Development of a Cantonese-English Code-mixing Speech Recognition System

CAO, Houwei

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

The Chinese University of Hong Kong
May 2011

UMI Number: 3497756

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent on the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI 3497756

Copyright 2012 by ProQuest LLC.

All rights reserved. This edition of the work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

To my parents, for their love, consideration, and support.

Acknowledgements

First of all, I am heartily thankful to my supervisor, Prof. Pak-Chung CHING, for providing me the chance to work on this interesting topic. This thesis would not have been possible without his encouragement, guidance and support. I would like to express my sincere gratitude to Prof. Tan LEE for his advice and suggestions throughout this research. His encouragement and patience made me less frustrated when I was in the face of the difficulties. Prof. CHING and Prof. LEE also supported me to attend overseas academic activities.

I would also like to thank Dr. Frank K. SOONG, Dr. Yao QIAN, Prof. William Shi-Yuan WANG, Prof. Helen MENG and Prof. Wing-Kin MA for their precious suggestions. Special gratitude goes to Prof. De-Yuan CHENG, who introduced me to the area of speech and language processing when I was an undergraduate student.

Many thanks are due to all my friends and colleagues in DSP and the Speech Technology Laboratory, who have helped and encouraged me a lot in different ways. In particular, Mr. Yu Ting YEUNG shared me with lots of experience on this research project. Ms. Ning WANG always encouraged me when I was down. Dr. Yvonne Siu Wa LEE and Dr. Wai-Man NG were always kind enough to discuss matters and help me. Dr. Nengheng ZHENG has given me many suggestions both in relation to my study and life. I am also grateful to Kin On LUK for his technical support. Without these friends, my life in CUHK would have not been so colourful and interesting all these years. Also I would like to thank all colleagues of MSRA, for their support during my internship.

Finally, I wish to express my deepest gratitude to my parents, for their continuous love, support and understanding.

Abstract of thesis entitled:

Development of a Cantonese-English Code-mixing Speech Recognition System

submitted by **CAO, Houwei**

for the degree of **Doctor of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong** in

April 2011.

Code-mixing is a common phenomenon in bilingual societies. It refers to the intra-sentential switching of two languages in a spoken utterance. This thesis addresses the problem of the automatic recognition of Cantonese-English code-mixing speech, which is widely used in Hong Kong.

While automatic speech recognition (ASR) of either Cantonese or English alone has achieved a great degree of success, recognition of Cantonese-English code-mixing speech is not as trivial. Unknown language boