep Neural Networks

#### 3.4.2.1 Shared-hidden-layer Multilingual DNN for ASR

Shared-hidden-layer multilingual deep neural networks were proposed by [Huang et al. \[2013\]](#) for improving performance of monolingual ASR. The method involved training hidden layers of DNN on multilingual data simultaneously and then fine-tuning the hidden layers on new language data using back propagation algorithm in order to obtain the final (softmax) layer. The method was carried out in such a way as hidden layers with multilingual information could discriminate phonetic classes in other languages. The architecture of the shared hidden layer in DNN used in this work is depicted in Figure 3.2.

Method	ENU (WER%)			CHN (CER%)		
	3h	9h	36h	3h	9h	36h
Baseline	38.9	30.9	23.0	45.1	40.3	31.7
FRA HLs + Train Softmax layer	-	27.3	-	-	-	-
SHL-MDNN + Train Softmax layers	28.0	25.3	22.4	35.6	33.9	26.6

TABLE 3.3: DNN results obtained by [Huang et al. \[2013\]](#) in their experiment for cross-language knowledge transfer using shared-hidden-layers (SHL-MDNN) for ASR with limited conditions. Target languages : American English (ENU) and Mandarin Chinese (CHN)

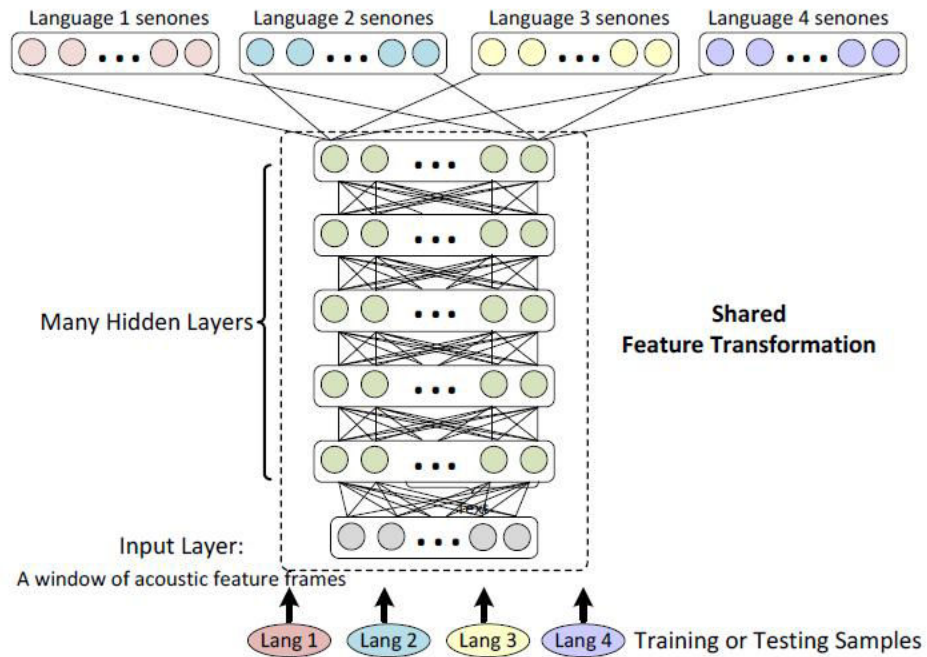


FIGURE 3.2: Architecture of the shared-hidden-layer multilingual DNN. Reprinted from Huang et al. [2013].

The authors conducted the study for Microsoft internal speech recognition task, using four data from European languages, French (FRA-138 h), German (DEU-195 h), Spanish (ESP-63 h) and Italian (ITA-63 h). All four datasets were used to create 5 hidden layers with 2048 units in each layer. The shared-hidden-layers or SHL, were then finetuned to American English (ENU) and Mandarin (CHN) data ( ENU, CHN: 3, 9 and 36 h). Based on the baseline (monolingual) results, the DNN improved as more training data is available. More importantly, model transfer using multilingual pretrained model succeeded in improving the monolingual results. In fact, adding several language data as input for training the hidden layers proved to be more effective than using a single language data for acoustic model transfer (in the case of using FRA DNN for ENU). After fine-tuning multilingual hidden layers on ENU data, ASR performance improved with 5.6% WER reduction from the ENU baseline. Interestingly, the Mandarin ASR, where the language has no relation with European languages, also gained recognition accuracy after cross-lingual approach was applied. The Mandatin systems achieved 8.3 to 21.1% relative character error rate reduction from the baseline results. Table 3.3 presents DNN baseline and improved results obtained in the experiments.

### 3.4.2.2 Cross-lingual and Multilingual DNN experiments using GlobalPhone corpus

Recently, several studies on cross-lingual and multilingual DNN have been conducted using GlobalPhone [Schultz et al., 2013] corpus (e.g. [Ghoshal et al., 2013], [Swietojanski et al., 2013], [Miao and Metze, 2013], [Vu et al., 2014]). The GlobalPhone is a multilingual data developed by Karlsruhe Institute of Technology. It has more than 400 hours of read speech in 20 languages, spoken by more than 2,000 native speakers. Each language has 10 to 30 hours of data for training and test (development and evaluation sets). The package includes the audio data, speech transcripts, pronunciation dictionaries and language models. Three studies done by Swietojanski et al. [2013], Miao and Metze [2013] and Vu et al. [2014] exploited the corpus for building systems with limited conditions and applying cross-lingual approaches for improving their baseline results.

Swietojanski et al. [2013] explored cross-lingual DNN using four language data in the GlobalPhone corpus for building a German system with limited amount of transcribed data. Two approaches were used for the study, DNN based on tandem and hybrid settings. For initializing the DNNs, they employed an unsupervised pretraining using layer-wise restricted Boltzmann machine (RBM) [Hinton et al., 2006]. The pretraining was done by training the RBMs on untranscribed multilingual data. Subsequently, the RBMs or Deep Belief Network (DBN) were refined using training data in target language.

The setup of their experiments were as follows: the German (GE) ASRs had 1h, 5h and 15h of transcribed training data, and the untranscribed data for pretraining contain speeches in German (GE-15h), Portuguese (PO-26h), Spanish (SP-22h), Swedish (SW-22h) or all four languages (85h) mentioned. As shown in Table 3.4, the WER results for DNN with pretraining outperformed HMM/GMM and DNN with random initialization. The results also showed performance of tandem systems. By comparing performance of tandem and hybrid systems, the latter provided lower WER particularly when training data is very limited. However, from the DNN results it is hard to distinguish which language data was better for pretraining. The WERs are almost consistent for different pretraining data used. Furthermore, adding more data for pretraining (85h), showed very little effect to the DNN results. Nevertheless, it is worth

to note from their findings that pretraining is language independent and the outputs of pretraining can be used for initializing DNN for other languages.

System	Tandem			Hybrid		
	15h	5h	1h	15h	5h	1h
HMM/GMM	24.53	27.56	34.08	24.13	27.08	33.11
DNN random initialized	22.05	25.10	31.84	21.52	25.03	33.54
DNN pretrained on GE	21.39	24.60	30.91	20.09	22.78	28.70
DNN pretrained on PO	<b>21.21</b>	24.43	31.29	<b>20.00</b>	<b>22.44</b>	28.79
DNN pretrained on SP	21.48	<b>24.23</b>	30.74	20.03	22.64	<b>28.4</b>
DNN pretrained on SW	21.62	24.44	<b>30.52</b>	20.20	22.89	28.92
DNN pretrained on ALL	21.48	24.49	30.98	20.14	22.70	28.72

TABLE 3.4: Tandem and Hybrid WER results on German [Swietojanski et al., 2013]

Another study on cross-lingual / multilingual DNN was conducted by Miao and Metze [2013] who also used German as the target language. Their first goal of the study was to improve DNN finetuning by utilizing "dropout", which could help to prevent overfitting when using back propagation algorithm and the second goal was to use multilingual DNN to boost performance of low-resource systems. They used a dropout strategy as proposed by Hinton et al. [2012b] involving hidden dropout factor (HDF) and input dropout factor (IDF), defined in the DNN training. The dropout discards each hidden unit based on the dropout factors, averaging multiple networks which can help to enhance model observations. For training multilingual DNN, their method was similar to the strategy used by Huang et al. [2013] where the models were trained simultaneously on several language data.

Method	2h	5h
DNN (random initialized)	29.0	24.6
DNN	27.8	24.1
dropout_DNN	26.3	23.1
DNN_SP	26.2	23.6
DNN_SP+PO	25.8	23.4
dropout_DNN_SP+PO	24.6	22.5

TABLE 3.5: DNN results based on WER for German ASR with dropout and multilingual effects [Miao and Metze, 2013]

The dropout and multilingual strategies were deployed to build two German ASRs with limited training conditions; 2h and 5h labeled data. Standard HMM/GMM systems were obtained to get the HMM structures. Then, DNN systems were built using target language data for acquiring baseline results. Subsequently, the dropout strategy was applied to the monolingual system and results showed improvement of at least 1% WER

from the baseline (see Table 3.5). For the multilingual strategy, Spanish (SP-5h) and Portuguese (PO-5h) data were employed for creating hidden layers of DNN, which were then used for fine-tuning on German data. The strategy succeeded in improving the monolingual DNN results. In fact, using Spanish data in the DNN reduced the WERs by 1.2% for 2h system and 0.5% for 5h system and by adding Portuguese data, the multilingual DNN outperformed the baseline by 2% and 0.7% for 2h and 5h systems, respectively. They also applied the dropout strategy in the multilingual approach which resulted further improvement to the multilingual DNN results.

The third study on multilingual DNN using the same corpus was carried out by Vu et al. [2014]. They investigated the effects of concatenating phones for multilingual DNN training and using multilingual DNN based on Kullback-Leibler (KL) divergence method. The first study employed two strategies. The first strategy was to use a universal phone set created by concatenating all phone sets (including language ID) from the languages involved. The second strategy used a set created by concatenating all phones sets and phones that shared the same IPA symbol. Then, the DNN was fully trained on multilingual data using the shared phone set. The last layer of the DNN was removed and then the hidden layers of the DNN were ported to a new language for fine-tuning. The second study was on using KL-HMM models trained on out-of-language data in low-resource systems. In this investigation, Vu et al. exploited ten language datasets from the GlobalPhone corpus: Bulgarian (BG), Czech (CZ), French (FR), Japanese (JP), German (GE), Hausa (HA), Mandarin (MAN), Portuguese (PO), Spanish (SP), and Vietnamese (VN) and English data set from the Wall Street Journal corpus (WSJ10).

Method	PO-17h	PO-5h	PO-1h	CZ-1h	HA-1h	VN-1h
DNN	13.2	15.2	20.5	16.9	16.1	32.1
DNN-MUL*	12.9	13.9	17.8	14.0	13.6	27.1
DNN-MUL*+KL	<b>12.6</b>	13.8	17.7	<b>13.1</b>	<b>12.0</b>	<b>26.6</b>
DNN-MUL*-IPA	12.9	<b>13.7</b>	17.4	13.9	13.3	27.0
DNN-MUL*-IPA+KL	12.7	<b>13.7</b>	17.1	13.4	12.3	26.8

TABLE 3.6: Summary of WER (%) results by Vu et al. [2014] for multilingual DNN using modified phone sets and Kullback-Leibler divergence method. Note that MUL\* indicates that the multilingual datasets vary for Portuguese and Czech, Hausa, Vietnamese experiments.

Their first experiment involved datasets containing Indo-European languages (source languages: FR, GE and SP, target language: PO). Full (17h) and randomly selected (1h



and 5h) training sets for PO were used. The DNNs were trained using RBM pretraining, stochastic gradient descent and discriminative training. Based on their findings, the baseline results were outperformed by the multilingual DNNs. As shown in Table 3.6, the best results were obtained when using multilingual DNN based on KL method and using merged phone set. The gains are significant particularly for very limited training data in target language (1h). The second experiment used datasets from languages that come from different language families (source languages: BG, GE, SP, JP, MAN, target language : CZ, HA, VN). Similar approaches were used except that the multilingual DNNs were trained using greedy-wise-layer supervised training and 1h training data for each target language. The hidden layers of the DNN were later fine-tuned to the target language data. Results showed similar effect; WER reductions for Czech, Hausa and Vietnamese systems when multilingual approach was used. Although the source languages were not related to the target languages, the approach boosted the performances of the low-resource ASRs.

### 3.5 Summary

This chapter described on-going challenges that are faced by researchers when handling under-resourced languages for speech recognition. Not only they are dealing with technical difficulties, but also some social issues that can influence the procedure of data acquisition, which can deteriorate the quality of the corpus. In this chapter, we also summarized previous and most recent initiatives that had been done by scientists in the quest to build and improve systems in low-resourced conditions, particularly for new languages. This led to our interest of two most recently used techniques for dealing with under-resourced languages, the SGMM and DNN. From the studies which we have reviewed, both techniques are useful for conducting cross-lingual acoustic modelling. The UBM in SGMM and, DBN and hidden layers in DNN are language independent thus can be trained multilingually and transferred across systems. Furthermore, universal phone sets are not necessary for performing cross-lingual strategy using these two techniques. Also, reviewed studies on cross-lingual and multilingual SGMM / DNN for low-resource systems showed very promising results. Due to this, we are motivated to employ SGMM and DNN framework in our ASR experiments. The following chapter introduces Iban, an under-resourced language, as well as several languages involved in our study.





## Chapter 4

# Languages from Malaysia

### 4.1 Introduction

Malaysia is a multilingual country where it is common to hear conversations in several languages in one place. The official language is Malay and English is the second language in the country. However, languages such as Mandarin and Tamil are also widely spoken in the country as these two languages are taught in vernacular schools. Besides these local dominant languages, there are other languages spoken by the native people, such as Iban, Bidayuh, Dusun, Melanau, Kelabit, Penan and many more. Yet, not many of these languages have writing system and some of the languages are only spoken by minority groups in the country. Ethnologue [Lewis et al., 2014] reported that there are in fact 138 living languages spoken by 24 million people in the country. Of the living languages, 101 are classified as “In Trouble”, based on a measurement called the EGIDS or Expanded Graded Intergenerational Disruption Scale [Lewis and Simons, 2010]. This tool investigates the level of development or endangerment that occurs in languages. The “In Trouble” class defines a language that is going through a broken intergenerational transmission. The child-bearing generation (or parents) is still able to use the language and it is suggested that parents could help to revitalize the language at home. Also in the report, 11 languages are institutional, 6 are developing, 5 are vigorous and 15 are dying. Figure 4.1 shows a map of Malaysia with language distribution according to language families. It shows that languages from the Malayic language group are spread in most part of Malaysia compared to other groups. In the Malayic group, there are two

languages of our concern and they are Malay and Iban. Hereafter, we will describe the two languages in detail.

## 4.2 Malay language

Malay is a member of the Austronesian language family, which is a family that has languages spread from the mountainous area in the northern part of Taiwan to the south of New Zealand, and from the west of Madagascar to the southeast of Easter Island [Omar, 1989]. There are four primary branches under the Austronesian family, Rukai, Tsou, Puyuma and Malayo-Polynesian [Bust, 1999]. Malay belongs to the Malayo-Polynesian branch. Figure 4.2 shows the full path in the linguistic family tree for Austronesian languages - focusing on Malay and Iban. The path is described from the highest level classification to the lowest, as provided by Ethnologue. There are several Malay languages spoken across Indonesia, Singapore, Brunei and Malaysia. Our study concerns with the one that is used in Malaysia, which is called, Standard Malay (for simplicity, the term “Malay” will be used in this thesis).

### 4.2.1 Phonology

Phonology is one of the important concepts to describe language properties. It studies the sounds of a language such as function, behaviour and organization of sounds as a linguistic item. It also studies the sound system and the general characteristics of the system. Phonemes are used to describe the speech sounds of a language. A phoneme is the smallest unit in the sound system. Typically, linguists use IPA symbols as shown in Appendix B to represent the phonemes and phonemes are placed between slashes (e.g /k/ for “k”) when written.

Malay has 27 consonants, six vowels and three diphthongs [Maris, 1979]. Among the 27 consonants, 8 of them are borrowed consonants, due to many words borrowed from English and Arabic languages.

#### 4.2.1.1 Vowel phonemes

There are two front vowels /i, e/. For example:

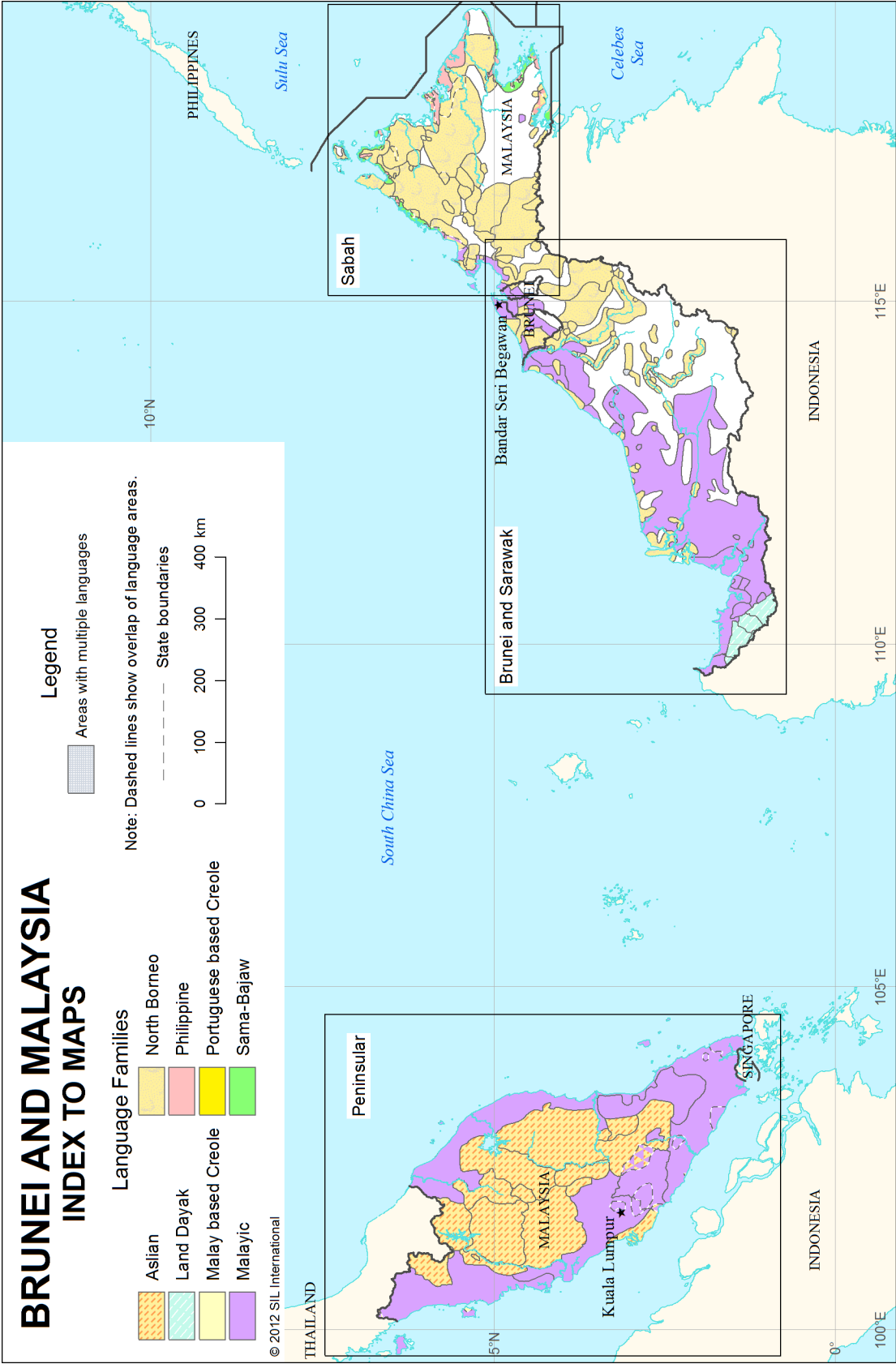


FIGURE 4.1: Language Map of Malaysia and Brunei with language distribution. Reprinted from Ethnologue with permission.

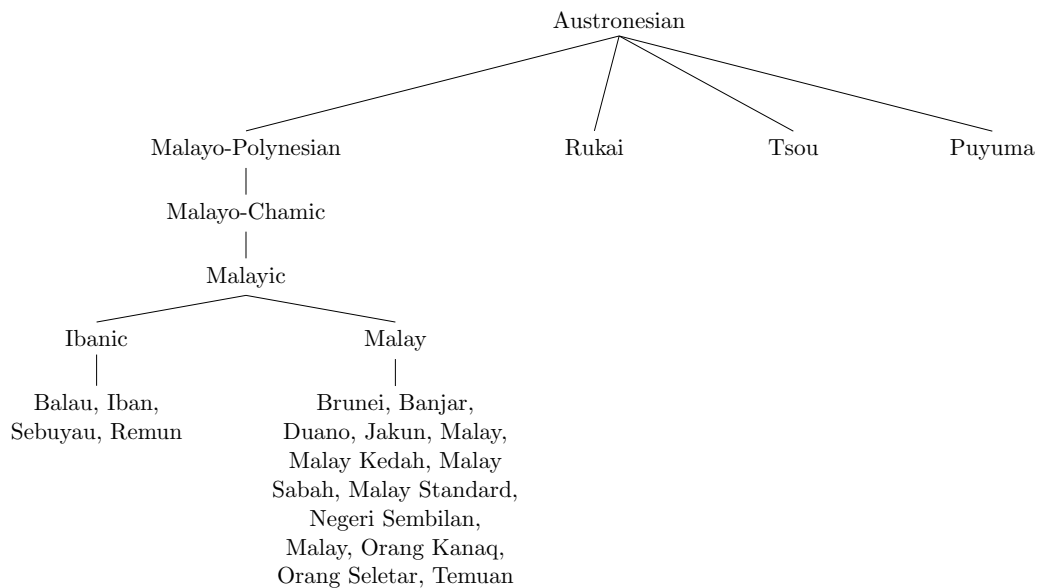


FIGURE 4.2: Family tree of the Austronesian languages - focusing on Malay and Iban

- /i/ : /ikan/ ‘*ikan*’ (fish), /itu/ ‘*itu*’ (that), /kini/ ‘*kini*’ (now)
- /e/ : /ekor/ ‘*ekor*’ (tail of an animal), /eja/ ‘*eja*’ (spell)

For the back vowels, there are /o/ and /u/. For example:

- /o/ : /otak/ ‘*otak*’ (brain), /orang/ ‘*orang*’ (person or people), /kota?/ ‘*kotak*’ (box)
- /u/ : /gula/ ‘*gula*’ (sugar), /umpan/ ‘*umpan*’ (bait), /lampu/ ‘*lampu*’ (light)

Finally, there are two central vowels /ə/ and /a/. For example:

- /ə/ : /əma?/ ‘*emak*’ (mother), /kəna/ ‘*kena*’ (hit), /kitə/ ‘*kita*’ (we)
- /a/ : /aku/ ‘*aku*’ (I/me/myself), /kita/ ‘*kita*’ (we), /tiba/ ‘*tiba*’ (arrive)

Some Malay words come with open final syllable, for example ‘*kita*’, is often pronounced as /kitə/ by Malaysians that originate from the west. Those who are from the east often pronounce the word as /kita/.

#### 4.2.1.2 Consonant phonemes

There are 19 original consonants in Malay and the consonants are categorized as plosive, nasal, affricate, fricative, trill, lateral and semi-vowel. The consonants are:

- Plosive: /p, b, t, d, k, g, ʔ/
- Affricate: /tʃ, dʒ/
- Nasal: /m, n, ɲ, ŋ/
- Fricative: /s, h/
- Trill: /r/
- Lateral : /l/
- Semi-vowel : /w, j/

Table 4.1 presents the consonant phonemes and examples of using the sounds in Malay. All the consonants can be pronounced at the beginning, middle or final position of a word. The glottal stop, /ʔ/ is mostly used in word-final position. It is often written as ‘k’ or not written explicitly, for example, /masaʔ/ is written as ‘*masak*’ (cook) and /nasiʔ/ is written as ‘*nasi*’ (rice). It can be also pronounced in the middle of a word, for example, in borrowed words from Arabic language e.g; /taʔat/ ‘*taat*’ (loyal).

Consonant	Examples
/p/	/lupa/ ‘ <i>lupa</i> ’ (forget), /pindʒam/, ‘ <i>pinjam</i> ’ (borrow), /atap/ ‘ <i>atap</i> ’ (roof)
/b/	/batu/ ‘ <i>batu</i> ’ (rock), /cuba/ ‘ <i>cuba</i> ’ (try), /ləmbab/ ‘ <i>lembab</i> ’ (weak)
/t/	/təpi/ ‘ <i>tepi</i> ’ (side), /tətapi/ ‘ <i>tetapi</i> ’ (however), /dapat/ ‘ <i>dapat</i> ’ (get)
/d/	/dəpan/ ‘ <i>depan</i> ’ (front), /sudah/ ‘ <i>sudah</i> ’ (already), /had/ ‘ <i>had</i> ’ (limit)
/k/	/kaja/ ‘ <i>kaya</i> ’ (rich), /cuka/ ‘ <i>ʃuka</i> ’ (vinegar), /tembaʔ/ ‘ <i>tembak</i> ’ (shoot)
/g/	/guna/ ‘ <i>guna</i> ’ (use), /lagu/ ‘ <i>lagu</i> ’ (song), /beg/ ‘ <i>beg</i> ’ (bag)
/ʔ/	/masaʔ/ ‘ <i>masak</i> ’ (cook), /nasiʔ/ ‘ <i>nasi</i> ’ (rice)
/m/	/makan/ ‘ <i>makan</i> ’ (eat), /lama/ ‘ <i>lama</i> ’ (old), /pitam/ ‘ <i>pitam</i> ’ (faint)
/n/	/nasiʔ/ ‘ <i>nasi</i> ’ (rice), /anaʔ/ ‘ <i>anak</i> ’ (child), /lapan/ ‘ <i>lapan</i> ’ (eight)
/ɲ/	/ɲapi/ ‘ <i>nyanyi</i> ’ (sing), /miɲaʔ/ ‘ <i>minyak</i> ’ (oil)
/ŋ/	/ŋaja/ ‘ <i>nganga</i> ’ (jaw drop), /seŋat/ ‘ <i>sengat</i> ’ (sting), /matan/ ‘ <i>matang</i> ’ (mature)
/s/	/saja/ ‘ <i>saya</i> ’ (I/me/myself), /pasar/ ‘ <i>pasar</i> ’ (market), /putus/ ‘ <i>putus</i> ’ (cut off)
/h/	/harap/ ‘ <i>harap</i> ’ (hope), /bahu/ ‘ <i>bahu</i> ’ (shoulder), /rumah/ ‘ <i>rumah</i> ’ (house)
/tʃ/	/tʃinta/ ‘ <i>cinta</i> ’ (love), /katʃa/ ‘ <i>kaca</i> ’ (glass), /matʃ/ ‘ <i>Mac</i> ’ (March)
/dʒ/	/dʒam/ ‘ <i>jam</i> ’ (hour/watch), /badʒa/ ‘ <i>baja</i> ’ (baja), /garadʒ/ ‘ <i>garaj</i> ’ (garage)
/r/	/rehat/ ‘ <i>rehat</i> ’ (rest), /bərat/ ‘ <i>berat</i> ’ (heavy), /bənar/ ‘ <i>benar</i> ’ (true)
/l/	/lari/ ‘ <i>lari</i> ’ (run), /malas/ ‘ <i>malas</i> ’ (lazy), /bantal/ ‘ <i>bantal</i> ’ (pillow)
/w/	/wajan/ ‘ <i>wayang</i> ’ (cinema), /bawa/ ‘ <i>bawa</i> ’ (bring), /kalaw/ ‘ <i>kalau</i> ’ (if)
/j/	/ja/ ‘ <i>ya</i> ’ (yes), /sajan/ ‘ <i>sayang</i> ’ (darling), /pakaj/ ‘ <i>pakai</i> ’ (take)

TABLE 4.1: Malay consonant phonemes

The remaining eight Malay consonants are borrowed from Arabic and English languages. This is due to the historical influence from the Arab merchants who travelled

to Southeast Asia in the early 15th century and then the British during the colonization era in 1940s. All of the consonants are fricatives and occur in borrowed words. The consonants are:

- /f/: /fikir/ ‘*fikir*’ (think), /wafat/ ‘*wafat*’ (death)
- /v/: /novel/ ‘*novel*’ (novel), /ven/ ‘*van*’ (van)
- /θ/: /hadiθ/ ‘*hadith*’ (hadith)
- /ð/: /ðarab/ ‘*darab*’ (multiply), /ðadi/ ‘*kadi*’ (qadi)
- /z/: /zaman/ ‘*zaman*’ (era) /zu/ ‘*zoo*’ (zoo)
- /χ/: /χabar/ ‘*khabar*’ (news),
- /ʃ/: /ʃarat/ ‘*syarat*’ (rule), /ʃabas/ ‘*syabas*’ (excellent)
- /ʁ/: /ʁaib/ ‘*ghaib*’ (invisible), /loʁat/ ‘*loghat*’ (accent)

#### 4.2.1.3 Diphtongs

Malay diphtongs are /aj/, /aw/ and /oj/ and they occur in open syllables for example /aj/ : /pandaj/ ‘*pandai*’ (clever), /aw/ : /kərbaw/ ‘*kerbau*’ (buffalo) and /oj/: /amboj/ ‘*amboi*’ (an expression equivalent to wow).

#### 4.2.2 Writing system

Orthography defines the set of symbols to be used in writing to represent spoken words. It helps to build a writing system for a language by setting up rules of spelling and using hyphenation or punctuation.

Malay is written using Latin alphabets and it is considered as an agglutinative language. An agglutinative language has words that are formed by adding affixes onto a root word and has word composition or reduplication. Furthermore, it is not a tonal language like Mandarin or Vietnamese languages. Also, speakers can distinguish general pronunciations directly from the written form. All of the phonemes presented previously are associated to the letters as shown in Table 4.2. Almost all of the phonemes can be

represented with a unique letter or pair of letters except for vowels /e/ and /o/. The letters ‘i’ and ‘u’ are /e/ and /o/ respectively, when in close positions, for example, ‘-ih’, ‘-ik’, ‘-uh’ or ‘-um’.

Phoneme	Capital letter	Small letter	Example
/a/	A	a	<i>akar</i> (root)
/b/	B	b	<i>batu</i> (rock)
/ɕ/	C	c	<i>cawan</i> (cup)
/d/	D	d	<i>dekat</i> (near)
/ə/	E	e	<i>emak</i> (mother)
/e/	E	e	<i>enak</i> (delicious)
/e/	-	i	/bileʔ/ <i>bilik</i> (room)
/f/	F	f	<i>Februari</i> (February)
/g/	G	g	<i>gajah</i> (elephant)
/h/	H	h	<i>haiwan</i> (animal)
/i/	I	i	<i>ikan</i> (fish)
/dʒ/	J	j	<i>jahat</i> (bad)
/k/	K	k	<i>kawan</i> (friend)
/l/	L	l	<i>lupa</i> (forget)
/m/	M	m	<i>makan</i> (food)
/n/	N	n	<i>nama</i> (name)
/ɲ/	Ny	ny	<i>banyak</i> (a lot)
/ŋ/	Ng	ng	<i>bunga</i> (flower)
/o/	O	o	<i>bola</i> (ball)
/o/	-	u	/ɲamok/ <i>nyamuk</i> (mosquito)
/p/	P	p	<i>jumpa</i> (meet)
/r/	R	r	<i>berat</i> (heavy)
/s/	S	s	<i>rasa</i> (taste)
/t/	T	t	<i>datang</i> (come)
/u/	U	u	<i>tumbang</i> (fall down)
/v/	V	v	<i>vot</i> (vote)
/w/	W	w	<i>wayang</i> (cinema)
/j/	Y	y	<i>kaya</i> (rich)
/z/	Z	z	<i>bazar</i> (bazaar)
/ð/	Z	z	<i>zalim</i> (evil)
/θ/	-	th	<i>mithali</i>
/ʃ/	Sy	sy	<i>syarat</i> (rule)
/χ/	Kh	kh	<i>khabar</i> (news)
/ʁ/	Gh	gh	<i>ghaib</i> (invisible)

TABLE 4.2: List of phonemes and alphabets used in Malay



### 4.3 Iban language

Sarawak is the largest state in Malaysia and has an estimated population of 2.3 million in 2013<sup>1</sup>. It is situated on the northwest of the Borneo Island and shares borders with the state of Sabah, Brunei and Kalimantan from Indonesia. There are many languages spoken in Sarawak such as Iban, Sarawak Malay (Malay dialect), Melanau, Bidayuh, Penan, Kelabit and others. The Iban language is mostly spoken by the Iban, an ethnic group with more than 690,000 people living in the state. Historically, during the British colonization, the rajah of Sarawak, Sir James Brooke, referred them as “Sea Dayak” due to their activities related to the sea. Similar to Malay, the language belongs to the Malayo-Polynesian branch of the Austronesian language family and Iban falls in the Ibanic language group [Lewis et al., 2014] (See Figure 4.2 for the linguistic path for Iban). Speakers can also be found in other parts of the Borneo Island such as Kalimantan, Indonesia; however, we limit our focus to Iban speakers in Sarawak.

Based on the EGIDS scale, Iban is among six Malaysian languages in status 3 (Wider communication). This means that there are language developments and it is used and sustained by certain bodies or institutions beyond the home and community [Lewis and Simons, 2010]. Since the early 90s, Iban is taught in schools in Sarawak at the primary and secondary level as an elective subject. Teaching effort has also recently spread to the university level, where several universities offer courses to undergraduate students for learning basic Iban. Local radio stations such as *Radio Televisyen Malaysia* (RTM) and Cats Radio, broadcast news and programs in Iban. Local newspaper such as “The Borneo Post” writes news in several languages including Iban. Several organizations such as Tun Jugah Foundation<sup>2</sup>, *Majlis Adat Istiadat Sarawak*<sup>3</sup> (Sarawak Native Customs) and the state’s education department play important role in standardizing the Iban system and gathering information on culture and literature.

According to Omar [1989], Iban is historically related to Malay where both languages belong to one branch called, Malay-Iban. This hypothesis was made due to the fact that there are many similarities between the two compared to other languages from other subgroups. Furthermore, a Malay speaker can guess the words are from Iban (or a Malay dialect), given a text or speech.

---

<sup>1</sup>Source from the Department Statistics Malaysia Portal [www.statistics.gov.my](http://www.statistics.gov.my)

<sup>2</sup>Official website of TJF : [tunjugahfoundation.org.my](http://tunjugahfoundation.org.my)

<sup>3</sup>Official website of Majlis Adat Istiadat Sarawak : [www.nativecustoms.sarawak.gov.my](http://www.nativecustoms.sarawak.gov.my)

### 4.3.1 Phonology

Omar [1981] provided the first description of the language in 1981 where among her work included phonological descriptions for Iban. Based on the author's study, the language has 19 consonants, 6 vowels and 11 vowel clusters.

#### 4.3.1.1 Vowel phonemes

Following describes the type of vowels in Iban and examples on how they are used.

Front vowels:

- /i/: /galiʔ/ = to lie down, /diŋa/ = listen
- /e/: /deʔ/ = you, /meh/ = an emphasizing word

Back vowels:

- /o/: /koŋ/ = tin or plastic cup, /noan/ = you
- /u/: /untuəŋ/, /nuan/ = you

Central vowels:

- /ə/: /bərap/ = to embrace, /kə/ = to, who, which, that
- /a/: /mata/ = eye(s), /atap/ = roof

#### 4.3.1.2 Consonant phonemes

Iban has 19 original consonants and we describe them using examples from the author's work, as shown in Table 4.3.

#### 4.3.1.3 Vowel clusters

Each vowel cluster in Iban consists of only two vowels. There are three types of vowel clusters in Iban; fronting, backing and centering [Omar, 1981]. Table 4.4 shows the phonemes and word examples.

Classification	Phoneme	Example
Plosive	/p/	/pandʒaj/ (long), /pupuəs/ (completed)
	/b/	/badas/ (good), /bini/ (wife)
	/t/	/tətaʔ/ (to cut up), /mataʔ/ (raw)
	/d/	/dampiəh/ (near), /diaʔ/
	/k/	/kaki/ (foot), /makaj/ (eat)
	/g/	/gaga/ (happy), /gəliʔ/ (amused, eerie feeling)
	/ʔ/	/gəmuʔ/ (fat), /ŋagaʔ/ (doing)
Affricate	/tʃ/	/tʃiru/ (clear, bright), /tʃəlap/ (cold)
	/dʒ/	/dʒalaj/ (walk, road), /pədʒam/ (close)
Nasal	/m/	/mandiʔ/ (shower, bath), /mənsia/ (human, people)
	/n/	/niŋa/ (listen), /nama/ (name)
	/ɲ/	/ɲaliən/ (copying), /məɲadiʔ/ (sibling)
	/ŋ/	/magaŋ/ (all)
Fricative	/s/	/sajaw/ (love), /sinuʔ/ (to feel pity)
	/h/	/humah/ (house), /tusah/ (difficult)
Trill	/r/	/rumeah/ (house), /bərat/ (heavy)
Lateral	/l/	/ləlaʔ/ (tired), /bulu/ (hair)
Semi-vowel	/w/	/gawaj/ (a Dayak festival) /ɲawa/ (voice, mouth, life)
	/j/	/gaju/ (long life), /ukuj/ (dog)

TABLE 4.3: Iban consonant phonemes

Vowel cluster type	Phoneme	Example
Fronting	/ai/	/kumbai/ (call)
	/ui/	/ukui/ (dog)
Backing	/ia/	/kiaʔ/ (going there)
	/ea/	/rumeah/ (house)
	/ua/	/kuap/ (mould)
	/oa/	/menoa/ (village, place)
	/iu/	/niup/ (to blow)
	/au/	/taun/ (year)
Centering	/iə/	/biliəʔ/ (room)
	/uə/	/puən/ (at the beginning)
	/oə/	/boəʔ/ (hair)

TABLE 4.4: Iban vowel clusters and examples

### 4.3.2 Writing system

The writing system for Iban is similar to Malay where it also uses Latin alphabets. All phonemes are represented with the following letter(s) as presented in Table 4.5. Letters ‘f’, ‘q’, ‘v’, ‘x’ and ‘z’ are also used to present borrowed words such as ‘*sulfur*’, ‘*Quran*’, ‘*universiti*’, ‘*xeroks*’ and ‘*zoo*’. Like Malay, the glottal stop ‘ʔ’ can be expressed using ‘k’ or not written at all. It occurs at word-final position such as /jakuʔ/ ‘*jaku*’ (speak) or /biaʔ/ ‘*biak*’ (child, kid).

Phoneme	Capital letter	Small letter	Example
/a/	A	a	<i>apai</i> (father)
/b/	B	b	<i>batu</i> (rock)
/ʈ/	C/Ch	c/ch	<i>chabut</i> (cup)
/d/	D	d	<i>dapur</i> (kitchen)
/ə/	E	e	<i>empai</i> (not yet)
/e/	E	e	<i>chire</i> (tasteless)
/g/	G	g	<i>gerah</i> (loose)
/h/	H	h	<i>hari</i> (day)
/i/	I	i	<i>mali</i> (forbidden)
/dʒ/	J	j	<i>jalai</i> (walk, road)
/k/	K	k	<i>kami</i> (we, us)
/l/	L	l	<i>malam</i> (night)
/m/	M	m	<i>makai</i> (eat)
/n/	N	n	<i>nama</i> (name)
/ɲ/	Ny	ny	<i>nyamuk</i> (mosquito)
/ŋ/	Ng	ng	<i>ngajat</i> (dance)
/o/	O	o	<i>chunto</i> (example)
/o/	-	u	/ɲebot/ <i>nyebut</i> (said)
/p/	P	p	<i>apai</i> (father)
/r/	R	r	<i>raban</i> (group, community)
/s/	S	s	<i>asuh</i> (instruct)
/t/	T	t	<i>datai</i> (come)
/u/	U	u	<i>ketup</i> (bite)
/w/	W	w	<i>lawa</i> (arrogant)
/j/	Y	y	<i>kaya</i> (rich)
/ai/	Ai	ai	<i>kumbai</i> (call)
/ui/	-	ui	<i>ukui</i> (dog)
/ia/	-	ia	<i>kiak</i> (going there)
/ea/	-	a	<i>rumah</i> (house)
/ua/	-	ua	<i>kuap</i> (mould)
/oa/	-	ua	<i>menua</i> (village, place)
/iu/	-	iu	<i>niup</i> (to blow)
/au/	-	au	<i>taun</i> (year)
/iə/	-	i	<i>bilik</i> (room)
/uə/	-	u	<i>pun</i> (at the beginning)
/oə/	-	u	<i>buk</i> (hair)

TABLE 4.5: List of phonemes and alphabets used in Iban

## 4.4 Malay and Iban relationship

Based on the phonological descriptions in previous sections, we compare the list of Malay and Iban phonemes to see the common and different phonemes between the two. As shown in Table 4.6, both has common consonants and vowels if we exclude borrowed phonemes from other languages. The obvious difference is the vowel cluster, which appears more in Iban than in Malay. The common vowel clusters (as defined by Omar

[1981], or diphtongs by Maris [1979]) are Iban /ai/ and /au/ and with Malay /aj/ and /aw/.

Iban vs. Malay	Phonemes
Common	Consonants: p, b, m, w, t, d, n, ʈ, dʒ, s, l, r, ɲ, j, k, g, ŋ, h, ʔ Vowels : a, e, ə, i, o, u V. Clusters/Diphtongs: ai, au
Difference	<b>only appear in Malay</b> Consonants: f, v, θ, z, ʃ, ʒ, ð, ʃ <b>only appear in Iban</b> V. Clusters : ui, ia, ea, ua, oa, iu, iə, uə, oə

TABLE 4.6: Iban and Malay common and different phonemes

In terms of orthography, both languages are written using Latin alphabets. There are common words (same surface form) in Malay and Iban, though sometimes they are pronounced rather differently in order to distinguish the word origin. In Table 4.7, we show some examples of common words and their phoneme sequences for Malay and Iban. Although they are written the same manner, some pronunciations are distinct due to the vowel clusters and glottal stop.

No.	Word [meaning(s) I : M]	Iban	Malay
1	ke [I=M : to]	/kə/	/kə/
2	nya [that : him/her]	/ɲaʔ/	/ɲə/ or /ɲa/
3	kayu [I=M: wood]	/kaʃuʔ/	/kaʃu/
4	bilik [I=M: room]	/biliəʔ/	/bileʔ/
5	dua [I=M: two]	/duwa/	/duwə/ or /duwa/
6	kepala [head, leader : head]	/kepalaʔ/	/kepala/
7	puluh [I:M : -ty (e.g. quantity)]	/puluəh/	/puluh/
8	raban [group : rambling speech, talk rapidly]	/raban/	/raban/
9	lalu [then, pass-by : then, before (time), unwell ]	/lalu/	/lalu/
10	orang [I=M : person, people]	/uraŋ/	/oraŋ/

I : Iban, M : Malay, I = M : same meaning in Malay and Iban

TABLE 4.7: Malay-Iban examples with their pronunciations

A study on lexical cognates between Malay and Iban using lexicostatistics approach has been conducted by Yusof [2003]. The method was developed by Swadesh [1952] to investigate the distance between close languages (languages that belong to the same group) for defining the status of the languages as dialect or non-dialect. Cognates are words evolved from the same origin. They can be classified as true cognates (*vrais*

*amis*) or false cognates. Words that are spelled similarly in both languages and has the same meaning are true cognates such as *accent*, *cigarette* and *ordinaire-ordinary*, for instance, are true French-English cognates. On the other hand, false cognates (*faux amis*) are words that are similar in form but are quite different in meaning. The word pair *actuellement* (at the present time) in French and *actually* (in fact) in English, or, *blanco* (white) in Spanish and *blank* (empty) in English are classic examples of false cognates. Lexicostatistics approach measures the distance between languages by comparing lexical cognates quantitatively. A word list such as the Swadesh list can be used by linguist to find words in the languages involved, which are close to the meanings in the list. Then the linguist decides the cognacy level for each pair of words (true, false or undetermined). If the cognates falls below 85% (more false cognates), this means that the status for the two subjects (languages) is *different*. Meanwhile, if the result is above 85% (more true cognates), the languages are considered as *dialects*. Yusof [2003] used 100 words from the Swadesh list [Lehmann, 1993] and translated the words in Malay and Iban. From the Malay-Iban list, the author found that the level of cognacy of both languages is 69%. This implies that more than half of word pairs in the list are similar in spelling and also meaning (Other examples: French and English are connected by 24%, while German and English are similar by 57% [Dyen et al., 1992]).

Another interesting study on Malay-Iban relationship is by Ng et al. [2009] for measuring the orthographic similarities between Malay and Iban as well as with other Sarawak languages. One of the methods that they used was the Levenshtein distance. It was used for calculating edit distance on 200 Malay-Iban word pairs. The word pairs were also translations of the Swadesh list. A score scale was proposed for classifying the word pairs as true or false cognates. For word pairs with edit distance below a threshold value, they were considered as true cognates otherwise, they were false cognates. Subsequently, the number of word pairs that fall in between the two intervals were counted for finding the level of cognacy. Based on their final outcome, Iban has the highest cognate percentage with Malay (61%) compare to other Sarawak languages like Kelabit, Melanau and Bidayuh. Table 4.8 shows a summary of their evaluation results on eight languages.

Language	Malay	Iban	Melanau	Bidayuh	Kelabit	Penan	Sa’ban	English
Iban	61.0	-	-	-	-	-	-	-
Melanau	40.7	34.7	-	-	26.4	-	-	0.5
Bidayuh	33.5	24.0	30.2	-	22.5	-	-	0.6
Kelabit	25.3	17.2	-	-	-	-	-	1.7
Penan	20.5	20	27.7	13.1	26.3	-	16.3	0
Sa’ban	11.3	10.6	15.5	9.9	28.7	-	-	0.7
English	1.5	0.5	-	-	-	-	-	-

TABLE 4.8: Level of cognacy for eight languages calculated using normalized Levenshtein Distance method. As presented in [Ng et al., 2009]

## 4.5 Data for ASR tasks

Previous sections in this chapter have mentioned a brief introduction to the languages from Malaysia and described the phonology and writing system for Malay and Iban. Here, we report about the available data for our ASR experiments. Three databases: MASS (Malay), TED-LIUM (English) and non-native English are existing data while the Iban database was created as a part of this research. All audio data used in this study have sampling frequency of 16kHz and a quantization of 16 bit to fit the format of Kaldi.

### 4.5.1 MASS

The MASS [Tan et al., 2009] corpus contains clean, read speech in Malay of about 140 hours. The data collection was carried out through a collaborative work between three universities: Universiti Sains Malaysia, Multimedia University Malaysia and Nanyang Technological University Singapore. The database was designed to be utilized for conducting speech recognition research tasks (see [Tan and Rainavo-Malançon, 2009], [Xiao et al., 2010], [Juan et al., 2012] for research on Malay ASR). The read texts were obtained from newspaper articles concerning local and international news, economy, entertainment, sports and others. A total of 199 speakers had their voices recorded and each of them read about 45 minutes. Table 4.9 summarizes the details of the Malay speech data based on ethnicity and experiment setting. The pronunciation dictionary for the system has about 64K words with 76K pronunciations (pronunciation variants included). Each pronunciation was generated using a rule-based Malay grapheme-to-phoneme (G2P) tool [Tan and Rainavo-Malançon, 2009]. The G2P tool

utilized 34 phonemes to transcribe the graphemes. The language model for decoding is a trigram model which was built from news data dated from 1998 to 2011, with a vocabulary size of 59K words. We obtained a perplexity of 185 after verifying the model on the test data.

		Malay	Chinese	Indian	Other
Train	Hours	40.1	73.9	5.9	2.3
	Sentences	18,298	29,475	2,638	965
	Words	277K	451K	39K	15K
	Speakers	64	102	8	3
Test	Hours	7.6	7.7	1.6	-
	Sentences	3,231	3,150	663	-
	Words	48K	48K	10K	-
	Speakers	10	10	2	-

TABLE 4.9: Size of the Malay data set for different speaker origins (ethnicities) in hours, number of sentences and number of speakers

#### 4.5.2 TED-LIUM

The second database is the English TED-LIUM [Rousseau et al., 2012] corpus. The corpus was built by the Laboratoire d’Informatique at Université du Maine (LIUM) for the 2011 evaluation campaign on spoken language translation in the International Workshop on Spoken Language Translation (IWSLT). The task was to decode and translate English speeches from TED (Technology, Entertainment, Design) conferences to French. The training set contains speeches that were extracted from the video talks via the TED website. Table 4.10 provides information on the corpus. It has 118h of speech for 774 talks by 666 speakers. The development set as shown in the table is a manually transcribed data from the IWSLT 2010 development and test corpora, which is used to test our English ASR system. Besides that, the pronunciation dictionary for ASR was also developed by the team. It has a vocabulary size of 157.6K words. The phoneme sequences in the dictionary were based on CMU pronunciation dictionary (using Arpabet). Out-of-vocabulary (OOV) words were treated using the Festival Speech Synthesis System [CSTR, 2012] for generating missing pronunciations. The trigram language model has 50K words, is trained on TED and WMT11 (Workshop on Machine Translation 2011) data with SRILM toolkit [Stolcke, 2002]. The model has perplexity of 220 after we estimated it on the dev set.



	Male	Female	#speakers	#talks	#segments	#words
Train	81h 53m 7s	36h 11m 41s	666	774	56.8k	2.56M
Dev	3h 13m 57s	58m 58s	19	19	2k	47k

TABLE 4.10: TED-LIUM corpus information

### 4.5.3 Iban

Set	Speakers	Gender (M:F)	Sentences	Words	Minutes
Test	6	2:4	473	11K	71
Train	17	7:10	2659	61K	408

TABLE 4.11: Amount of Iban manually transcribed speech

The Iban speech corpus is a small news database, which we obtained from the *Radio Televisyen Malaysia Berhad*, a local radio and television station in Malaysia. We have almost eight hours of speech, spoken by 23 speakers. After manually transcribing the speech data, we have more than 3K sentences as shown in Table 4.11. Chapter 5 will further explain on the development of the Iban corpora, language model and pronunciation dictionary. To the best of our knowledge, this is the first Iban database created for ASR experiments.

### 4.5.4 Non-native English

The fourth corpus is an English speech data, spoken by Malaysians. The data was collected by Universiti Sains Malaysia for conducting a recent speech recognition research (see [Tan et al., 2014]). It contains 14h of speech spoken by 24 speakers of different origin: Malay, Chinese and Indian. The read texts were taken from news articles from a local newspaper. Table 4.12 shows the train and test sets for our experiments. Our training corpus contains 2h of transcribed (Trans.) data for building a very low-resource ASR. The untranscribed (Untrans.) data with 9h of speech is used for training the UBM in our SGMM experiments. For decoding, we use a trigram language model that was built with SRILM toolkit, using texts from *The Star* (a local newspaper) news articles. The model has a perplexity of 188 after evaluating on the test data. Lastly, we utilize the CMU dictionary with more than 100K words (no non-native pronunciations) for our ASR experiments.

Data	Speakers	Gender (M:F)	Amount of data	#words
Train (Untrans.)	12	5:7	9h	-
Train (Trans.)	6	3:3	2h	12K
Test	6	2:4	4h	29K

TABLE 4.12: Overview of the non-native (Malaysian) English corpus

## 4.6 Summary

This chapter provided a brief introduction on the spoken languages in Malaysia and we described two important languages related to our study, Malay and Iban. We presented the general view of the sound and writing system of each language. Furthermore, we revealed the relationship between Malay and Iban based on our observations and past studies. In terms of phonology, both has similar set of consonants but more vowel clusters exist in Iban than in Malay. Orthographically, both are written using Latin alphabets and there exist Malay and Iban words with same surface forms. Also in this chapter, we reported the characteristics of the speech databases for our ASR experiments. In the following chapter, we will report on the Iban data collection and development of resources for the ASR system.



## Part II

# Data collection methodology and Iban automatic speech recognition development



## Chapter 5

# Rapid development of Iban resources

### 5.1 Introduction

To conduct our research, we built Iban speech and text corpora as there were none available. This chapter describes the methods used for obtaining the resources. The first section explains about the source of the Iban speech and text data while the second section demonstrates our proposed strategy to quickly build a pronunciation dictionary for Iban.

### 5.2 Quick collection of text and speech data

#### 5.2.1 Initial available data

Collecting speech data can be a difficult task especially when there is lack of equipment or staff for conducting the task. We were given the permission from a local radio station to use Iban news data for building the speech corpus. The news recording was done between March to June 2012 at the *Radio Televisyen Malaysia Berhad* station in Kuching, Sarawak. We received 3GB of audio data for pre-processing, unfortunately, not all recordings were in good quality. Some of the wav files had music or commercial breaks that were too long and some of the recorded news had too much distortion and

the speeches were not very clear. We were left with 1.1 GB of audio data to work on after discarding the low quality files and removing noises.

### 5.2.2 Transcribing speech corpus through a collaborative workshop

Speech transcription data was required for building an ASR system. To transcribe the news data, a data collection workshop was conducted in the Faculty of Computer Science and Information Technology in Universiti Malaysia Sarawak, Malaysia. The goal of the workshop was to have users with basic skills on using Transcriber<sup>1</sup> [Barras et al., 2000] for annotating speech signals. We hired eight native speakers to take part in the workshop and to produce the transcription data.

The workshop lasted for three days, from 17 to 19 July 2013. We provided basic tutorial such as segmenting speech signals according to sentences, discarding noise such as music, page turns, etc; and annotating speech to get transcription. All of the speakers were computer users, therefore, the tutorial session was easily conducted. However, transcribing the whole corpus took almost two weeks to complete because each user had different speed in finishing their task. During the process, some of them were unsure of the correct way to spell the words. To ease this problem, we provided a hardcopy of an Iban dictionary [Ensiring et al., 2011] for them.

During this workshop, we have gathered more than 3K sentences which were uttered by 25 speakers in the eight hours of clean speech. The corpus statistics have been described in Table 4.11 of Section 4.5.3.

### 5.2.3 Collecting large amount of text data

We found an online news website<sup>2</sup> that publishes Iban articles over the past few years. From this website, we extracted articles dated from 2009 to 2012 through web crawling approach. In total, we obtained 7K articles on sports, entertainment and general matters. The next procedure was normalizing the raw data. We employed the following steps to clean the data: (1) remove HTML tags, (2) convert dates and numbers to words (e.g: 1973 to *sembilan belas tujuh puluh tiga*), (3) convert abbreviations to full terms (e.g:

---

<sup>1</sup>Available at: <http://trans.sourceforge.net/en/presentation.php>. There is also another version called TranscriberAG at <http://transag.sourceforge.net/> but we used the former version for this work.

<sup>2</sup><http://www.theborneopost.com/news/utusan-borneo/berita-iban/>

Dr. to *Doktor*, Prof. to *Profesor*, Kpt. to *Kapten*), (4) split paragraphs to sentences, (5) change uppercase characters to lowercase and (6) remove punctuation marks (except hyphen / '-'). After completing these steps, the text contains 2.08M words with 36,358 unique words. We used the text to build a trigram language model with modified Kneser-Ney discounting applied on SRILM toolkit. Following that, we evaluated the language model over the speech transcription data and obtained a perplexity of 158 (2.3% OOV rate).

### 5.3 Semi-supervised approach for building the Iban pronunciation dictionary

The next task was to build an Iban pronunciation dictionary for ASR. Most of the time, linguists are required to produce phonemic transcripts for a word lexicon. If a large vocabulary of words is involved, it would require a lot of time to complete if the work is done manually. Moreover, humans are prone to make mistakes if they lose focus. In our scenario, we proposed a fast bootstrapping strategy for creating the phonetic lexicon for Iban without defining a new phoneme set for the language.

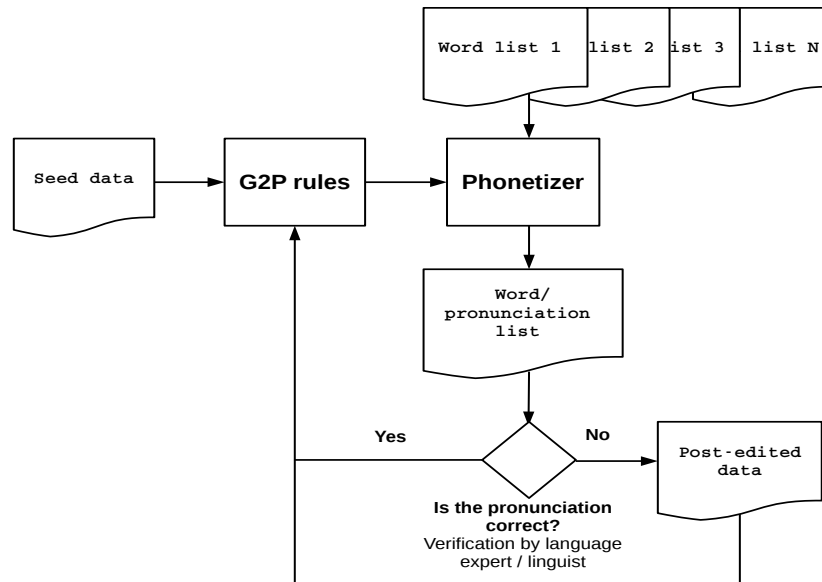


FIGURE 5.1: General process of our bootstrapping approach



### 5.3.1 Bootstrapping grapheme-to-phoneme (G2P) conversion

Bootstrapping G2P is a strategy to reduce effort of producing phonetic transcriptions for all the words in the vocabulary list from scratch. Maskey et al. [2004] introduced the method in their work to produce a pronunciation dictionary for Nepali and English. Figure 5.1 illustrates the concept of bootstrapping G2P. This semi-supervised method requires a small transcript that has words and pronunciations in a target language, usually prepared by a native speaker or a linguist. This transcript then becomes the seed data for building G2P rules. Then, the pronunciation model is used to predict new entries in the vocabulary and post-editing can be carried out later, if needed. The process of updating the model can be repeated by adding the post-edited list into the model. The method was also used by Davel and Barnard [2003] by incorporating machine learning and human intervention for building a list of pronunciations. Their strategy was tested over German and Afrikaans [Davel and Barnard, 2004] languages.

Instead of using Iban data for the seed model, our proposed approach used an existing Malay G2P for obtaining initial transcript for Iban. This transcript can be improved to acquire Iban pronunciations. Since both languages have language contact (share phonemes and words), completing the task can be rather quick. Besides that, using Malay phoneme set can be an advantage for Iban since both have almost similar phonological description.

### 5.3.2 Deeper analysis on our Iban lexicon

Prior to that, we made an analysis to find the amount of words that are similar to Malay or English in the Iban lexicon. We also verify for English as we observed that the crawled texts contain English words. The Iban lexicon was validated using the Malay MASS [Tan et al., 2009]) and English CMU dictionaries.

Corpus	Vocab. size	Identical words	
		with English	w/o English
Malay	63,900	13,774	8,472
Iban	36,358		

TABLE 5.1: Number of identical (same surface form) words found in Iban and Malay lexicons - size of Iban lexicon (36K)

Table 5.1 shows that more than thirteen thousand words are shared between Malay, Iban and also surprisingly English. After removing the English words, the amount of Malay-Iban common surface forms is 8,472. Conclusively, 23% (8,472) of the Iban lexicon is shared with Malay, 19% (6,707) with English, while the remaining 58% (21,179) purely belong to Iban. In other words, 42% of this lexicon is found shared not only with one, but, with two languages, English and Malay. This gave us an idea of language contact and code switching issues related to Iban language.

### 5.3.3 Measuring pronunciation distance of Malay and Iban

Since we wanted to use Malay as our source for G2P task, we were concerned over the cost of transforming a Malay phoneme sequence to an Iban one. Past study [Ng et al., 2009] has applied Levenshtein distance to estimate the difference between Malay and Iban based on orthography. Here, we applied the same method but using phoneme sequences for measuring pronunciation distances. We can relate to the study by Heeringa and de Wet [2008], who have applied this approach for comparing Germanic languages and Dutch dialects using pronunciations. Their results showed that Dutch and Afrikaans have closer pronunciations than Dutch and Frisian or, Dutch and German.

#### 5.3.3.1 Obtaining a Malay G2P

The first step is to build a Malay G2P system. The system was built in order to generate an initial pronunciation transcript for Iban. We built the phonetizer using data from a Malay pronunciation dictionary - used to build Malay ASR using MASS corpus. The dictionary contains 63.9K unique words with 76.05K pronunciations, including pronunciation variants. Initially, the Malay phonetic lexicon was developed using a rule-based G2P [Tan and Rainavo-Malançon, 2009] but we did not have access to this tool. We developed a G2P on Phonetisaurus [Novak et al., 2011] using 68K pronunciations. The pronunciation model is then built by training an  $N$ -gram model using grapheme-phoneme alignments produced by the tool. We used a seven-gram model with original Kneser-Ney smoothing, built on SRILM. Subsequently, we evaluated the phonetizer using 8.05K Malay words (not from training data). Our results are 6.20% phoneme error rate (PER) and 24.98% word error rate (WER).

### 5.3.3.2 Evaluating using Levenshtein distance

We focused on Malay-Iban common surface forms (hereafter we address this as Malay-Iban words) only for analyzing pronunciation distances. We selected 100 most frequent Malay-Iban words for phonetization. Subsequently, we used Malay G2P system to phonetize the Malay-Iban graphemes. The phonemic transcripts were later corrected by an Iban speaker<sup>3</sup>. Then, we calculated the costs for converting Malay G2P output to the post-edited one based on the number of insertions, substitutions and deletions. As an outcome, we obtained 17% PER and found out that 53 of the pronunciation pairs were equivalent (no change)! This result confirms that the use of a Malay G2P can be a good starting point to bootstrap an Iban G2P system.

The results were analyzed to determine which phonemes were frequently edited. Phoneme /o/ was frequently substituted with /uə/. For example, the word *puluh* is transcribed as /puluh/ in Malay while in Iban, it is /puluəh/. This substitution occurred because the vowel cluster /uə/ appears in Iban and not in Malay. Also, we observed that phoneme /e/ was frequently substituted by /iə/. For example, /pəsiser/ in Malay converts to /pəsisier/ in Iban, for the word *pesisir*. In the post-edited transcript, glottal stop /ʔ/ was inserted at almost all words ending with a vowel. For instance, *kepala* is transcribed as /kəpala/ for Malay, but it needs a glottal stop at the final vowel to change to /kəpalaʔ/ to sound like Iban.

### 5.3.4 Obtaining Iban G2P training data via post-editing Malay G2P output

Subsequently, we selected 500 Malay-Iban and 500 pure Iban (not shared with neither Malay nor English orthography) words for training. After that, Malay G2P was used to produce the phonemic transcript and later the outputs were manually post-edited. The task took about 1 hour and 30 minutes to complete. We evaluated Malay G2P outputs with the post-edited transcript and found the G2P performances are 11.88% PER and 48.9% WER. Using the new references, we developed a first Iban G2P system on Phonetisaurus. Figure 5.2 illustrates the process to build the phonetizer.

---

<sup>3</sup>the author of this thesis

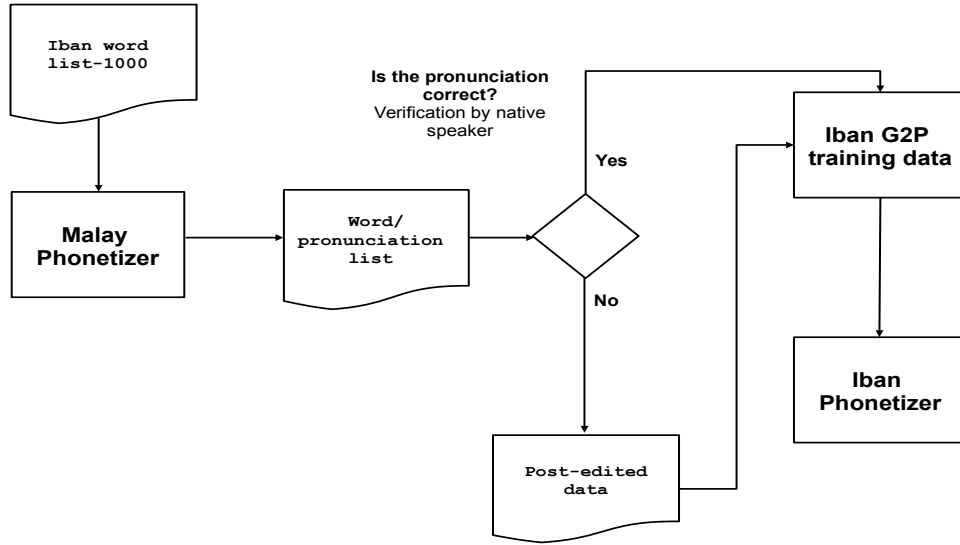


FIGURE 5.2: Steps to develop an Iban phonetizer through bootstrapping approach

The next step was to evaluate the two phonetizers on an Iban test data. We selected another 1K words from the Iban lexicon using similar setting (500 pure Iban and 500 Malay-Iban) for this task. Then, each G2P system was used to transcribe the test set. Both set of outputs were manually corrected in order to get a final transcript. Subsequently, we checked all transcripts to evaluate the performance level of the G2Ps. Table 5.2 shows our results based on phoneme and word error rates (PER and WER), as well as the amount of time taken to complete post-editing task. The time spent to correct Iban G2P output for pure Iban words was shorter than post-editing Malay G2P output. We also discovered that Malay G2P system performed best for Malay-Iban matching words and the result is consistent with Malay G2P performance on Malay test set. On the other hand, Iban G2P system gave better results for pure Iban words. Despite of the small amount of data used to train the Iban G2P (1000 words), it was able to perform less than 10% PER. However, it seems that Iban G2P is not suitable to phonetize common Malay-Iban words (PER increases from 6.52% to 13.6%).

We further analyzed each G2P outputs for identifying wrongly substituted or deleted phonemes. For both pure Iban and Malay-Iban words, Malay G2P substituted phonemes /uə/, /iə/, /ea/, with /o/, /e/ and /a/, respectively, while the glottal stop /ʔ/ and phoneme /r/ were missing. On the contrary, Iban G2P substituted phonemes /ə/ for /e/, /iə/ for /i/, /uə/ for /u/ and /uə/ for /o/. Moreover, the Iban system wrongly

Phonetizer	Corpus	PER (%)	WER (%)	Post-edit (mins)
Malay G2P	500 <sub>IM</sub>	6.52	27.2	30
	500 <sub>I</sub>	15.8	56.0	42
Iban G2P	500 <sub>IM</sub>	13.6	44.2	45
	500 <sub>I</sub>	8.2	31.8	32

Note: *IM* for common Malay-Iban words and *I* for pure Iban words

TABLE 5.2: Malay G2P and Iban G2P performances for Iban phonetization task

inserted glottal stop. Yet, the system gave better phoneme sequences than Malay G2P for pure Iban because the latter has no phoneme sequences that can represent vowel clusters in Iban. The impact is different for Malay-Iban because Malay G2P conversion was more accurate than Iban G2P. Many outputs were retained after the post-editing process. This was mainly due to adopted words from Malay which we found in the data (e.g; *parlimen* (parliament), *menteri* (minister) and *muzik* (music)).

### 5.3.5 Phonetization of the whole Iban lexicon

From previous experiments, we observe that Malay G2P and Iban G2P are equally important for transcribing the two groups of Iban words. For that reason, we proposed to transcribe the whole Iban lexicon using both G2P modules. The following strategy was applied: Malay G2P phonetizes all Malay-Iban, while Iban G2P phonetizes all pure Iban words. Also, Malay G2P contains English data, too. Therefore, to deal with English words in Iban lexicon, we used Malay G2P to phonetize as English data is available<sup>4</sup>.

For measuring the accuracy of the G2P outputs, first, we post-edited 2K random phonemic transcripts. The random list had 1426 pronunciations from pure Iban and 574 pronunciations from Malay-Iban. Then, we made a comparison between the post-edited and original outputs and obtained 8.1% PER and 29.4% WER. These results show improvements in phoneme and word accuracies when compared to Iban G2P and Malay G2P performances as reported in Table 5.2. In terms of phoneme accuracy, we achieved 2.8% PER reduction from Iban G2P performance and 3% PER reduction from Malay G2P performance (average of 500<sub>IM</sub> and 500<sub>I</sub> results). Finally, Table 5.3 summarizes the results of Malay and Iban phonetizers on Malay and Iban phonetization tasks<sup>5</sup>.

<sup>4</sup>see English-Malay phoneme mapping conducted in Tan and Rainavo-Malançon [2009] as apart of their work to build the rule-based G2P tool

<sup>5</sup>see Appendix C for our experiment and results on Iban phoneme-to-phoneme (P2P) system

Phonetizer	#words	PER	WER
Malay G2P	8050 (Malay)	6.2	24.98
Malay G2P & Iban G2P	2000 (Iban)	8.1	29.4

TABLE 5.3: Performance of Malay and Iban phonetizers based on phoneme error rate (PER) and word error rate (WER) (%)

## 5.4 Summary

In this chapter, we have described our strategies for obtaining Iban speech and text corpora to build ASR system. We have applied several steps in acquiring the data: (1) made contacts with a local radio station to provide news speech, (2) organized a collaborative workshop for transcribing the news data, (3) implemented web-crawling approach for online data acquisition and (4) obtaining Iban pronunciation dictionary through semi-supervised approach. In the process of building the Iban pronunciation dictionary, we have discovered that pronunciation distance between Malay-Iban same surface forms are close. Our initial investigation showed 53 pronunciation pairs were equivalent for 100 Malay-Iban words. This large number was due to the close phonological relationship between Malay and Iban (shared phonemes). Motivated by this finding, we have successfully built our first Iban G2P using 1K post-edited Malay G2P outputs. The newly acquired system was tested along with Malay G2P. The G2P evaluation results showed that both Malay G2P and Iban G2P were necessary for phonetizing different word groups (Malay-Iban or pure Iban). Thus, both G2Ps were employed in our final strategy to produce Iban pronunciation dictionary for ASR. In summary, we have 8 hours of transcribed speech, 2M words in the text corpus and 36K words in the pronunciation dictionary. Besides that, we have trained a trigram language model using the text corpus for ASR decoding. Most importantly, we have made our Iban data and Kaldi scripts available for download in github<sup>6</sup> for the speech community to for research.

<sup>6</sup><https://github.com/sarahjuan/iban>



## Chapter 6

# First Iban ASR system developed

### 6.1 Introduction

In Chapter 5, we have presented the data collection strategies for building our Iban speech and text corpora, the first database ever created for ASR experiments in this language. To measure the performance of the recently acquired pronunciation dictionary, we build several more to test on our Iban ASR. This chapter describes our first approaches for developing Iban ASR using our data and analyzing the performances.

### 6.2 Setup for experimentation

#### 6.2.1 Text and speech corpus

Full description of the train and test set for speech recognition experiments is available in Section 4.5.3. Briefly, we used 7h to train a speech recognition system and the remaining data (1h) was used for evaluation. For decoding, we utilized a trigram language model which was built on news data.

#### 6.2.2 Pronunciation dictionaries

We tested several pronunciation dictionaries on the speech recognition system. In Section 5.3, we have demonstrated a semi-supervised approach via bootstrapping a



grapheme-to-phoneme (G2P) system for obtaining Iban G2P system. Iban G2P has 1K pronunciations and performs better in converting grapheme to phoneme for pure Iban. On the other hand, Malay G2P, a system which was trained on 68K pronunciation data, is useful for transcribing Malay-Iban (Malay and Iban same surface forms). Hence, we have phonetized the whole Iban lexicon using both G2P systems and we called it as Hybrid G2P. To evaluate the performance of Hybrid G2P on our Iban ASR, we developed two Iban pronunciation dictionaries. Both dictionaries were acquired through applying Malay G2P and Iban G2P, respectively. We labelled each dictionary accordingly as Malay G2P and Iban G2P.

### 6.2.2.1 Analyzing phoneme sequences

In order to know the difference between the three dictionaries, we conducted out a comparison study on the phoneme sequences. The investigation was done by comparing phonemic transcripts from two dictionaries at a time. Using these data, we calculated the number of pronunciation pairs that were different.

Our findings are shown in Table 6.1. Here, we denote the pairs using the following labels for simplicity: Malay G2P as  $S1$ , Iban G2P as  $S2$  and Hybrid G2P as  $S3$ . Let  $C_{\mathbf{AB}}$  has elements that are **not** common to  $\mathbf{A}$  and  $\mathbf{B}$ . We found that 67% of pronunciations obtained by applying Malay G2P are different from the ones generated by Iban G2P. Besides that, the hybrid version ( $S3$ ) is similar to Iban G2P ( $S2$ ) by 71%.

$C_{\mathbf{AB}}$	No. of diff. pronunciations	%
$C_{S1S2}$	24,587	67.6
$C_{S1S3}$	14,162	39.0
$C_{S2S3}$	10,593	29.1

TABLE 6.1: Comparison results between two pronunciation dictionaries (total words 36K)

We analyzed further to know the origin of the words that have this set of different pronunciations (elements of  $C_{\mathbf{AB}}$ ). The words were identified according to three languages, English, Malay and pure Iban (or not English and Malay). English and Malay were identified using vocabulary lists from Malay G2P and English CMU systems. Results obtained after this identification process are presented in Table 6.2. After comparing  $S1$  (Malay G2P) with  $S2$  (Iban G2P), majority of the words belong to pure

Iban. The same trend was found after comparing Malay G2P with Hybrid G2P. As for English words, there was no difference for Malay G2P and Hybrid G2P outputs. This happened after English words in Hybrid G2P were phonetized by Malay G2P, which resulted similar phoneme sequences. The dissimilarities only showed after comparing phoneme sequences generated by Iban G2P with the other two dictionaries ( $C_{S1S2}$  and  $C_{S2S3}$ ) because the phonetizer lack of English references.

Language	$C_{S1S2}$	$C_{S1S3}$	$C_{S2S3}$
English	5,605	0	5,605
Malay	5,031	202	4,912
pure Iban	13,951	13,960	76

TABLE 6.2: Word statistics in Table 6.1 based on three languages

### 6.3 Baseline Iban speech recognizers

We experimented Kaldi [Povey et al., 2011b] for building all our ASR systems. We extracted 13 MFCCs from the training data and used Gaussian mixture models for monophone and triphone trainings. For the triphones, we employed 2,998 context-dependent states and 40,000 Gaussians. Subsequently, we implemented delta delta coefficients on the MFCCs, linear discriminant analysis (LDA) transformation and maximum likelihood linear transform (MLLT) [Gopinath, 1998]. Then, we applied feature-space maximum likelihood linear regression (fMLLR) [Gales, 1998] for speaker adaptation. The Iban ASRs were respectively called Malay G2P, Iban G2P and Hybrid G2P, after applying the proposed pronunciation dictionaries in the systems. Lastly, we set language model weights from 5 to 20 in Kaldi for decoding. As a result, we acquired 16 sets containing hypotheses from an ASR system. We report only the best results in the following sections.

#### 6.3.1 Results and analysis

The baseline results are presented in Table 6.3. After using context independent models, ASR performed with an average of 42% WER. The accuracies gradually increased when triphone models and different features employed. Employing speaker adaptive method gave the best results by providing 6-7% improvements for all systems. Generally, the

Training approach	Dictionary		
	Malay G2P	Iban G2P	Hybrid G2P
Monophone	42.17	41.79	41.97
Triphone	36.00	36.44	36.11
Triphone + $\Delta$ + $\Delta$	36.47	36.98	36.77
+ MLLT + LDA	27.24	27.71	26.80
+ SAT(fMLLR)	20.82	21.90	<b>20.60</b>

TABLE 6.3: Iban ASR performances based on word error rate (WER%) for different approaches

performances of the three systems are similar. Although Hybrid G2P has the best performance (20.6%), the result is only  $\sim 1\%$  different from Malay G2P or Iban G2P result.

### 6.3.1.1 Correlation between ASR and G2P performances

Consequently, we conducted an investigation to study the relationship between ASR and G2P systems based on recent results. To perform this, best results from the three decoders and estimated G2P results from Table 5.2 were used. Unfortunately, we found that the correlation is rather weak, as shown in Figure 6.1. From this graph, we observe that the performance for Hybrid system is the best among the other two systems, with G2P and ASR accuracies have 29.8% and 20.6% WERs, respectively. However, the bad news is that all our efforts to improve our Iban lexicon (including the hybrid approach) did not have a strong effect on ASR. The good news is that using G2P of a similar well-resourced language (such as Malay) seems to be a good starting point to generate pronunciations and build an ASR system for a very under-resourced language (such as Iban).

### 6.3.1.2 System combination

Kaldi supports system combination based on minimum Bayes risk (MBR) [Goel et al., 2003] for decoding. The method combines lattices from a number of systems and produces sequences that have least expected losses. We applied this approach to combine decoding results from the three systems. Using lattices from the best WERs (see Table 6.3), we combined the systems based on the strategies described in Table 6.4. Results

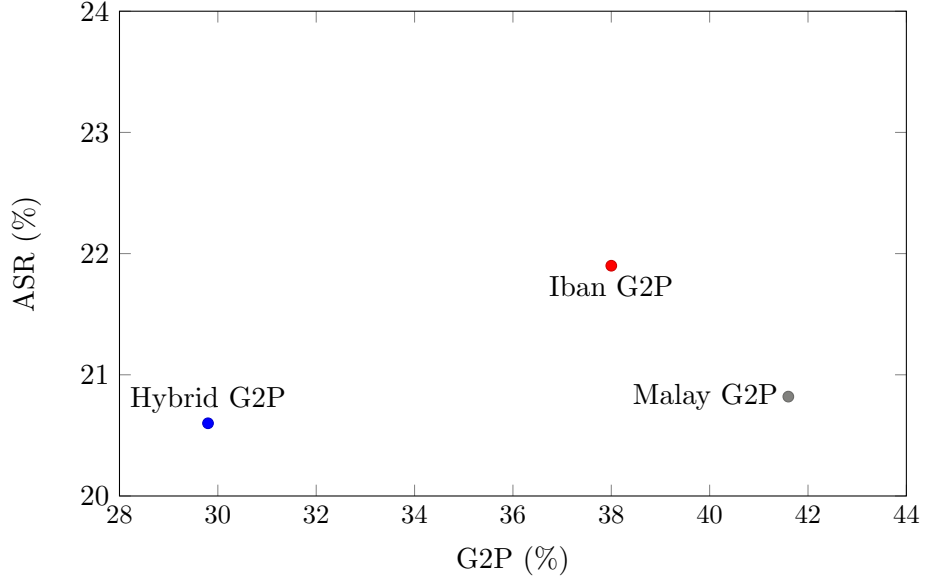


FIGURE 6.1: Iban ASR vs G2P based on WER (%) results

show that our baseline systems were outperformed by the combined systems. Moreover, adding more system in the combination provided the best result. However, we observe no significant difference between the new results (less than 1% different).

Combination	%WER
Hybrid G2P + Iban G2P	19.83
Malay G2P + Iban G2P	19.76
Hybrid G2P + Malay G2P	19.55
Hybrid G2P + Malay G2P + Iban G2P	19.22

TABLE 6.4: System combination and WERs

### 6.3.1.3 Confusion pairs

As a final evaluation, we conducted an error analysis to observe decoding outputs (HYP) and the test references (REF). The analysis was done by identifying confusion pairs between REF and HYP. Table 6.5 presents the top ten most frequent confusion pairs for Iban based on outputs from four ASR systems. Words on the left are words in REF while words on the right are from HYP. The first three columns have HYP from the single lattice decoders. The last column indicates results for HYP after system combination (Hybrid G2P + Malay G2P + Iban G2P) applied.

From the errors, we observe that there are normalization and morphological issues. For instance, the word *rakyat* (people) is a Malay word but the system recognized *rayat*,

Hybrid	Malay	Iban	Combine (H+M+I)
rakyat => rayat	rakyat => rayat	rakyat => rayat	rakyat => rayat
ka => ke	ka => ke	ari => hari	ka => ke
ti => ke	ari => hari	ka => ke	ari => hari
ari => hari	serta => sereta	serta => sereta	ti => ke
urang => orang	ti => ke	ti => ke	serta => sereta
serta => sereta	urang => orang	urang => orang	urang => orang
mohamad => mohd	datuk => dato	ke => ka	ke => ka
ka => madahka	ka => madahka	mohamad => mohd	mohamad => mohd
ke => bejalailka	ke => bejalailka	agensi => ijinsi	agensi => ijinsi
antara => entara	mohamad => mohd	ka => madahka	ka => madahka

**Bold** : normalization problem; Normal : morphological problem

TABLE 6.5: Top ten confusion pairs from Hybrid, Malay, Iban systems and system combination

which is actually correct for Iban. The former word exists in the REF causing a false substitution in the ASR performance. A possible reason for this orthographical mistake may be that the transcribers could have been influenced by Malay as the language is used frequently. Other examples are such as *urang* and *orang* (person), *agensi* and *ijinsi* (agency), and, *serta* and *sereta* (as well as / join). Common spelling variations are such as mohamad and mohd, *penerbai* and *penerebai* (airline), and, *ka* and *ke*. For the case of *ti* and *ke*, it appears that some of the sounds corresponding to *ke* were wrongly interpreted to *ti*. Honorific titles that are normally used in salutations such as *dato* (here apostrophe is neglected, original is *Dato'*) and *datuk* can be often misspelled because both have similar pronunciations (/dato?/). We have also identified morphological problems such as *ka* with *madahka* (also with *bejalaika* or *ngambika*), *waifm* and *fm*, as well as *sehari* and *tu* (full form *isseharitu* or *saritu* - found two versions in the reference). In Iban language, *ka* is a suffix that forms transitive verbs just like the suffix *kan* in Malay. It is concatenated with a root word but we found that it is frequently separated from the root word in the REF list.

## 6.4 Experimenting with cross-lingual and grapheme-based phonetizers for Iban ASR

Typically, phonetizers are language dependent where the rules for G2P conversions are carefully developed to present the right pronunciations of a target language. Previously, we showed that it is possible to use data from a language that is close to Iban for developing its pronunciation lexicon. We have also shown that using a pronunciation dictionary that contains Malay and Iban pronunciations in Iban ASR has helped in obtaining the best ASR hypotheses.

Our next task is to observe the impact of using a G2P that has been used to transcribe another language that is not close to Malaysian languages. Also, we investigate a grapheme-based G2P that has graphemes as representation of phonemes. Thus, our experiments involve G2Ps with phoneme or grapheme set that are different than the one used in Malay or Iban G2P.

To perform the first study, we used an existing English G2P based on CMU pronunciation data from Phonetisaurus. The system was downloaded directly from the

website and has been used as a demo tool. It has more than 100K training data, using 39 phonemes in Arpabet symbols. For the second task, we utilized a grapheme-based system built using Malay segmentation rules [Tan and Rainavo-Malançon, 2009].

Training approach	Grapheme	Dictionary			
		English G2P	Malay G2P	Iban G2P	Hybrid G2P
Monophone	40.04	48.8	42.17	41.79	41.97
Triphone + $\Delta$ + $\Delta$	33.85	39.91	36.47	36.98	36.77
+ MLLT + LDA	26.52	30.20	27.24	27.71	26.80
+ SAT(fMLLR)	21.43	23.79	20.82	21.90	<b>20.60</b>

TABLE 6.6: Iban ASRs performances (WER%) after using five different pronunciation dictionaries.

We produced phoneme/grapheme sequences for Iban lexicon using the G2P systems mentioned which resulted two pronunciation lexicons called, English G2P and Grapheme, respectively. Then, we trained two Iban ASRs using the dictionaries. Table 6.6 shows our final results for Iban ASR after employing five pronunciation dictionaries. Hybrid G2P still outperforms the rest of the systems. As expected, the two new systems did not perform better but it is interesting to note that one of them, the grapheme-based system, gave similar results with Malay G2P and Iban G2P. This shows that Malay segmentation rules can be used for building Iban pronunciation dictionary. English G2P results are the worst but only  $\sim 3\%$  different (23.79% WER) from other results. As a conclusion, using phonetizers with out-of-language data for creating Iban pronunciation dictionary is also a good starting point.

## 6.5 Solving text normalization issues

Using several number of human transcribers to transcribe speech data could result human errors or inconsistencies in transcription data. We have shown in our analysis on confusion pairs in Section 6.3.1.3 that the references in our test data have normalization issues. Normally, to solve this problem, a language expert is hired to check the transcription data again. In our case, we proposed a different approach that requires collaborative work from people to verify the data.

In the past, crowdsourcing has been used for collecting speech transcripts. The strategy uses the advantage of direct access to huge amount of Internet users (workers)

who can perform tasks given by a requester. Wikipedia<sup>1</sup> is an example of a crowdsourcing system that has been receiving continues contributions (ideas, facts) from volunteers. There are also systems that pay workers to complete simple task such as Amazon's Mechanical Turk (MTurk<sup>2</sup>) or solve technical problems that require some level of programming skills like TOPCODER<sup>3</sup>.

It might be less useful to hire professional transcriber since there have been studies which showed transcriptions by non-professional and professional have no significant impact to ASR performance. Novotney and Callison-Burch [2010] demonstrated that non-professional annotation through this application can meet professional quality. The impact of using transcripts that were transcribed by professional and Turkers to English ASR performance was not significant (about 1% difference). The same trend was found for Korean after comparing the performance of two ASR systems. A study for evaluating transcriptions quality for African languages (Swahili and Amharic) found that the ASR results were similar for system with transcription data that was collected through MTurk and system with reference transcription [Gelas et al., 2011].

Survey data collection is a standard method used in empirical research for many fields such as social sciences, marketing and statistics. Researchers could perform data collection through online surveys which is less expensive than conducting door-to-door interviews [Wright, 2005]. When conduction online, this is also considered crowsourcing where the owner of the survey collect responses from Internet users voluntarily or with payment, the former is often found, for example, opinion polls in government or news websites.

To create and publish online surveys, one may need to search a suitable tool to use. Simple systems such as Limesurvey<sup>4</sup> is an example of a tool for creating surveys. It is an open source tool, which offers unlimited number of surveys, questions and participants to a survey. Unlike MTurk, Limesurvey is not a platform for users (respondents) to earn money but to provide opinions on questions given to them.

Why not use MTurk? Hiring Turkers to complete task does not guarantee good quality transcription data. Gelas et al. [2011] have identified several problems in their

---

<sup>1</sup><http://www.wikipedia.org>

<sup>2</sup><https://www.mturk.com/mturk/welcome>

<sup>3</sup><http://www.topcoder.com/>

<sup>4</sup><https://www.limesurvey.org/>



transcription data in African languages. The data was collected through MTurk by hiring native speakers as Turkers. After analyzing the quality of the data produced by Turkers, they found incomplete transcriptions, different writing system used and copy-paste instructions in the data. More issues regarding quality of work produce by Turkers are also discussed in [Adda et al., 2014].

Hence, we proposed an online survey for correcting our transcription data to investigate data normalization through *volunteerism*. The idea was to create a survey with simple tasks to complete and publish it on social media sites that were targeted to Iban speakers for attracting people to participate. We used a multiple choice type of survey where respondents were required to choose words to correct the target words in sentences. To prepare the survey content, we conducted the following steps: (1) choose candidates for the target words; (2) select sentences for the survey; (3) create and publish the survey and (4) collect responses and update reference data.

### 6.5.1 Obtaining target words and candidates

To obtain a list of target words to be corrected, first we measured the edit distance between two words for all words in the Iban lexicon. This was done by employing Levenshtein method. Then, we removed word pairs with distance more than one ( $> 1$ ) to limit the amount of words for our experiment. After that, we sorted the list to group the candidates to target words as shown in Figure 6.2, which shows an example of seven target words and their candidates. Subsequently, we filtered the list again to keep only target words that were seen in the test data.

The candidates were options for respondents to choose to replace the target words. However, we found that the number of candidates for some target words were too many (more than 5). Thus, we selected top 5 most frequently used words in our text data as candidates.

In total, we had 658 target words for normalization. From the data, we found that almost 50% of the target words had only one candidate, as presented in the bar graph in Figure 6.3. Therefore, we predicted that respondents would be able to complete the survey in a reasonable time. Finally, we selected sentences from the test data, where each sentence contains a target word.

```

tinggal ['tungal']
wan ['tan', 'wang']
lagu ['lalu', 'agu']
long ['song']
lagi ['bagi', 'magi', 'lai', 'lagu', 'pagi', 'agi']
kadiatur ['kediatur']
cukai ['bukai', 'ukai']

```

FIGURE 6.2: Example of target word and [candidates]

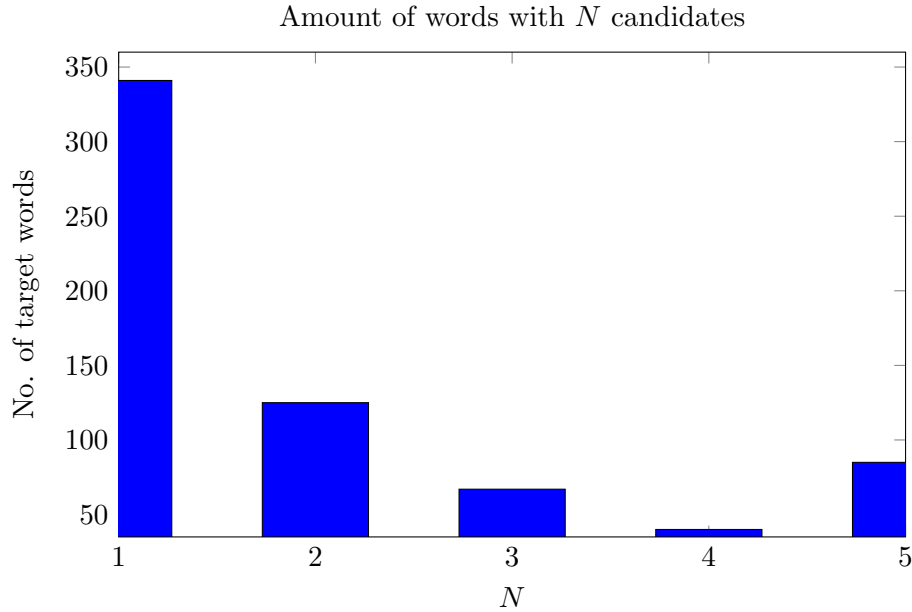


FIGURE 6.3: Bar graph showing total amount of words based on number of candidates

### 6.5.2 Creating and publishing the online survey

Sentences and candidates (hereafter we call them as questions) were added into a survey template (.txt file), which we downloaded from Limesurvey. We divided the questions into four sections. Each section had less than 200 questions and each question had one sentence with the respective candidates. In the list of candidates, we also added an empty dialog box and the original target word. This was done as to allow respondents to provide a new word when all options were not true or choose the original word if it was valid.

When our survey content was ready, we published the online survey<sup>5</sup> and promoted it on several Facebook<sup>6</sup> pages such as, “IBAN PEOPLE”, “IBAN FACEBOOK” and “Iban”. The task was completed after running the survey for one month but we received

<sup>5</sup>see Appendix D for the system’s user interface

<sup>6</sup><https://www.facebook.com/>

very few contributors. Some of the participants corrected less than 40 sentences or did not complete any task at all. The outcome of this approach was a let down due to lack of participation. Despite the fact that there are thousands of members on the Facebook pages targeted for Iban speakers, we failed to attract them to participate.

### 6.5.3 Impact on ASR results

Based on the survey, only 59 original target words were replaced. Consequently, we updated the references in our test data and then used the updated test data to compare with our ASR hypotheses. Table 6.7 presents the old and new ASR results.

Reference	Grapheme	Dictionary			
		English G2P	Malay G2P	Iban G2P	Hybrid G2P
Old	21.43	23.79	20.82	21.90	<b>20.60</b>
New	20.50	22.90	19.8 0	20.9 0	<b>19.70</b>

TABLE 6.7: Old and new ASR results for speaker adapted (fMMLR) based on WER (%)

Results only show  $\sim 5\%$  relative improvement from our previous results. Nevertheless, we have solved some of the normalization issues in the transcription data. This latest version of test data will be used in the following ASR experiments.

Target word	Replaced with	Target word	Replaced with
madahkan	madahka	pelajai	pejalai
negri	nengeri	organisation	organization
pengerja	pengereja	nyangkum	nyengkaum
serta	sereta		
rakyat	rayat	pulis	polis

TABLE 6.8: Words that were replaced with options chosen by respondent

Table 6.8 indicates examples of target word and the chosen candidate which we obtained from the survey. We found words that have spelling mistakes such as *madahkan*, *nyangkum* and *pelajai*. Others were due to spelling variant or personal choice. For instance, the word *sereta* is often written as *serta* because the second *e* or the sound /ə/ is excluded in pronunciation. Moreover, *serta* is a Malay word which has similar meaning with *sereta* for Iban. Another example is *pulis* and *polis* (police), where the former is close to how it is pronounced in Iban but the latter is commonly used in

writting, and it is also used in Malay. On the contrary, the word *rakyat* is a Malay word but Iban writes *rayat* although the pronunciation /k/ is kept. Finally, English word "organisation" is written following American format but most Malaysian texts follow British format which is "organization".

## 6.6 Summary

In this chapter, we have demonstrated our first efforts in obtaining Iban ASR, the first system for this language. The system was built on Kaldi and we investigated several approaches for building the acoustic models and as tested several pronunciation dictionaries on ASR. Our results showed that the speaker adapted system with Hybrid G2P dictionary gave the best ASR performance (lowest WER). Nevertheless other systems that used the other four dictionaries had results close to Hybrid G2P system. This shows that using Malay or English G2P as first strategy in building Iban pronunciation dictionary for ASR is a good approach. Moreover, we showed that our grapheme-based system outperformed language dependent Iban G2P system.

Furthermore, our system combination strategies further improved the baseline results and the best combination was Hybrid G2P + Malay G2P + Iban G2P. We also performed confusion pair analysis on decoding results where normalization and morphological issues were identified in the transcription data. We found spelling variations, orthography mistakes as well as stems and affixes that were not concatenated. Thus, these problems led to our proposed solution for improving the data. The approach involved applying Levenshtein distance to select target words, candidates and sentences from test data for post-editing, creating an online survey using selected sentences and candidates as survey content, and lastly, publishing the survey to collect responses and updating references. Unfortunately, we did not receive many volunteers to help complete the task. In the future, we should focus on strategies to attract people to volunteer in providing data.



## Part III

# Cross-lingual acoustic modelling for bootstrapping ASR in very low-resource settings



## Chapter 7

# Using resources from a closely-related language to develop ASR for a very under-resourced language

### 7.1 Introduction

In the previous part on this thesis, we have succeeded in developing the first speech recognition system ever built for Iban language. We are aware that the current system for the under-resourced language has very low amount of training data, thus scientific solutions are proposed for increasing the speech recognition accuracy. In this chapter, we present approaches for cross-lingual acoustic modelling for improving the Iban ASR. Particularly, we exploit out-of-language data in our strategies to prove that using a closely-related language data as source data in the cross-lingual approach is important for improving the Iban system.

### 7.2 Related work

ASR with limited training data typically suffers low recognition accuracy. This is because current state-of-the-art systems are designed for large vocabulary continuous speech



recognition (LVCSR) systems. Furthermore, conventional acoustic modelling approaches such as Gaussian Mixture Modelling requires large amount of data for training. Past studies have shown that cross-lingual or multilingual acoustic models can help to boost the performance of language-specific systems by providing universal phone units that cover several spoken languages (e.g. [Schultz and Waibel, 1997], [Schultz and Waibel, 1998], [Lin et al., 2009], [Wang et al., 2002]). With these universal phone units available, all speech data can be pooled for acoustic model training in hope to replace the missing sounds from target language, if training data is not available. However, mapping source phone units to target units can be tricky, especially for very under-resourced languages that are poorly described.

Lately, Subspace Gaussian Mixture Model (SGMM) has shown to be very promising for ASR in limited training conditions, as shown in [Lu et al., 2014] and [Imseng et al., 2014]. Both studies presented cross-lingual and multilingual work using SGMM for improving ASR with very limited training data. The authors carried out the approach by employing UBM trained on source language data, either monolingual or multilingual data, in SGMM training of their target language. Applying this technique improved the ASR performance of their low-resource system.

In this improved technique for HMM/GMM system, the acoustic units in SGMM are all derived from a common GMM called Universal Background Model (UBM). This UBM, which somehow represents the acoustic space of the training data, can be estimated on large amount of untranscribed data from one or several languages. The globally shared parameters do not need the knowledge about the phone units used in the source language(s). Without this constraint of source-target mapping of acoustic units, UBM can be easily used in cross-lingual or multilingual settings. Furthermore, UBM that is trained on data that has many speakers also help to increase speaker diversity in the acoustic space.

In the meantime, Deep Neural Networks (DNNs) have been increasingly employed for building efficient ASR systems in the very recent years. HMM/DNN hybrid systems clearly outperform HMM/(S)GMM systems for many ASR tasks [Hinton et al., 2012a] which include dealing with low-resource systems ([Miao and Metze, 2013], [Vu et al., 2014], [Huang et al., 2013]). Several studies have shown that multilingual DNNs can be achieved by utilizing multilingual data for conducting unsupervised RBM pretraining

[Swietojanski et al., 2013] or training the whole network ([Huang et al., 2013], [Vu et al., 2014], [Heigold et al., 2013]).

Optimizing hidden layers of deep neural network (DNN) can be done through pretraining using Restricted Boltzmann Machines (RBM) [Hinton, 2010]. The generative pretraining strategy builds stacks of RBMs corresponding to the number of desired hidden layers and provides better starting point (weights) for DNN fine-tuning through backpropagation algorithm. Pretraining a DNN can be conducted in a unsupervised manner because it does not involve specific knowledge (labels, phone set) of a target language<sup>1</sup>. However, it has been shown by Swietojanski et al. [2013] that using untranscribed data for RBM pretraining as a multilingual strategy has little effect on improving monolingual ASR performance.

The *transfer learning* [Heigold et al., 2013] approach has shown large recognition accuracy improvements. The technique involves removing the top layer of a multilingual DNN and fine-tuning the hidden layers to a specific language. The hidden layers are considered language independent thus are transferable across systems. The softmax layer is the only part of the DNN structure that is sensitive to the target language. It is added on top of the hidden layers during fine-tuning and its output corresponds to the HMM states of the target language. This cross-lingual approach has been applied in several recent studies for improving low-resource systems. Such studies can be found in [Miao and Metze, 2013], [Vu et al., 2014] and [Huang et al., 2013].

### 7.2.1 Motivation

Most of the cross-lingual works cited above used one or several "source" language to help the design of "target" language ASR. However, the choice of the source language(s) was not always legitimate while we believe that the use of a closely-related (well resourced) language is the best option in most cases. Thus, we attempt to answer to the following question: is there a clear benefit when using resources from closely-related language in building acoustic models for very under-resourced language? We evaluate this on ASR system for Iban and we systematically compare the use of resources from a closely-related language (Malay) with resources from a non closely-related language (English).

---

<sup>1</sup>In that sense, RBM pretraining (for DNN) and UBM training (for SGMM) are both unsupervised methods to get an initial representation of the acoustic space before modelling the speech units

## 7.3 Cross-lingual approaches for building Iban acoustic models

### 7.3.1 Training ASR (SGMM, DNN) on Iban data only

We trained acoustic models on two conditions of (transcribed) training speech: 1h and 7h Iban data. The 1h data was randomly picked from the 7h training data. We built triphone acoustic model (39 MFCC with deltas and deltas deltas) on the respective training data for obtaining GMM baseline systems. The 1h system had 664 context-dependent states and 5K Gaussians, while the 7h system had 2998 context-dependent states and 40K Gaussians. Note that we used the Hybrid G2P pronunciation dictionary and the Iban language model, both of these we have acquired and applied in the experiments described in the previous chapter.

To build SGMM models, we used the same decision trees as the ones used in the GMM systems. We did not include speaker adaptive training at this point, because we want to observe only the cross-lingual effect in the experiments described in this section. The UBM was trained on 7h of untranscribed (training) data for initializing SGMMs of both systems, using 600 UBM Gaussians. Then, the SGMMs were derived from this UBM and the phonetic subspace dimension was set to 40. The acoustic model in the 1h system had 805 substates and the 7h system had 10K substates. For acquiring DNNs, we trained the network using state-level minimum Bayes risk [Kingsbury, 2009] (sMBR) and the network has seven layers, each of the six hidden layers has 1024 hidden units. The network was trained from 11 consecutive frames (5 preceding and 5 following frames) of the same MFCCs as in the GMM systems. Furthermore, same HMM states were used as targets of the DNN. The initial weights for the network were obtained using Restricted Boltzmann Machines (RBMs) that resulted in a deep belief network with 6 stacks of RBMs. Then, fine tuning was done using Stochastic Gradient Descent with per-utterance updates, and learning rate 0.00001 kept constant for 4 epochs. To run our DNN experiments, we utilized a GPU machine and CUDA toolkit to speed up the computations.

The ASR results for monolingual SGMM and DNN are presented in Table 7.1. Both modelling techniques provided better ASR performance than GMM for the two systems with different amount of training data. For the 1h system, with SGMM and DNN

Training approach	Amount of training data	
	1-hour	7-hour
GMM	40.3	36.0
SGMM	37.8	18.9
DNN	26.9	18.4
# of states	661	2998

TABLE 7.1: Monolingual ASR results (WER%) on our Iban (1h) test set - no speaker adaptation at this stage.

we achieved 2.5% and 13.4% absolute improvements on the WER, respectively. On the other hand, both approaches resulted almost equal performance in the 7h system. We achieved 17.1% and 17.6% reductions in WER for the SGMM and DNN systems, respectively.

### 7.3.2 Out-of-language resources for Iban ASR

Our research question is whether using data from Malay in cross-lingual acoustic modelling for Iban can improve the performance of the low-resource ASR. To answer this question, we employed two out-of-language databases in our experiments. We used Malay speech corpus as data from a closely-related language and English corpus for data from a *non* closely-related language. Given these two scenarios, we hope to evaluate Malay and English performances for cross-lingual SGMM and language-specific DNN. We explain in brief regarding the databases used for this study. Full details have been described in Chapter 4 (see page 62).

#### 7.3.2.1 Closely-related language (Malay) data

The MASS corpus [Tan et al., 2009] contains Malay speech of about 140 hours. The corpus has training data of 120h and the remaining 20h data is dedicated for testing. The database has been used for speech recognition research tasks (see [Tan and Rainavo-Malançon, 2009], [Xiao et al., 2010]). We have also recently used this database for conducting a study on Malay speech spoken by speakers from different origin [Juan et al., 2012] (see abstract in Appendix A).

### 7.3.2.2 Non closely-related language (English) data

For English data, we used the first release of TED-LIUM [Rousseau et al., 2012] corpus. The corpus contains speeches excerpted from video talks of the TED website. For this study we used 118h transcribed speech for training acoustic models.

Cross-lingual SGMM	Amount of training data	
	1h	7h
Using monolingual UBM:		
a. Malay	28.3	19.4
b. English	30.8	19.2
Using multilingual UBM:		
a. Iban + Malay	27.2	19.6
b. Iban + English	29.8	19.2
c. English + Malay	29.4	19.1
d. Iban + Malay + English	28.3	19.2
# of substates	805	10K

TABLE 7.2: Results of cross-lingual SGMM (WER %) in 1h and 7h systems on 1h test data.

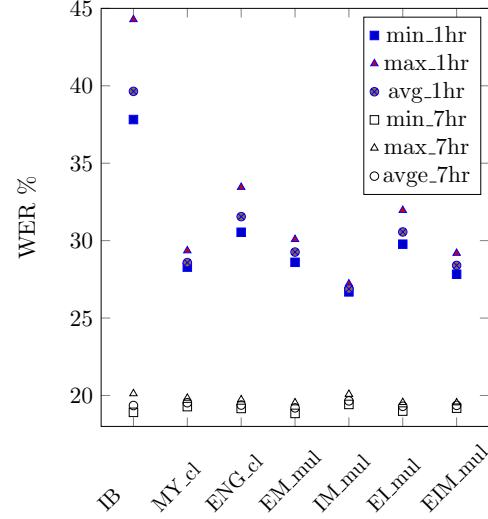


FIGURE 7.1: Minimum, maximum and average results for 1h and 7h Iban ASR using cross-lingual SGMM with different number of substates applied. Test data (1h)

### 7.3.3 Cross-lingual SGMM using monolingual and multilingual data

To obtain cross-lingual SGMM, first we trained monolingual UBMs on Malay and English data. We used training corpus available in the MASS and TED-LIUM corpora to build the models and the respective datasets contain 120h Malay and 118h English speeches. We set 600 Gaussians in the UBM and phonetic subspace dimension to 40. Then, each UBM was used in SGMM training for Iban. We applied the same training conditions used in the monolingual experiment for obtaining cross-lingual SGMM. Apart from using monolingual (English or Malay) UBM, we employed multilingual UBM in the cross-lingual method. We proposed four multilingual UBMs built using data from Iban, Malay and English. Each UBM was trained simultaneously on multilingual data that contain two or three language data. The performance of these different data combinations will help us to observe which language has better impact on Iban ASR.

The combinations of data are stated in Table 7.2 along with the results for SGMM systems with 805 and 10K substates applied.

As shown in the table, we found that the cross-lingual approaches improved our baseline results for the very low-resource setting only (1h training data) where we gained 7% to 11% absolute WER improvement against the SGMM baseline (37.8%). Between English and Malay, the latter gave greater impact to the Iban system. For example, employing Malay UBM yielded 9.5% absolute WER improvement while using English UBM resulted 7%. In addition, using UBM trained on multilingual data that had English speech did not give good results while Iban + Malay combination was the best.

Figure 7.1 shows additional results on monolingual, cross-lingual and multilingual SGMM systems for Iban. In the graph, we present the minimum, average and maximum WER values from our observations after applying different number of substates in the training. For the 1h system evaluation, we used substate values ranging from 800 to 8700, while for the 7h system, we employed 4200 to 56000 substates.

The results in the graph also showed that the cross-lingual approach based on monolingual/multilingual data was effective only for ASR with 1h training condition. In this setting, it is clearly shown that Malay was more beneficial than English for the Iban system. We observed (not shown in the graph) that our results for system with 1h training data degraded when the number of substates were increased. However, for the 7h system we obtained lower WERs as we used more substates.

#### 7.3.4 Language-specific top layer for DNN

Based on the observations of cross-lingual SGMM for Iban, Malay acoustic features seem to be useful for improving Iban ASR. Thus, we hypothesize that DNN trained on Malay data is also beneficial for improving acoustic models in Iban system. For conducting cross-lingual approach in the DNN framework Figure 7.2 illustrates our process.

First, we obtained two speaker adapted DNN systems, respectively trained on Malay and English data. The DNN targets were 3K context-dependent triphone states for Malay and English, which were obtained from HMM/GMM systems. For acquiring the GMM systems, we employed 39 MFCC with deltas and deltas deltas and applied LDA, MLLT and fMLLR. We trained six-hidden-layer DNNs, which have 1024 units in each

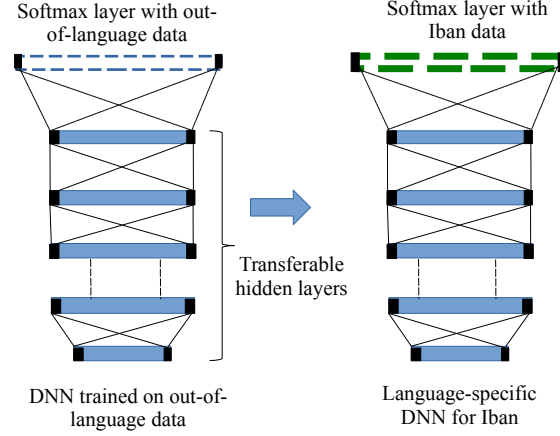


FIGURE 7.2: Process of obtaining language-specific DNN for Iban (right) using hidden layers from DNN trained on out-of-language data

hidden layer. Then, we removed the last layer of each DNN to keep only the hidden layers for cross-lingual approach.

Following this, we obtained DNN targets for Iban by acquiring a speaker adapted HMM/GMM system. We trained new Iban triphone models on new feature vectors by using the same feature transformation methods described above. During the LDA+MLLT training, one important trick is to use feature transforms acquired from the source (Malay or English) corpus (with large number of speakers). This is because merging DNNs with different feature transforms is not a good approach (for which we have observed no improvement). Finally, we built language-specific DNN for Iban by fine-tuning the hidden layers from Malay and English on 1h and 7h Iban training data. Then, the DNN systems were evaluated on the Iban test set.

DNN with language-specific top layer	Amount of train data	
	1h	7h
Hidden layers from English	19.1	15.2
Hidden layers from Malay	18.9	15.2

TABLE 7.3: WERs of cross-lingual DNNs - with speaker adaptation

The results of the DNN systems are presented in Table 7.3. Applying both speaker adaptation and language-specific top layer technique significantly improve our DNN baselines (reported in Table 7.1, second last line). For comparison, training a monolingual and speaker adapted Iban DNN lead to 15.8% WER with 7h train condition. The results also showed few language effect (English or Malay) even if for 1h training condition, the hidden layers from English were slightly less useful for Iban ASR.

## 7.4 Towards zero-shot ASR using a closely-related language

Our final work for Iban ASR is developing a zero-shot ASR for our target language using data from a closely-related language. The zero-shot system was developed through unsupervised training on Iban data. Here, we assume that we only have a language model and lexicon that are dependent to the target language. The Iban training transcripts, however, are automatically generated using a Malay acoustic model.

To perform this task we built a Malay acoustic model on Malay 120h training data, using the SGMM approach. The SGMM model was initialized by UBM with 600 Gaussians and trained using the same decision trees in HMM/GMM system for the experiment described in the previous section. Furthermore, we applied speaker adaptive training in the SGMM. Then, the Malay SGMM acoustic model and the Iban language model as well as lexicon were employed for decoding the Iban training data. We compared the ASR outputs with the references in the Iban training data and obtained 27.3% WER. Subsequently, we employed the hypothesized transcripts for training Iban ASR systems using GMM, SGMM and DNN.

The Iban GMM system was trained on 7h data for obtaining a triphone acoustic model with LDA and MLLT, as well as speaker adaptive training applied. Then, we acquired SGMM and DNN models based on the HMM targets in the GMM system using the same parameter settings applied for monolingual experiment described in Section 7.3.1. All three systems were later evaluated on Iban test data.

ASR system (7h)	Training approach		
	GMM	SGMM	DNN
Supervised (no spkr adapt.)	36.0	18.9	18.4
Supervised (with spkr adapt.)	19.7	16.6	15.8
Unsupervised (with spkr adapt.)	21.4	18.6	18.9

TABLE 7.4: Performance of Iban ASR with Supervised and Unsupervised transcripts  
- Training on 7h of Iban speech

Table 7.4 presents a performance comparison of supervised and zero-shot (unsupervised) ASR systems. Note that we did not perform cross-lingual acoustic modelling in this last experiment. In the first row of results, we present the ASR results as indicated in Table 7.1. The next line shows the performance of supervised systems



with speaker adaptive training applied. The SGMM and DNN acoustic models of the supervised systems were built on the GMM system which yielded the best performance in the pronunciation dictionary evaluation (last row and column of Table 6.7).

In general, the performance of the unsupervised system was quite close to the performance of the supervised system. We observed only 2% to 3% WER difference between each supervised and unsupervised system. The small difference suggest that the zero-shot system and the language dependent system were able to produce almost the same transcripts. However, since no difference between the performance of SGMM and DNN is observed for zero-shot ASR, we hypothesize that DNN training might be less robust to the use of noisy transcripts.

## 7.5 Summary

In this chapter, we have demonstrated approaches for cross-lingual acoustic modelling for improving performance of Iban ASR in very low-resource settings. We have applied two strategies that employed out-of-language data to build our systems: (1) cross-lingual acoustic model training based on SGMM and DNN, and (2) unsupervised training of a zero-shot ASR. The first approach significantly helped us to improve our monolingual systems. For cross-lingual SGMM, we gained more WER improvements when Malay was employed and the effect was only pronounced for systems with very low-resource setting (1h). This outcome was also observed in [Lu et al., 2014] and [Imseng et al., 2014] work for cross-lingual SGMM. Then, fine tuning hidden layers from Malay DNN also improved our DNN baselines for Iban, particularly for 1h training condition. Last but not least, the second (unsupervised) approach provided a system with promising performance. However, it seems that using automatic transcripts deserves DNN training since we observe no difference in the results of the unsupervised SGMM and DNN systems.

## Chapter 8

# Fine merging of native and non-native speech for low-resource accented ASR

### 8.1 Introduction

In this chapter, we describe two approaches for fine merging of native and non-native data for improving low-resource non-native ASR. We introduce a novel strategy for dealing with unbalanced corpora in multilingual/multi-accent SGMM, where in our case, we have a large amount of native data and a small amount of non-native data for training. The second approach involves obtaining accent-specific DNN in low-resource setting. We applied the above methods on Malaysian English (non-native) ASR system and used data from TED-LIUM corpus as source for obtaining multi-accent and accent-specific acoustic models.

### 8.2 Related work

Performance of non-native automatic speech recognition (ASR) is poor when few (or no) non-native speech is available for training / adaptation. Many approaches have been suggested for handling accented-speech in ASR, such as acoustic model merging ([Morgan, 2004], [Bouselmi et al., 2005], [Tan and Besacier, 2007], [Tan et al., 2014]),

applying maximum likelihood linear regression (MLLR) for adapting models to each non-native speaker [Huang et al., 2000], or adapting lexicon ([Arslan and Hansen, 1996], [Goronzy, 2002]).

Recently, cross-lingual acoustic modelling based on SGMM and DNN framework have been applied for improving non-native ASR systems for instance, in [Mohan et al., 2012], [Tong et al., 2014] for SGMM, and in [Huang et al., 2014], [Chen and Cheng, 2014] for DNN. Both studies on SGMM applied similar cross-lingual approach where the authors adapted large amount of native speech data to non-native (low-resource) systems through training a UBM on both non-native and native data, simultaneously for obtaining multi-accent acoustic models. In the case of cross-lingual DNN, the above mentioned studies proposed multi-accent DNN acoustic model with an accent specific softmax layer for dealing non-native speech in ASR systems. A full network was trained on large amount of native data and then, the last layer of the network was removed thus left only the hidden layers. These layers were subsequently fine-tuned to non-native speech data. All of these investigations required training multilingual data simultaneously for obtaining a language independent UBM or hidden layers.

However, is pooling data for training a UBM an optimal approach? In the recent investigation, unbalanced corpora (large amount of native data and small amount of non-native data) were used for obtaining a multi-accent SGMM. Thus, can we *finely* merge a large amount of native speech with a small quantity of non-native data for achieving an optimized multi-accent SGMM? Also, can the accent-specific DNN be applied in low-resource non-native ASR systems?

We try to respond to these questions using both SGMM (less efficient than DNNs but more compact for embedded applications) and DNN (state-of-the-art) frameworks. We apply our methods to Malaysian English ASR, where a large amount of English<sup>1</sup> data is available (TED-LIUM corpus), while only 2h of non-native speech (Malaysian English corpus) is available. More precisely, we propose one strategy for each framework: (1) language weighting for multi-accent SGMMs and (2) accent-specific top layer for DNN. The first strategy is novel and involves manipulating the number of Gaussians of each native / non-native model for (multi-accent) UBM merging. In the second approach,

---

<sup>1</sup>We are aware that TED-LIUM is not a truly native English corpus (non-native speakers of multiple origins) but we consider here that the corpus permit to build an efficient system to decode native English ASR. Thus, in the next experiments, we address it “excessively” a native corpus

we build accent-specific DNN similarly to last year’s work of Huang et al. [2014] but we make it work for a very low-resource setting and with speaker adaptation on top of it.

### 8.3 Experimental Setup

This section reports non-native and native speech databases used in our investigation. Furthermore, we present the baseline results for non-native ASR based on GMM, SGMM and DNN.

#### 8.3.1 Data

The non-native speech corpus contains 15h of English speech spoken by 24 Malaysians (of Malay, Chinese and Indian origin). The data were collected by Universiti Sains Malaysia for conducting research on acoustic model merging for ASR (see [Tan et al., 2014] for more details). Table 8.1 shows the amount of data used to train and evaluate the non-native ASR. We employed 2h of transcribed data for training the system and evaluate its performance on 4h of transcribed speech. For SGMM training, 9h of untranscribed data were added to the 2h of transcribed speech to build the UBM. Our system used the CMU pronunciation dictionary (no non-native adaptation of the lexicon) which has more than 100k words. Furthermore, we used a trigram language model for decoding. The model was trained on news data, taken from a local English news website<sup>2</sup>. After evaluating the LM on the test transcription data, the LM perplexity is 189 while the OOV rate is 2.5%.

Train		Test
Untranscribed	Transcribed	
9h	2h	4h

TABLE 8.1: Statistics of the non-native speech data for ASR.

To obtain a baseline for native ASR, we used the first release of TED-LIUM [Rousseau et al., 2012] corpus. We employed 118h to train the system and 4h for evaluation. We also used a pronunciation dictionary which was included in the package. For decoding, a trigram language model built on TED and WMT11 (Workshop on Machine Translation 2011) data was utilized. The perplexity of the language model was 220 after we evaluated on the test data.

<sup>2</sup><http://www.thestar.com.my/>

### 8.3.2 Baseline systems

For the non-native ASR system, we trained a triphone acoustic model (39 MFCC with deltas and deltas deltas) using 776 states and 10K Gaussians. Then, we trained SGMM using the same decision trees as in the previous system. The SGMM was derived from a UBM with 500 Gaussians and phonetic subspace dimension was  $S = 40$ . The UBM was trained on 11h data. We built a DNN based on state-level minimum Bayes risk [Kingsbury, 2009] (sMBR) and the network had 7 layers, each of the 6 hidden layers had 1024 hidden units. The network was trained from 11 consecutive frames (5 preceding and 5 following frames) of the same MFCCs as in the GMM system. Besides that, the same HMM states were used as targets of the DNN. The initial weights for the network were obtained using Restricted Boltzmann Machines (RBMs) that resulted in a deep belief network with 6 stacks of RBMs. Fine tuning was done using Stochastic Gradient Descent with per-utterance updates, and learning rate 0.00001 which was kept constant for 4 epochs.

For the native ASR system, we built a triphone acoustic model with 3304 states and 40K Gaussians. Subsequently, we built SGMM system using the same decision trees and 500 UBM Gaussians. Lastly, we trained a DNN with 7 layers using the same setting for building non-native DNN. The three systems were evaluated on native speech (TED task) and we achieved the following WER results: 30.55% for GMM, 28.05% for SGMM and 19.10% for DNN.

Acoustic Models	Non-native	Native
GMM	41.47	57.09
SGMM	40.41	45.84
DNN	32.52	40.70

TABLE 8.2: Word error rates (WER %) of ASR with non-native (2h) and native (118h) acoustic models on non-native evaluation data (4h test).

Table 8.2 presents the baseline results of systems that used non-native and native<sup>3</sup> acoustic models, evaluated on accented speech. For non-native acoustic modelling, SGMM and DNN systems outperformed the GMM system. The systems gave 3% and 22% relative improvement, respectively. Using these non-native models (trained on 2h only!) to decode non-native speech resulted lower word error rate (WER) compared to the pure native ASR systems (trained on 118h). In the following sections, we try to

<sup>3</sup>We used the acoustic model from the native baseline systems previously described

take advantage of both corpora (*large* native and *small* non-native) by merging acoustic models (or data) efficiently.

## 8.4 Language weighting for multi-accent Subspace Gaussian Mixture Models

### 8.4.1 Proposed Method

Recall that in SGMM, the system is initialized by a Universal Background Model (UBM) which is a mixture of full-covariance Gaussians. This single GMM is trained on all speech classes that are pooled together. The advantage of this model is that it can be trained on large amount of untranscribed data or multiple languages, as shown in [Lu et al., 2014] for cross-lingual SGMM in low-resource conditions. The authors showed that the SGMM global parameters are transferable between languages, especially when the parameters are trained in multilingual fashion. Thus, this gives an opportunity for low-resource systems to borrow UBM trained from other sources.

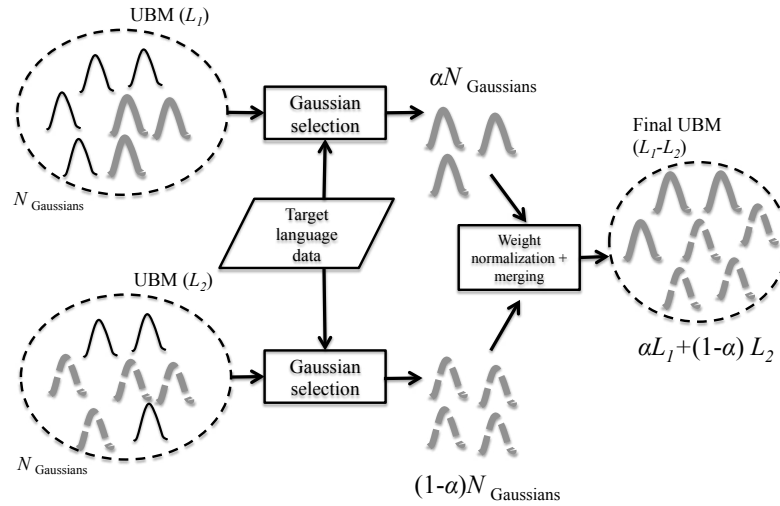


FIGURE 8.1: An Illustration of UBM merging through language weighting

Figure 8.1 illustrates the process of UBM merging through language weighting. The first step is to choose a language weight,  $\alpha$  to  $L_1$  in order to determine the number of Gaussians to be kept for merging ( $(1-\alpha)$  is given to  $L_2$ ). Intuitively, a larger  $\alpha$  should be given to the less represented source data. Then, we use data that are representative of the ASR task in order to find the top  $\alpha N$  Gaussians in  $L_1$  UBM using maximum likelihood

criterion. The same process is done for the  $L_2$  UBM but only  $(1-\alpha)N$  Gaussians are selected. The final step applies weight normalization before merging all the Gaussians in a single GMM. The final UBM should have the same number of Gaussians if both initial UBMs are the same size.

For experiments, we built a multi-accent UBM by merging native and non-native models using our language weighting strategy. To implement this, we used UBM of native speech (trained on 118h) and UBM of non-native speech (trained on 11h). Each of the UBMs has 500 Gaussians. Using the two models, we employed the language weighting approach for obtaining several multi-accent UBMs. Thereafter, these UBMs were used to estimate the parameters of non-native SGMM systems. Subsequently, we trained multi-accent SGMMs with different number of substates, ranging from 800 to 8750.

#### 8.4.2 Results

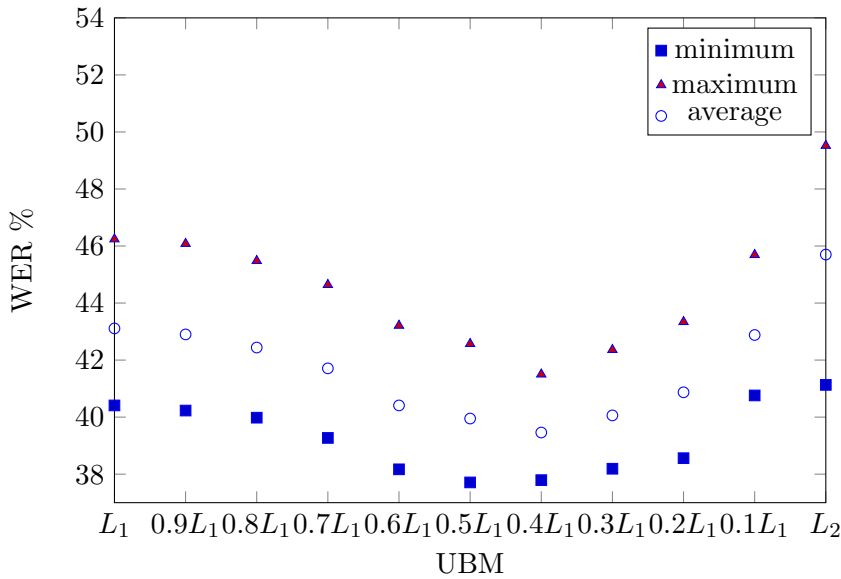


FIGURE 8.2: Min, max and average performance ( WER (%) ) of multi-accent SGMM based on language weighting strategy for non-native ASR (4h test). Note: non-native ( $L_1$ ) native ( $L_2$ ) and  $\alpha = 0.1, \dots, 0.9$ .

Our findings show that using the proposed strategy resulted significant improvement from the SGMM baseline, as presented in Figure 8.2. We reach the lowest WER when the SGMM system was obtained from a multi-accent UBM with 250 Gaussians from native and 250 Gaussians from non-native ( $\alpha = 0.5$ , WER=37.71%). This result proves that carefully controlling the contribution of two (unbalanced) data as sources

for UBM training is a way to optimize ASR performance. In this experiment, the optimal  $\alpha$  obtained tells us that non-native (Malaysian) data (in small quantity but very representative of the ASR task) and native (TED-LIUM) data (bigger corpus with speaker diversity) contribute equally to the acoustic space representation.

Furthermore, we did not gain WER improvements when the amount of substates increased. The minimum WERs shown in the figure are results for SGMMs with 800 substates. We extended our investigation to evaluate ASR performance for very compact (smaller number of Gaussians) UBM. The UBM was built with only 50 Gaussians using native/non-native data and then we applied the same language weighting strategy to obtain multi-accent UBMs.

SGMM	WER (%)
Non-native UBM500	40.41 (baseline)
Native UBM500	41.13
For $\alpha = 0.5$ ,	
a. Multi-accent UBM500	37.71
b. Multi-accent UBM50	34.24

TABLE 8.3: A summary of results from the SGMM experiments on non-native ASR (4h test). Different UBMs were employed for building SGMM with 2h of non-native training data.

The non-native system significantly improved after applying this method. Table 8.3 shows the comparison between multi-accent SGMM with UBM=500 and UBM=50. For  $\alpha = 0.5$ , the new multi-accent SGMM outperformed the one with more UBM Gaussians by 9% relative improvement on the WER. The result shows that deriving SGMM from a compact UBM gives better performance in very low-resource conditions. We also tried even smaller UBM but the WERs started to go back up (39.85% for UBM with 5 Gaussians!).

## 8.5 Accent-specific top layer for DNN

### 8.5.1 Proposed Method

We investigated two accent-specific DNNs with different training conditions for our non-native system: non speaker adapted and speaker adapted.



We began with a non speaker adapted DNN system that was already fine-tuned on native speech (last line and last column in Table 8.2). Then, we removed the softmax layer of native (source) DNN. Subsequently, a new softmax layer was added through fine-tuning the whole network on non-native (target) training data. For this condition, the targets of the DNN were obtained from the GMM baseline system for non-native.

As for the speaker adapted system, first we trained new native and non-native triphone models on new feature vectors using linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT), as well as speaker adaptive training using feature-space Maximum Likelihood Linear Regression (fMLLR). During LDA+MLLT training of non-native system, we employed feature transforms that were acquired from the native corpus. This technique was applied in our experiment for Iban ASR (see Chapter 7, page 102). Then, we trained DNN for native and later we removed the top layer of the model. Subsequently, we fine-tuned the remaining DNN layers on the non-native data.

### 8.5.2 Results

DNN with accent-specific top layer	WER (%)
a. No speaker adaptation	24.89
b. Speaker adaptation	21.48

TABLE 8.4: WERs of accent-specific DNN on non-native ASR task (4h test).

We tested the DNNs on the same non-native evaluation data (4h test). Table 8.4 presents our findings. Both results are significantly better than the pure non-native DNN baseline (last line in Table 8.2). For example, we achieved 24% and 34% relative improvement respectively over the non-native DNN baseline (32.52%). Thus, the hidden layers of the native DNN proved to be useful for improving the low-resource non-native ASR. Besides that, our approach for building DNN with speaker adaptation and accent-specific top layer provided the best result. We obtained 14% relative improvement over the accent-specific DNN without speaker adaptation.

## 8.6 Summary

We have proposed two approaches for fine (and optimal) merging of native and non-native data in order to improve accented ASR with limited training data. The first approach introduced a language weighting strategy for constructing multi-accent compact SGMM acoustic models. In this approach, we used language weights to control the number of Gaussians of each UBM involved in the merging process. Improvement of the ASR performance was observed with language weighting. The second approach involved fine-tuning the hidden layers of native DNN on non-native training data. We applied this approach for obtaining accent-specific DNN with and without speaker adaptation. For the former, we trained the DNN on HMM/GMMs that had feature transforms of the native speech data. Both DNNs outperformed the DNN baseline. Overall, the approaches used in this study resulted encouraging improvement in WER. Over the non-native baseline, we achieved relative improvement of 15% for SGMM (multi-accent UBM50) and 34% for DNN (accent-specific with speaker adaptation).



## Chapter 9

# Conclusions and Future Directions

### 9.1 Conclusions

Building automatic speech recognition system for under-resourced languages is a very challenging task. Collecting speech and text corpora as well as creating pronunciation dictionary for these languages can be expensive and time consuming process. In addition, one needs to understand some of the linguistic properties (phonology, morphology) for determining the best approaches to use. The above issues have gained increasing attention from the speech community in the past 10 years and many studies showed useful approaches for tackling resource-poor languages. Moreover, with several specific academic platforms available, interested researchers are able to share their methods and experiences related to this topic.

In this thesis, we have demonstrated several strategies that could be used for rapid development of a state-of-the-art ASR in an under-resourced language. We have achieved in obtaining the first ASR system for Iban language, one of the under-resourced languages in Malaysia using open source tools with latest techniques. Our work primarily focused on investigating Iban speech recognition and using resources from Malay - a closely-related language - for building Iban resources and improving performance of Iban ASR.

Existing studies showed that Malay and Iban are closely-related languages. We have presented a brief introduction of both languages in Chapter 4 and we showed that both are from the same language family, have similar phonological description and writing system but both are distinct languages, based on the lexicostatistics result from past research. Moreover, Iban vocabulary is highly influenced by Malay as the latter is dominantly used which resulted many Malay words adopted to Iban. In addition, there are also many Iban and Malay original words that are written similarly. Thus, our intuition to use Malay as source language for Iban ASR is strongly supported by these connections.

For studying Iban speech recognition, we obtained a few hours of news data from a local radio station. Our first challenge began from here as no speech transcripts were initially provided. To expedite the process of transcribing the data, we held a three-day workshop for native speakers to collaborate in labelling the speech corpus. In return for their effort, we taught them basic skills on using a toolkit for transcribing the speeches. After successfully acquiring the transcription data, the following task was to build an Iban pronunciation dictionary.

Creating a pronunciation dictionary from scratch is a tedious task. We faced this problem for Iban as we had very limited resource to guide us and we had more than 30K words to phonetize. We have proposed an approach that quickly produces pronunciation transcripts for our Iban lexicon. The lexicon contains list of words which were extracted from Iban online news articles. Prior to creating the dictionary, we conducted a study to measure the pronunciation distances between Malay and Iban words. We focused on words with same surface forms, which we labelled as Malay-Iban. A number of Malay-Iban words were selected and each word was transcribed in order to obtain Malay and Iban pronunciations. Our findings showed that there were many words with similar pronunciations. Hence, this gave us the motivation to use Malay resource as a basis for the Iban pronunciation dictionary. We proposed a semi-supervised approach for Iban pronunciation dictionary development, in which we employed Malay G2P to generate pronunciation transcripts for Iban words and later post-editing was applied to correct the outputs. The latter was used to train an Iban G2P. We demonstrated this work by employing *Phonetisaurus* - an open source toolkit for building G2P. Due to our analyses on outputs generated by Malay and Iban G2P systems, we phonetized our Iban lexicon using both systems with language identification applied.

Using the database and pronunciation dictionary previously acquired we developed our baseline ASR system using Kaldi, an open source speech recognition toolkit. We built HMM/GMM systems and applied transformation methods such as LDA, MLLT and speaker adaptation based on fMLLR in our efforts to improve the performance of our system. However, the monolingual ASR had limited performance due to the low amount of training data. As we have shown above that our pronunciation dictionary was built using Malay, we hypothesized that Malay acoustic data was suitable for Iban ASR.

Past studies on cross-lingual acoustic modelling were mostly focused on phone mapping or phone clustering strategies. These strategies can be tricky if one maps or clusters the phones from source and target languages wrongly hence decoding performance could degrade. Moreover, performing these methods on under-resourced languages that lack of linguistic information can be a challenge. Latest acoustic modelling methods for estimating emission probabilities such as the Subspace Gaussian Models (SGMM) and Deep Neural Networks (DNN) have shown to be promising techniques for conducting cross-lingual modelling. Without using universal phoneme sets for training acoustic models, certain model parameters from SGMM and DNN are transferable across systems. In SGMM, the UBM is language independent and used for initializing SGMM. While in DNN, the hidden layers are transferable and can be fine-tuned on different target languages. Therefore, we conducted cross-lingual experiments based on these two frameworks using out-of-language data. We showed how to train SGMM and DNN shared parameters, mainly the UBM and hidden layers, on different language data and used these parameters to initialize model training for Iban ASR. In this work, we utilized the MASS training corpus, which was used for Malay speech recognition tasks. At the same time, we contrast the performance of Malay using English data from the TEDLIUM corpus that contains speeches from TED Talks. As a result, our monolingual system was outperformed by the cross-lingual systems in terms of WER. We observed that the cross-lingual effects were evident for Iban ASR with very limited training data (1h speech). Interestingly, using Malay data did show better results than employing English in our approach thus shows the importance of closely-related language as source data.

This observation relates to our final contribution to Iban speech recognition, which is the development of a zero-shot ASR for Iban using SGMM and DNN methods.

In most cases, speech transcripts are not available for training acoustic models. We investigated the effect of building Iban ASR using automatic transcripts. The transcripts were generated by Malay ASR - trained on Malay data while using Iban pronunciation dictionary and language model. After evaluating several systems on Iban test data, we found that the WER differences between unsupervised and supervised systems were small. This finding is rather interesting as it suggests that both systems are able to produce similar outputs. Furthermore, we observed that using automatic transcripts deserves DNN training since we observe no difference in the results of the unsupervised SGMM and DNN systems.

The final work of this thesis concerned low-resource ASR with non-native speech. We demonstrated a language weighting strategy for merging non-native (limited data) and native (large amount of data) acoustic models in an attempt to find an optimal multi-accent acoustic model for the low-resource system. The SGMM modelling was used to test our approach on Malaysian English ASR. Significant improvements in WER were observed with language weighting. We obtained an optimal model after merging equal amount of UBM Gaussians from native and non-native. Using the same strategy, we developed compact acoustic models where in this case, we merged very small amount of UBM Gaussians from native and non-native models. This method further improved our results which suggests that large amount of UBM Gaussians are not needed for non-native ASR with very limited training data. Finally, we developed an accent-specific DNN by fine-tuning hidden layers which we trained on large amount of native data, to the non-native data. Our results showed that the accent-specific DNN was able to significantly outperform monolingual DNN.

## 9.2 Future Directions

In order to provide automatic speech recognition systems for under-resourced languages from Malaysia or other countries, we have presented in this thesis various approaches for rapid resource and ASR development. Particularly, we have shown how to construct a pronunciation dictionary with less manual work and improve decoding performance of low-resource systems using our cross-lingual approaches.

Our work could be further improved by addressing the following limitations of this thesis. The language models that were used in our systems were trained using  $N$ -gram approach. Lately, neural network based language models (NNLMs) such as the Recurrent NNLM (RNNLM) and Feed-Forward NNLM (FFNNLM) are increasingly used in ASR as they provide significant improvements in speech recognition performance. In [Gandhe et al., 2014], authors showed their investigation on NNLMs works for low-resource languages. Based on their experiments on language models for Tagalog, Pashto, Cantonese, Turkish and Vietnamese, NNLMs provide lower perplexity than  $N$ -gram backoff models for small amount of training data. As the training data increases, they found that interpolated models using NNLM and  $N$ -gram model are necessary for obtaining lower perplexity and better recognition accuracy. These findings are inspiring hence comparing NNLMs with our  $N$ -gram models would be an interesting study .

In our work on the zero-shot ASR, we used automatically generated transcripts which were obtained directly from Malay ASR. However, we have not considered to select best hypotheses for training the system. A more thorough analysis such as obtaining confidence scores for each hypothesis may help to decide which data should be used for training the system. Also, in the context of acoustic modelling we could also apply cross-lingual approaches to study the effects of using out-of-language data.

The language weighting strategy which we have proposed for handling multi-accent acoustic model was carried out by manually setting the weights on each model in order to find an optimal model. Based on our experiments, choosing the number of Gaussian components for merging is important and it can significantly improve performance if the right portions are used. Thus, we hope to extend this approach by considering more dynamical methods to do the weighting.

The methods which we have proposed for investigating speech recognition for Iban and non-native speech could be applied for studying other low-resource languages particularly those from Sarawak. The state currently has 47 living languages, mostly under-studied and never been exploited for natural language processing tasks. These languages may have some linguistic similarities with the languages which we have already explored, in terms of phonology, morphology, syntax or writing system if it exists. Hence, the techniques described in this thesis could be used for building several speech recognition systems for close languages.



To achieve this goal, we could incorporate emerging methods for addressing problems concerning under-resourced languages. For example, data collection strategy using mobile applications as proposed by Bird et al. [2014b] or de Vries [2011] is useful for collecting speeches from native speakers who live in far-to-reach places. We have already presented the Sarawak map in 4, which contains the distribution of speakers according to their native languages. Due to poor road access and limited electricity, researchers face challenges to reach to the villages of the native speakers for conducting data collection. Thus, mobile phones or tablettes could be useful to replace heavy laptops that are not power-efficient.

Keyword spotting (KWS) systems are based on ASR technologies and are typically used for information retrieval. Recently, KWS systems have been proposed for low-resource languages (see [Gales et al., 2014] [Titariy et al., 2014]). Researchers have proposed phonetic search [Titariy et al., 2014], identifying in-vocabulary and out-vocabulary words [Gales et al., 2014] and ranking keywords [Zhang and Glass, 2009]. Since our work concerns ASR for close languages, it would be interesting to try KWS approaches for identifying spoken terms in languages that are related.

Finally, strategies for addressing code-switching could be studied as this phenomenon commonly occurs in Malaysian conversations. Malaysian speakers tend to switch words that are from different languages. Hence, ASR systems that are built on monolingual data will suffer performance degradation due to code-switching in continuous speech. Recent studies have demonstrated methods to respond to the issue of code-switching in ASR ([Vu et al., 2012], [Adel et al., 2013] ) which we could use as motivational work.

## Appendix A

# Personal Bibliography

Paper 1: S. S. Juan, L. Besacier, and T.-P. Tan. Analysis of malay speech recognition for different speaker origins. In *Proceedings of International Conference on Asian Language Processing (IALP)*, pages 229–232. IEEE, 2012

This paper explores speech recognition performance for Malay language with multi accents from speakers of different origins or ethnicities. Accented speech imposes accuracy problem in automatic speech recognition systems. This frequently occurs to non-native speakers of a language due to insufficiency of the non-natives data in the recognizers. In this study, we investigate the mentioned problem by building a Malay model in our recognizer and test its performance for speakers of various ethnicities. Our Malay corpora consist of read speeches and texts that are collected from local newspapers in Malaysia. Speakers who contributed the speeches are of different ethnic backgrounds. We employ context dependent models by applying linear discriminant analysis for our acoustic model and a trigram based language model. Our experiments show improved results when linear discriminant analysis technique was employed in our model while our recognizer performed worst for speakers with accent that are not available in the training data.

Paper 2: S. S. Juan and L. Besacier. Fast bootstrapping of grapheme to phoneme system for under-resourced languages - application to the iban language. In *Proceedings of 4th Workshop on South and Southeast Asian Natural Language Processing 2013*,

Nagoya, Japan, October 2013

This paper deals with the fast bootstrapping of Grapheme-to-Phoneme (G2P) conversion system, which is a key module for both automatic speech recognition (ASR), and text-to-speech synthesis (TTS). The idea is to exploit language contact between a local dominant language (Malay) and a very under-resourced language (Iban - spoken in Sarawak and in several parts of the Borneo Island) for which no resource nor knowledge is really available. More precisely, a pre-existing Malay G2P is used to produce phoneme sequences of Iban words. The phonemes are then manually post-edited (corrected) by an Iban native. This resource, which has been produced in a semi-supervised fashion, is later used to train the first G2P system for Iban language. As a by-product of this methodology, the analysis of the “pronunciation distance” between Malay and Iban enlighten the phonological and orthographic relations between these two languages. The experiments conducted show that a rather efficient Iban G2P system can be obtained after only two hours of post-edition (correction) of the output of Malay G2P applied to Iban words.

Paper 3: S. S. Juan, L. Besacier, B. Lecouteux, and T.-P. Tan. Using closely-related language to build an ASR for a very under-resourced language: Iban. In IEEE, editor, *Proceedings of Oriental COCOSDA I 2014*, pages 71–76, September 2014a

This paper describes our work on automatic speech recognition system (ASR) for an under-resourced language, namely the Iban language, which is spoken in Sarawak, a Malaysian Borneo state. To begin this study, we collected 8 hours of speech data due to no resources yet for ASR concerning this language. Following the lack of resources, we employed bootstrapping techniques on a closely-related language to build the Iban system. For this case, we utilized Malay data to bootstrap the grapheme-to-phoneme system (G2P) for the target language. We also developed several G2Ps to acquire Iban pronunciation dictionaries, which were later evaluated on the Iban ASR for obtaining the best version. Subsequently, we conducted experiments on cross-lingual ASR by using subspace Gaussian Mixture Models (SGMM) where the shared parameters obtained in either monolingual or multilingual

fashion. From our observations, using out-of-language data as source language provided lower WER when Iban data is very imited.

Paper 4: S. S. Juan, L. Besacier, and S. Rossato. Semi-supervised G2P bootstrapping and its application to asr for a very under-resourced language: Iban. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*, May 2014b

This paper describes our experiments and results on using a local dominant language in Malaysia (Malay), to bootstrap automatic speech recognition (ASR) for a very under-resourced language: Iban (also spoken in Malaysia on the Borneo Island part). Resources in Iban for building a speech recognition were nonexistent. For this, we tried to take advantage of a language from the same family with several similarities. First, to deal with the pronunciation dictionary, we proposed a bootstrapping strategy to develop an Iban pronunciation lexicon from a Malay one. A hybrid version, mix of Malay and Iban pronunciations, was also built and evaluated. Following this, we experimented with three Iban ASRs; each depended on either one of the three different pronunciation dictionaries: Malay, Iban or hybrid. Our best results (WER) for Iban ASR (with different lexicon) were as follows: 20.82% (Malay G2P), 21.90% (Iban G2P) and 20.60% (Hybrid G2P). Apart from that, we applied system combination using all of the systems and obtained an improved accuracy of 19.22%.

Paper 5: S. S. Juan, L. Besacier, and S. Rossato. Construction faiblement supervisée d'un phonétiseur pour la langue iban à partir de ressources en malais. In *Proceedings of Journée d'Etude sur la Parole (JEP)*, Le Mans, France, June 2014c

Cet article décrit notre collecte de ressources pour la langue iban (parlée notamment sur l'île de Bornéo), dans l'objectif de construire un système de reconnaissance automatique de la parole pour cette langue. Nous nous sommes plus particulièrement focalisés sur une méthodologie d'amorçage du lexique phonétisé à partir d'une langue proche (le malais). Les performances des premiers systèmes de reconnaissance automatique de la parole construits pour l'iban (< 20% WER) montrent que l'utilisation d'un phonétiseur déjà disponible dans une langue proche (le malais) est une option tout à fait viable

pour amorcer le développement d'un système de RAP dans une nouvelle langue très peu dotée. Une première analyse des erreurs fait ressortir des problèmes bien connus pour les langues peu dotées : problèmes de normalisation de l'orthographe, erreurs liées à la morphologie (séparation ou non des affixes de la racine).

Paper 6: Fine Merging of Native and Non-native Speech for Low-resource Accented ASR Authors: Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux and Tien-Ping Tan.

*Submitted to the INTERSPEECH Conference 2015*

This paper presents our recent study on low-resource automatic speech recognition (ASR) system with accented speech. We propose multi-accent Subspace Gaussian Mixture Models (SGMM) and accent-specific Deep Neural Networks (DNN) for improving non-native ASR performance. In the SGMM framework, we present an original language weighting strategy to finely merge the globally shared parameters of two models based on native and non-native speech respectively. In the DNN framework, a native deep neural net is fine-tuned to non-native speech. Over the non-native baseline, we achieved relative improvement of 15% for multi-accent SGMM and 34% for accent-specific DNN with speaker adaptation.

Paper 7: Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban Authors: Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux and Mohamed Dyab.

*Submitted to the INTERSPEECH Conference 2015*

This paper presents our strategies for developing an automatic speech recognition system for Iban, an under-resourced language. We faced several challenges such as no pronunciation dictionary and lack of training material for building acoustic models. To overcome these problems, we proposed approaches which exploit resources from a closely-related language (Malay). We developed a semi-supervised method for building the pronunciation dictionary and applied cross-lingual strategies for improving acoustic models

trained with very limited training data. Both approaches displayed very encouraging results, which show that data from a closely-related language, if available, can be exploited to build ASR for a new language. In the final part of the paper, we present a zero-shot ASR using Malay resources that can be used as an alternative method for transcribing Iban speech.



## Appendix B

# Reference to International Phonetic Alphabets



## THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

## CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

## CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ɘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ɟ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɡ Velar	kʼ Velar
ǁ Alveolar lateral	ɠ Uvular	sʼ Alveolar fricative

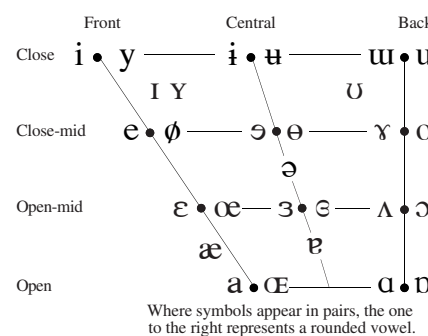
## OTHER SYMBOLS

ɱ Voiceless labial-velar fricative	ɕ ʑ Alveolo-palatal fricatives
ʋ Voiced labial-velar approximant	ɺ Voiced alveolar lateral flap
ɥ Voiced labial-palatal approximant	ɺ Simultaneous ʃ and x
ħ Voiceless epiglottal fricative	
ʕ Voiced epiglottal fricative	Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.
ʔ Epiglottal plosive	

## DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̥

◌ <sup>◌</sup> Voiceless	◌̥ ◌̜	◌ <sup>◌</sup> Breathy voiced	◌̤ ◌̩	◌ <sup>◌</sup> Dental	◌̪ ◌̫
◌ <sup>◌</sup> Voiced	◌̤ ◌̩	◌ <sup>◌</sup> Creaky voiced	◌̰ ◌̱	◌ <sup>◌</sup> Apical	◌̽ ◌̾
◌ <sup>◌</sup> Aspirated	◌̚ ◌̜	◌ <sup>◌</sup> Linguolabial	◌̼ ◌̽	◌ <sup>◌</sup> Laminal	◌̻ ◌̼
◌ <sup>◌</sup> More rounded	◌̙	◌ <sup>◌</sup> Labialized	◌̙ ◌̚	◌ <sup>◌</sup> Nasalized	◌̃
◌ <sup>◌</sup> Less rounded	◌̙	◌ <sup>◌</sup> Palatalized	◌̟ ◌̠	◌ <sup>◌</sup> Nasal release	◌̚
◌ <sup>◌</sup> Advanced	◌̟	◌ <sup>◌</sup> Velarized	◌̙ ◌̚	◌ <sup>◌</sup> Lateral release	◌̚
◌ <sup>◌</sup> Retracted	◌̠	◌ <sup>◌</sup> Pharyngealized	◌̙ ◌̚	◌ <sup>◌</sup> No audible release	◌̚
◌ <sup>◌</sup> Centralized	◌̠	◌ <sup>◌</sup> Velarized or pharyngealized	◌̙		
◌ <sup>◌</sup> Mid-centralized	◌̠	◌ <sup>◌</sup> Raised	◌̙ (ɹ̥ = voiced alveolar fricative)		
◌ <sup>◌</sup> Syllabic	◌̚	◌ <sup>◌</sup> Lowered	◌̙ (β̥ = voiced bilabial approximant)		
◌ <sup>◌</sup> Non-syllabic	◌̚	◌ <sup>◌</sup> Advanced Tongue Root	◌̙		
◌ <sup>◌</sup> Rhoticity	◌̙ ◌̚	◌ <sup>◌</sup> Retracted Tongue Root	◌̙		

## VOWELS



## SUPRASEGMENTALS

ˈ	Primary stress
ˌ	Secondary stress
ː	Long
ˑ	Half-long
◌̥	Extra-short
◌̚	Minor (foot) group
◌̚	Major (intonation) group
◌̚	Syllable break
◌̚	Linking (absence of a break)

## TONES AND WORD ACCENTS LEVEL CONTOUR

é or ˥	Extra high	ě or ˩	Rising
é	High	ě	Falling
ē	Mid	ē	High rising
è	Low	è	Low rising
è	Extra low	è	Rising-falling
↓	Downstep	↗	Global rise
↑	Upstep	↘	Global fall

FIGURE B.1: The International Phonetic Alphabets chart [IPA]

## Appendix C

# Converting Iban words using P2P system

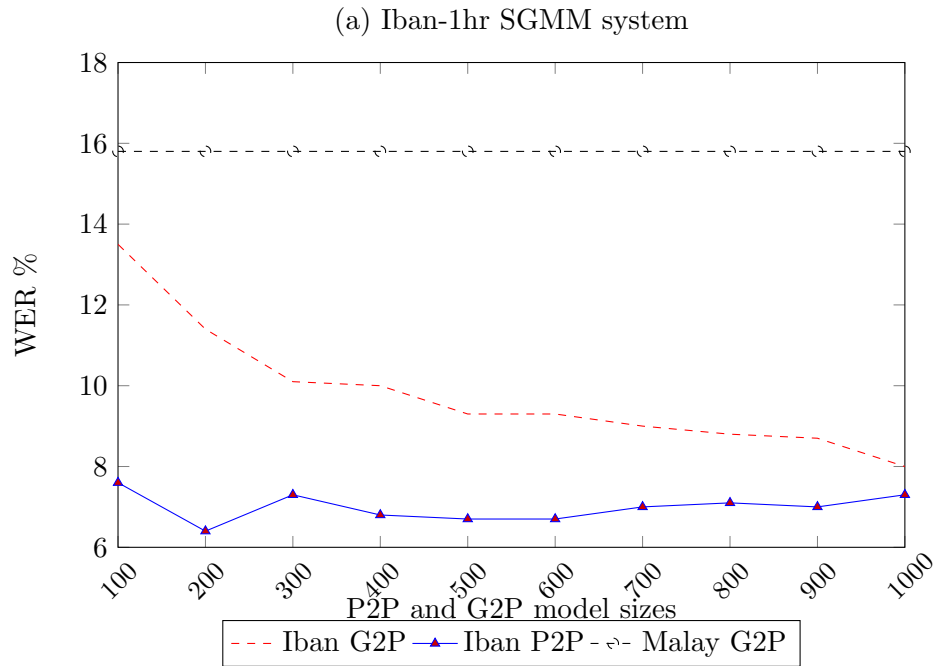


FIGURE C.1: Phoneme error rates (PER%) for 500 pure Iban words (500I)

Apart from using G2P, we used phoneme-to-phoneme (P2P) to perform phonetization tasks. To build our P2Ps, we used Malay G2P outputs and post-edited version from Iban G2P as training data. The data was divided into 10 sets with equal amount of pronunciation pairs in each set. Subsequently, we built 10 systems using different training data sizes (add one portion to the training corpus after each

model developed). We evaluated the systems on 500 pure Iban words. Additionally, we prepared 10 Iban G2P systems for comparing P2P performance.

As presented in Figure C.1, the P2P systems (see non-dotted line) have stable performance. We obtained phoneme error rates between 6.4% to 7.6%. In fact, we did not see significant effect when adding more training data to the P2P system. On the other hand, the G2P has low performance when training data is very limited. It is also interesting to see that the Iban P2P and G2P systems have almost the same performance level when using 1K data as shown in Table C.1.

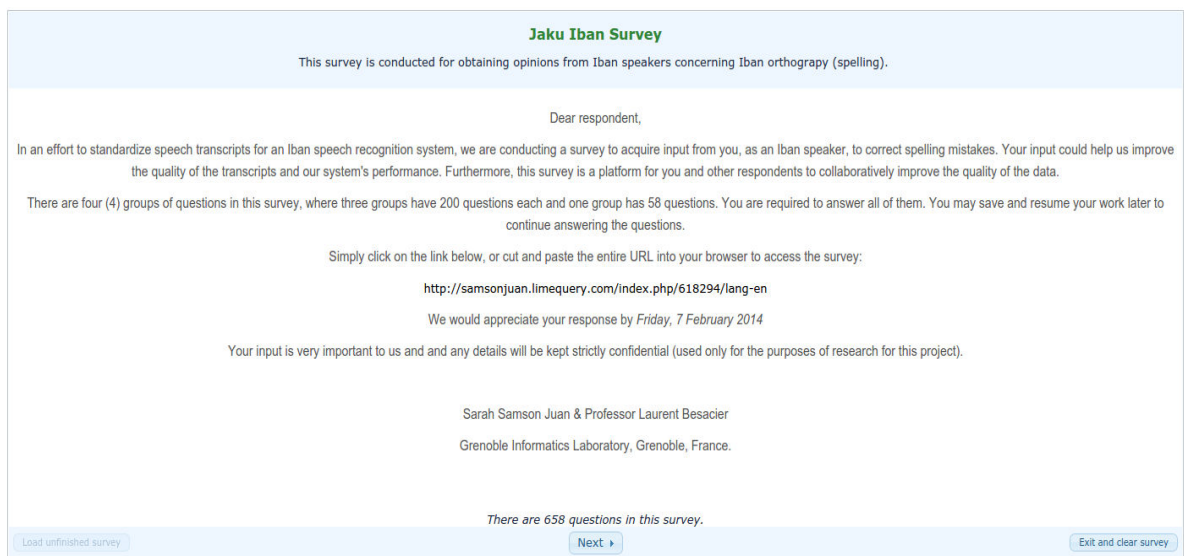
Phonetizer	Corpus	PER (%)	WER (%)	Post-edit (mins)
Malay G2P	500 <sub>IM</sub>	6.52	27.2	30
	500 <sub>I</sub>	15.8	56.0	42
Iban G2P	500 <sub>IM</sub>	13.6	44.2	45
	500 <sub>I</sub>	8.2	31.8	32
Iban P2P	500 <sub>IM</sub>	16.6	53.5	-
	500 <sub>I</sub>	7.3	31.9	-

Note: *IM* for common Malay-Iban words and *I* for pure Iban words

TABLE C.1: G2P and P2P performances for an Iban phonetization task

## Appendix D

# Online Survey System for Data Normalization



The screenshot shows the welcome page of the 'Jaku Iban Survey'. The page has a light blue header with the title 'Jaku Iban Survey' in green. Below the header, the text explains the purpose of the survey: to obtain opinions from Iban speakers on Iban orthography (spelling). It addresses the respondent, explaining that the survey aims to improve the quality of speech transcripts for an Iban speech recognition system. It mentions that there are four groups of questions, with three groups having 200 questions each and one group having 58 questions. It provides a URL to access the survey: <http://samsonjuan.limequery.com/index.php/618294/lang-en>. It also states that the response is appreciated by Friday, 7 February 2014, and that the input is confidential. The page is signed by Sarah Samson Juan & Professor Laurent Besacier from the Grenoble Informatics Laboratory, Grenoble, France. At the bottom, it states 'There are 658 questions in this survey.' and includes three buttons: 'Load unfinished survey', 'Next >', and 'Exit and clear survey'.

**Jaku Iban Survey**

This survey is conducted for obtaining opinions from Iban speakers concerning Iban orthography (spelling).

Dear respondent,

In an effort to standardize speech transcripts for an Iban speech recognition system, we are conducting a survey to acquire input from you, as an Iban speaker, to correct spelling mistakes. Your input could help us improve the quality of the transcripts and our system's performance. Furthermore, this survey is a platform for you and other respondents to collaboratively improve the quality of the data.

There are four (4) groups of questions in this survey, where three groups have 200 questions each and one group has 58 questions. You are required to answer all of them. You may save and resume your work later to continue answering the questions.

Simply click on the link below, or cut and paste the entire URL into your browser to access the survey:

<http://samsonjuan.limequery.com/index.php/618294/lang-en>

We would appreciate your response by *Friday, 7 February 2014*

Your input is very important to us and any details will be kept strictly confidential (used only for the purposes of research for this project).

Sarah Samson Juan & Professor Laurent Besacier  
Grenoble Informatics Laboratory, Grenoble, France.

*There are 658 questions in this survey.*

[Load unfinished survey](#) [Next >](#) [Exit and clear survey](#)

FIGURE D.1: The welcome page of the survey system

**Jaku Iban Survey**

This survey is conducted for obtaining opinions from Iban speakers concerning Iban orthography (spelling).

0%  100%

**Group 1**

You are now answering questions from Group 1.  
Please read this instruction carefully before answering the questions.  
There are 200 sentences taken from a speech transcript. You are required to choose the correct word to replace the word in CAPITAL letters in each sentence. For example,  
Q : dini nuan SEKULA  
Then, choose an option from the drop down list.  
sekula  
sekul  
sekla  
Other :  
If you choose Other, please write your answer in the empty box.  
After you have completed all questions in this group, click Next.

**Q1** kira satu poin empat juta iku orang ti gawa perintah nganti Jaku PADAH ari menteri besar datuk seri najib tun razak pasal skim gaji baru orang ti gawa perintah sbpa pegila  
Choose one of the following answers

Please choose...

**Q2** agih belanja kediberi dulu ari tu nya dikena sida bejalai ke projek kuasa karan ari PANCHAR mata panas tauka solar ti dipejalai ka ba rumah panjai di long lamai bakalalan enggau di larapan  
Choose one of the following answers

Please choose...

**Q3** Iya mansutka Jaku tu udah Iya ti nyuaka KUNCI lapanbelas pintu rumah pbr ba kampung baya sipat long tepah urun Iyanya sebengkah palan pengentap pendiau bansa penan di pelilih menua belaga  
Choose one of the following answers

Please choose...

**Q4** pehin seri talb ke lalu nyadi presiden parti pesaka bumiputera bersatu pbb mansutka Jaku tu lebuah Iya ti nyimpul JURAI nya ba aum konsil nengeri  
Choose one of the following answers

Please choose...

**Q5** nambahka nya kandang endur nya deka dikena alai mandangka main asal rambau MUSIM pengerami nyengkaum taun baru cina gawal tauka lebuah christmas  
Choose one of the following answers

Please choose...

**Q6** apin lama tu ke udah menteri ba opis menteri besal tan sri noor mohamed yaakub minta mayuh agi bala graduate BUMIPUTRA ngambi accounting program  
Choose one of the following answers

Please choose...

FIGURE D.2: Questions in the survey. User will need to scroll down to see more questions.

**Q650** api nya dipelabaka belabuh nyeraral ba ringkat atas rumah tu lalu ngeramptka BARUH  
Choose one of the following answers

Please choose...

**Q651** sida enda patut mai anak sida ngagai tempat ti sekut sereta nyaga pemeresi anak sida ari ti ngena sabun pemeresi JARI  
Choose one of the following answers

Please choose...

**Q652** taja pia datuk seri doktor rais nadai madahka atur enggau PEMANJAI awak ke deka diberi ngagai parti-parti politik  
Choose one of the following answers

Please choose...

**Q653** bala nemiak ke baru lepas sekula tikas ba tikas sijil pelajaran malaysia spm enggau sijil kemahiran malaysia SKM tau merasa peminta sida ti nampung pelajar ba institut pelajar teknikal enggau vokasional enggau institut pematih pengelandik ke sesi taun dua ribu dua belas dua ribu tiga belas  
Choose one of the following answers

Please choose...

**Q654** Iya madahka semua orang patut nemu bangsa melayu tuk nang bangsa asal ke nguan TANAH melayu  
Choose one of the following answers

Please choose...

**Q655** sebengkah pansik ti baru udah dipejalaka ngayanka pemakal ti diatur enggau manah sereta mungkur mayuh agi utal TANAM ulih nyendiaka utal ti diguna bala nemiak mit enggau orang ke udah besal tual  
Choose one of the following answers

Please choose...

**Q656** sida tau nanya ba HOTLINE kusung tiga lapan lapan tujuh kusung enam tujuh enam tujuh helpline kusung tiga lapan lapan tujuh kusung enam tujuh tujuh general line kusung tiga lapan lapan tujuh kusung enam lima lima tauka lapan lapan tujuh kusung enam enam tujuh enam enam  
Choose one of the following answers

Please choose...

**Q657** Iya madahka parti penyakal semina pandai ngaga SEMAYA bula kena ngulihka undi  
Choose one of the following answers

Please choose...

**Q658** ngena program tu mesgid enggau surau UKAI semina nyadi palan besamblang tang bela nyadi palan pemansang urang islam  
Choose one of the following answers

Please choose...

[Resume later](#) [Previous](#) [Submit](#) [Exit and clear survey](#)

FIGURE D.3: User is allowed to continue the task later or click submit once they have completed.

# Bibliography

- International phonetic alphabet (ipa) chart. electronic. URL [www.phon.ucl.ac.uk/home/johnm/sid/IPA\\_chart\\_\(C\)2005.pdf](http://www.phon.ucl.ac.uk/home/johnm/sid/IPA_chart_(C)2005.pdf). Retrieved 20 January 2015.
- G. Adda, J. Mairani, L. Besacier, and H. Gelas. Crowdsourcing for speech: Economic, legal and ethical analysis. 2014. URL <https://hal.archives-ouvertes.fr/docs/01/06/71/10/PDF/Ethics-6.pdf>.
- H. Adel, N. T. Vu, F. Kraus, T. Schilippe, H. Li, and T. Schultz. Recurrent neural network language modeling for code switching conversational speech. In *Proceedings of ICASSP*, pages 8411–8415. IEEE, 2013.
- M. J. Arslan and J. L. Hansen. A study of the temporal features and frequency characteristics in american english foreign accent. *Journal of the Acoustic Society*, 1996.
- J. Barnett, A. Corrada, G. Gao, L. Gillik, Y. Ito, S. Lowe, L. Manganaro, and B. Peskin. Multilingual speech recognition at dragon systems. In *Proceedings of ICSLP*, pages 2191–2194, Philadelphia, 1996.
- C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: development and use of a tool for assisting speech corpora production. In *Proceedings of Speech Communication special issue on Speech Annotation and Corpus Tools*, volume 33. available at : [trans.sourceforge.net/en/publi.php](http://trans.sourceforge.net/en/publi.php), 2000.
- N. Barroso, K. L. de Ipiña, M. Graña, and A. Ezeiza. Language identification for under-resourced languages in the basque context. In *Proceedings of 6th International Conference Soft Computing Models in Industrial and Environmental Applications*, volume 87, pages 475–483, 2011.

- L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. In *Proceedings of 3rd Symposium on Inequalities*, pages 1–8, University of California, Los Angeles, 1972. Academic Press.
- V. Berment. *Méthodes pour informatiser des langues et des groups de langues peu dotées*. PhD thesis, Université Joseph Fourier, 2004.
- L. Besacier, E. Barnard, A. Karpov, and T. Schultz. Automatic speech recognition for under-resourced languages : A survey. *Speech Communication Journal*, 56:85–100, January 2014.
- J. Billa, K. Ma, J. McDonough, G. Zavaliagkos, D. R. Miller, K. N. Ross, and A. El-Jaroudi. Multilingual speech recognition: the 1996 byblos callhome system. In *Proceedings of Eurospeech*, pages 363–366, Rhodes, Greece, 1997.
- S. Bird, L. Gawne, K. Gelbart, and I. McAlister. Collecting bilingual audio in remote indigenous communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1015–1024. Dublin City University and Association for Computational Linguistics, 2014a. URL <http://aclweb.org/anthology/C14-1096>.
- S. Bird, F. R. Hanke, O. Adams, and H. Lee. Aikuma: A mobile app for collaborative language documentation. *ACL 2014*, page 1, 2014b.
- M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008. doi: 10.1016/j.specom.2008.01.002.
- P. H. Bo-June and J. Glass. Iterative language model estimation: Efficient data structure & algorithms. In *Proceedings of INTERSPEECH*, pages 841–844, Brisbane, Australia, September 2008.
- G. Bouselmi, D. Fohr, and J. P. Haton. Fully automated non-native speech recognition using confusion-based acoustic model intergration. In *Proceedings of Eurospeech*, pages 1369–1372, Lisboa, 2005.
- L. Burget, P. Schwartz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas. Multilingual

- acoustic modeling for speech recognition based on subspace gaussian mixture models. In IEEE, editor, *Proceedings of ICASSP*, pages 4334–4337, 2010.
- R. Bust. Subgrouping, circularity and extinction : some issues in austronesian comparative linguistics. In *Selected papers from the Eighth International Conference on Austronesian Linguistics*, pages 31–94, Taipei, 1999. Taiwan: Academia Sinica.
- G. Caelen-Haumont. Towards the mo piu tonal system: first results on an undocumented south-asian language. In *Proceedings of Speech Prosody*, 2012.
- G. Caelen-Haumont and S. Sam. Comparison between two models of language for the automatic phonetic labeling of an undocumented language of the south-asia: the case of mo piu. In *Proceedings of LREC*, pages 956–962, 2008.
- G. Caelen-Haumont, S. Sam, and E. Castelli. Automatic labeling and phonetic assessment for an unknown asian language: The case of the "mo piu" north vietnamese minority (early results). In *Proceedings of International Conference on Asian Language Processing (IALP)*, pages 260–263, Penang, 2011.
- Ö. Çetin, M. Plauché, and U. Nallasamy. Unsupervised adaptive speech technology for limited resource languages: A case study for tamil. In *Proceedings of ICASSP*, 2007.
- S. F. Chen. Conditional and joint models for grapheme-to-phoneme conversion. In *Proceedings of EUROSPEECH*, pages 933–936, Geneva, Switzerland, September 2003.
- X. Chen and J. Cheng. Deep neural network acoustic modeling for native and non-native mandarin speech recognition. In *Proceedings of International Symposium on Chinese Spoken Language Processing*, 2014.
- P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward. Towards a universal speech recognizer for multiple languages. In *Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, pages 591–598, St. Barbara , CA, 1997.
- CSTR. The festival speech synthesis system, 2012. URL <http://www.cstr.ed.ac.uk/projects/festival>.
- G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):30–42, 2012.



- M. Davel and E. Barnard. Bootstrapping in language resource generation. In *Proceedings of 14th Annual Symposium of the Pattern Recognition Association of South Africa*, 2003.
- M. Davel and E. Barnard. The efficient generation of pronunciation dictionaries: Human factors during bootstrapping. In *Proceedings of INTERSPEECH*, 2004.
- S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing IEEE Transactions on*, 28(4):357–366, 1980.
- N. J. de Vries. Effective automatic speech recognition data collection for under-resourced languages. Master’s thesis, North-West University, South Africa, 2011.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–21, 1977.
- C. Dugast, X. Aubert, and R. Kneser. The philips large-vocabulary recognition system for american english, french and german. In *Proceedings of Eurospeech*, pages 197–200, Madrid, 1995.
- I. Dyen, J. B. Kruskal, and P. Black. *An Indoeuropean classification: a lexicostatistical experiment*, volume iii. Transactions of the American Philosophical Society, 1992.
- J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.
- J. Ensiring, J. Umat, and R. M. Salleh, editors. *Bup Sereba Reti Jaku Iban*. The Tun Jugah Foundation, 2011.
- V. Ferdiansyah and A. Purwarianti. Indonesian automatic speech recognition using english-based acoustic model. *American Journal of Signal Processing*, 2(4):60–63, 2012.
- R. A. Fisher. *The Use of Multiple Measures in Taxonomic Problems*. 1936.
- S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Acoustics, Speech and Signal Processing IEEE Transactions on*, 34(1):52–59, 1986.
- M. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. In *Computer Science and Language*, volume 12, pages 75–98, 1998.

- M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath. Speech recognition and keyword spotting for low resource languages: Babel project research at cued. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*, Russia, May 2014.
- A. Gandhe, F. Metze, and I. Lane. Neural network language models for low resource languages. In *Proceedings of INTERSPEECH*, 2014.
- H. Gelas, S. T. Abate, L. Besacier, and F. Pellegrino. Quality assessment of crowdsourcing transcriptions for african languages. In *Proceedings of INTERSPEECH*, 2011.
- A. Ghoshal, P. Swietojanski, and S. Renals. Multilingual training of deep neural networks. In *Proceedings of ICASSP*, pages 7319–7323, 2013.
- J. Glass, G. Flammia, D. Goodine, M. P. J. Polifroni, S. Sakai, S. Seneff, and V. Zue. Multi-lingual spoken language understanding in the mit voyager system. *Speech Communication*, 17:1–18, 1995.
- V. Goel, S. Kumar, and W. Byrne. Segmental minimum bayes-risk decoding for automatic speech recognition. In *Proceedings of IEEE Transactions on Speech and Audio Processing*, 2003.
- K. E. Goh and A. M. Ahmad. Malay speech recognition using self-organizing map and multilayer perceptron. In *Proceedings of Postgraduate Annual Research Seminar*, 2005.
- C. Gollan and M. Bacchiani. Confidence scores for acoustic model adaptation. In *Proceedings of ICASSP*, pages 4289–4292, Las Vegas, Nevada, USA, April 2008.
- I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3,4):237–264, 1953.
- R. A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *Proceedings of ICASSP*, pages 661–664, 1998.
- S. Goronzy. Robust adaptation to non-native accents in automatic speech recognition. *Springer*, 2002.

- S. Hahn, P. Vozila, and M. Bisani. Comparison of graphemeto-phoneme methods on large pronunciation dictionaries and lvcsr tasks. In *Proceedings of INTERSPEECH*, 2012.
- W. Heeringa and F. de Wet. The origin of afrikaans pronunciation: a comparison to west germanic languages and dutch dialects. In *Proceedings of Conference of the Pattern Recognition Association of South Africa*, pages 159–164, 2008.
- G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. Multilingual acoustic models using distributed deep neural networks. In *Proceedings of ICASSP*, 2013.
- G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012a.
- G. E. Hinton. A practical guide to training restricted boltzmann machines. Utml tr 2010-003, Dept. Computer Science, University of Toronto, 2010.
- G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep believe nets. *Neural Computation*, 18:1527–1554, 2006.
- G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. In *Proceedings of CoRR*, 2012b.
- C. Huang, E. Chang, J. Zhou, and K.-F. Lee. Accent modeling based on pronunciation dictionary adaptation for large vocabulary mandarin speech recognition. In *Proceedings of ICLSP*, volume 2, pages 818–821, 2000.
- J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Proceedings of ICASSP*, 2013.
- Y. Huang, D. Yu, C. Liu, and Y. Gong. Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation. In *Proceedings of INTERSPEECH*, 2014.

- T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. Moreno, and M. LeBeau. Building transcribed speech corpora quickly and cheaply for many languages. In *Proceedings of INTERSPEECH*, pages 1914–1917, 2010.
- D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner. Using out-of-language data to improve under-resourced speech recognizer. *Speech Communication*, 56(0):142–151, 2014.
- F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, London, 2001.
- F. Jelinek and R. L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of Workshop on Patter Recognition in Practice*, pages 381–397, North-Holland, Amsterdam, The Netherlands, May 1980.
- S. Jiampojarn, G. Kondrak, and T. Sherif. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April 2007. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/N/N07/N07-1047>.
- S. S. Juan and L. Besacier. Fast bootstrapping of grapheme to phoneme system for under-resourced languages - application to the iban language. In *Proceedings of 4th Workshop on South and Southeast Asian Natural Language Processing 2013*, Nagoya, Japan, October 2013.
- S. S. Juan, L. Besacier, and T.-P. Tan. Analysis of malay speech recognition for different speaker origins. In *Proceedings of International Conference on Asian Language Processing (IALP)*, pages 229–232. IEEE, 2012.
- S. S. Juan, L. Besacier, B. Lecouteux, and T.-P. Tan. Using closely-related language to build an ASR for a very under-resourced language: Iban. In IEEE, editor, *Proceedings of Oriental COCOSA I 2014*, pages 71–76, September 2014a.
- S. S. Juan, L. Besacier, and S. Rossato. Semi-supervised G2P bootstrapping and its application to asr for a very under-resourced language: Iban. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*, May 2014b.

- S. S. Juan, L. Besacier, and S. Rossato. Construction faiblement supervisée d'un phonétiseur pour la langue iban à partir de ressources en malais. In *Proceedings of Journée d'Etude sur la Parole (JEP)*, Le Mans, France, June 2014c.
- S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.
- B. Kingsbury. Lattice-based optimization of sequence classification criteria for neural network acoustic modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3761–3764, April 2009.
- R. Kneser. Grapheme-to-phoneme study. Technical report, Philips Speech Processing, Germany, 2000.
- R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proceedings of ICASSP*, pages 181–184, 1995.
- L. Lamel, M. Adda-Decker, and J. L. Gauvain. Issues in large vocabulary multilingual speech recognition. In *Proceedings of Eurospeech*, pages 185–189, 1995.
- V. B. Le and L. Besacier. Automatic speech recognition for under-resourced languages: application to vietnamese language. *IEEE Transactions on Audio, Speech and Language Processing*, 17(8):1471–1482, 2009.
- K. F. Lee, S. Hayamizu, H. W. Hon, C. Huang, J. Swartz, and R. Weide. Allophone clustering for continuous speech recognition. In *Proceedings of ICASSP*, pages 749–752, 1990.
- C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech and language*, 9(2):171–185, 1995.
- W. P. Lehmann. *Historical Linguistics*. Routledge London and New York, third edition, 1993.
- M. P. Lewis and G. F. Simons. Assessing endangerment: Expanding fishman's gids. *Revue Roumaine de Linguistique*, 55(2):103–120, 2010.

- M. P. Lewis, G. F. Simons, and C. D. Fennig. *Ethnologue : Languages of the world, Seventh Edition*. SIL International, 2014. URL <http://www.ethnologue.com>.
- H. Lin, L. Deng, D. Yu, Y. fan Gong, A. Acero, and C.-H. Lee. A study on multilingual acoustic modeling for large vocabulary asr. In *Proceedings of ICASSP*, pages 4333–4336, Taipei, 2009.
- J. Löff, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney. The rwth 2007 tc-star evaluation system for european english and spanish. In *Proceedings of INTERSPEECH*, pages 2145–2148, Antwerp, Belgium, August 2007.
- J. Löff, C. Gollan, and H. Ney. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. In *Proceedings of INTERSPEECH*, Brighton, U.K., 2009.
- L. Lu, A. Ghoshal, and S. Renals. Regularized subspace gaussian mixture models for cross-lingual speech recognition. In *Proceedings of IEEE ASRU*, 2011.
- L. Lu, A. Ghoshal, and S. Renals. Maximum a posteriori adaptation of subspace gaussian mixture models for speech recognition. In *Proceedings of ICASSP*, 2012.
- L. Lu, A. Ghoshal, and S. Renals. Cross-lingual subspace gaussian mixture models for low-resource speech recognition. In *IEEE/ACM Transactions on Audio, Speech and Language Processing*, volume 22, pages 17–27, January 2014.
- K. R. Mabokela, M. J. Manamela, and M. Manaileng. Modeling code-switching speech on under-resourced languages for language identification. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*, pages 225–230, St. Petersburg, Russia, 2014.
- Y. M. Maris. *The Malay Sound System*. Siri Teks Fajar Bakti, Kuala Lumpur, 1979.
- S. R. Maskey, A. W. Black, and L. M. Tomokiyo. Bootstrapping phonetic lexicons for language. In *Proceedings of INTERSPEECH*, pages 69–72, 2004.
- M. Maxwell and B. Hughes. Frontiers in linguistic annotation for low-density languages. In *Proceedings of Workshop on Frontiers in Linguistically annotated corpora*, pages 29–37. Association for Computational Linguistics, 2006.

- Y. Miao and F. Metze. Improving low-resource cd-dnn-hmm using dropout and multilingual dnn training. In *Proceedings of INTERSPEECH*, pages 2237–2241, 2013.
- T. Mikolov, M. Karafiát, L. Burget, J. H. Černocký, and S. Khudanpur. Recurrent neural network based language model. In *Proceedings of Interspeech*, pages 1045–1048, Chiba, Japan, September 2010.
- T. Mikolov, S. Kombrink, L. Burget, J. H. Černocký, and S. Khudanpur. Extensions of recurrent neural network language model. In *Proceedings of ICASSP*, pages 5528–5531, 2011.
- A. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 20, pages 14–22, 2012.
- A. Mohan, S. H. Ghalehjegh, and R. C. Rose. Dealing with acoustic mismatch for training multilingual subspace gaussian mixture models for speech recognition. In *Proceedings of ICASSP*, pages 4893–4896, Kyoto, March 2012. IEEE.
- R. Molapo, E. Barnard, and F. de Wet. Speech data collection in an under-resourced language within a multilingual context. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*, pages 238–242, 2014.
- J. J. Morgan. Making a speech recognizer tolerate non-native speech through gaussian mixture merging. In *Proceedings of ICALL’04*, Venice, 2004.
- J. Mugabe, P. Kameri-Mbote, and D. Mutta. Traditional knowledge, genetic resources and intellectual property protection: Towards a new international regime. *International Environmental Law Research Center*, 2001. URL <http://www.ielrc.org/content/w0105.pdf>.
- W. D. Mulder, S. Bethard, and M.-F. Moens. A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, 30(1): 61 – 98, 2015. ISSN 0885-2308. doi: <http://dx.doi.org/10.1016/j.csl.2014.09.005>. URL <http://www.sciencedirect.com/science/article/pii/S088523081400093X>.
- E. L. Ng, A. W. Yeo, and B. Rainavo-Malançon. Identification of closely-related indigenous languages: an orthographic approach. In *Proceedings of International Conference on Asian Language Processing (IALP)*, pages 230–235. IEEE, 2009.

- J. R. Novak, N. Minematsu, and K. Hirose. Evaluations of an open source wfst-based phoneticezer. PDF, General Talk No. 452, The Institute of Electronics, Information and Communication Engineers, 2011.
- S. Novotney and C. Callison-Burch. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 207–215, Los Angeles, California, June 2010. Association for Computational Linguistics.
- A. H. Omar. *Phonology*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia, 1981.
- A. H. Omar. Perkaitan bahasa melayu dengan bahasa iban dari segi sejarah. In *Persidangan Antarabangsa Pengajian Melayu: Persoalan Warisan dan Kini*, Universiti Malaya, Kuala Lumpur, 1989.
- H. F. Ong and A. M. Ahmad. Malay language speech recognizer with hybrid hidden markov model and artificial neural network (hmm/ann). *International Journal of Information and Education Technology*, 1(2):114–119, 2011.
- M. Pitz and H. Ney. Vocal tract normalization as linear transformation of mfcc. In *Proceedings of Eurospeech*, Geneva, Switzerland, 2003.
- C. Plahl, B. Hoffmeister, M.-Y. Hwang, D. Lu, G. Heigold, J. Löff, R. Schlüter, and H. Ney. Recent improimprove of the rwth gale mandarin lvcsr system. In *Proceedings of INTERSPEECH*, pages 2426–2429, Brisbane, Australia, September 2008.
- D. Povey, H.-K. J. Kuo, and H. Soltau. Fast speaker adaptive training for speech recognition. In *Proceedings of INTERSPEECH*, pages 1245–1248, Brisbane, Australia, 2008.
- D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas. Subspace gaussian mixture models for speech recognition. In *Proceedings of ICASSP*, 2010.
- D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. G. M. Karafiat, A. Rastrow, R. C. Rose, P. Schwartz, and S. Thomas. The subspace gaussian mixture model - a structured model for speech recognition. *Computer Speech and Language*, 25:404–439, 2011a.



- D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý. The kaldi speech recognition toolkit. In I. S. P. Society, editor, *Proceedings of Workshop on Automatic Speech Recognition and Understanding*, volume IEEE Catalog No. : CFP11SRW-USB, December 2011b.
- L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of IEEE*, volume 77, pages 257–286, 1989.
- A. Rousseau, P. Deléglise, and Y. Estève. Ted-lium: An automatic speech recognition dedicated corpus. In *Proceedings of LREC*, pages 125–129. European Language Resources Association (ELRA), 2012.
- D. Rybach, S. Hahn, C. Gollan, R. Schlüter, and H. Ney. Advances in arabic broadcast news transcription at rwth. In *Proceedings of ASRU*, pages 449–454, Kyoto, Japan, December 2007.
- D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Löff, Schlüter, and H. Ney. The rwth aachen university open source speech recognition system. In *Proceedings of INTERSPEECH*, pages 2111–2114, Brighton, U.K., September 2009.
- K. P. Scannell. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, pages 5–15, 2007.
- T. Schultz. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Proceedings of ICLSP*, pages 345–348, 2002.
- T. Schultz and A. Waibel. Fast bootstrapping of lvcsr systems with multilingual phoneme sets. In *Proceedings of Eurospeech*, pages 371–374. Citeseer, 1997.
- T. Schultz and A. Waibel. Multilingual and crosslingual speech recognition. In *Proceedings of DARPA workshop on Broadcast News Transcription and Understanding*, pages 259–262, 1998.
- T. Schultz and A. Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35 : 1:31–52, 2001.
- T. Schultz, N. T. Vu, and T. Schilp. Globalphone: A multilingual text and speech database in 20 languages. In *Proceedings of ICASSP*, 2013.

- F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Proceedings of INTERSPEECH*, pages 437–440, 2011.
- A. Sixtus and H. Ney. From within-word model search to across-word model search in large vocabulary continuous speech recognition. *Computer Speech and Language*, 16(2):245–271, May 2002.
- A. Stolcke. Srilm - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904, 2002.
- M. Swadesh. Lexico-statistic dating of prehistoric ethnic contacts. In *Proceedings of the American Philosophical Society*, volume 96, pages 452–463, 1952.
- P. Swietojanski, A. Ghoshal, and S. Renals. Unsupervised cross-lingual knowledge transfer in dnn-based lvcsr. In *Proceedings of ICASSP*, 2013.
- T.-P. Tan and L. Besacier. Acoustic model interpolation for non-native speech recognition. In *Proceedings of ICASSP*, 2007.
- T.-P. Tan and B. Rainavo-Malançon. Malay grapheme to phoneme tool for automatic speech recognition. In *Proceedings of Workshop of Malaysia and Indonesia Language Engineering (MALINDO) 2009*, 2009.
- T.-P. Tan, H. Li, E. K. Tang, X. Xiao, and E. S. Chng. Mass: a malay language lvcsr corpus resource. In *Proceedings of Oriental COCOSDA International Conference 2009*, pages 26–30, 2009.
- T.-P. Tan, L. Besacier, and B. Lecouteux. Acoustic model merging using acoustic models from multilingual speakers for automatic speech recognition. In *Proceedings of International Conference on Asian Language Processing (IALP)*, 2014.
- E. Titariy, N. Lotner, M. Gishri, and A. Moyal. A hybrid keyword spotting approach for combining lvcsr and phonetic search. In *Proceedings of Speech Processing Conference*, Tel Aviv, Israel, July 2014.
- R. Tong, B. P. Lim, N. F. Chen, B. Ma, and H. Li. Subspace gaussian mixture models for computer-assisted language learning. In *Proceedings of ICASSP*, pages 5347–5351. IEEE, 2014.

- P. Vozila, J. Adams, Y. Lobacheva, and T. Ryan. Grapheme to phoneme conversion and dictionary verification using graphonemes. In *Proceedings of Eurospeech*, 2003.
- N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li. A first speech recognition system for mandarin-english code-switch conversational speech. In *Proceedings of ICASSP*, pages 4889–4892, 2012.
- N. T. Vu, D. Imseng, D. Povey, P. Motlíček, T. Schultz, and H. Bourlard. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *Proceedings of ICASSP*, 2014.
- M. Walsh. Will indigenous languages survive? *Annual Review of Anthropology*, 34: 293–315, 2005.
- Z. Wang, T. Schultz, and A. Waibel. Towards universal speech recognition. In *Proceedings of International Conference on Multimodal Interfaces*, 2002.
- J. Weiner, N. T. Vu, D. Telaar, F. Metze, T. Schultz, D.-C. Lyu, E. Chng, and H. Li. Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*, pages 76–79, 2012.
- L. Welling, S. Kanthak, and H. Ney. Improved methods for vocal tract normalization. In *Proceedings of ICASSP*, volume 2, pages 761–764, Phoenix, Arizona, USA, March 1999.
- J. Wells, W. Barry, M. Grice, A. Fourcin, and D. Gibbon. Standard computer-compatible transcription. Doc. no. sam-ucl-037, Phonetics and Linguistics Dept, UCL, London, 1992.
- J. C. Wells. Computer-coding the ipa: a proposed extension of sampa, 1995. URL <http://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf>.
- J. C. Wells. *Handbook of Standards and Resources for Spoken Language Systems*, chapter SAMPA computer readable phonetic alphabet. Berlin and New York: Mouton de Gruyter, 1997. URL [www.phon.ucl.ac.uk/home/sampa](http://www.phon.ucl.ac.uk/home/sampa). Part IV, section B.
- P. J. Werbos. Backpropagation through time: what it does and how to do it. In *Proceedings of IEEE*, volume 78, pages 1550–1560, 1990.

- I. Witten and T. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37: 1085–1094, 1991.
- K. B. Wright. Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, 2005.
- S. Wurm. *Language Diversity Endangered*, chapter Threatened Languages in the Western Pacific Area from Taiwan to, and including Papua New Guinea, pages 374–390. De Gruyter Mouton, 2008.
- X. Xiao, E. S. Chng, T.-P. Tan, and H. Li. Development of a malay lvscr system. In *Proceedings of Oriental COCOSDA*, Kathmandu, Nepal, 2010.
- S. J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J. L. Gauvain, D. J. Kershaw, L. Lamel, D. A. Leeuwen, D. Pye, A. J. Robinson, H. J. M. Steeneken, and P. C. Woodland. Multilingual large vocabulary speech recognition: the european sqale project. *Computer Speech and Language*, 11:73–89, 1997.
- R. M. Yusof. Perkaitan bahasa melayu dan bahasa iban: Satu tinjauan ringkas. *Jurnal Bahasa*, 3(3), 2003.
- Y. Zhang and J. R. Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *Proceedings of IEEE ASRU*, pages 398–403. IEEE, 2009.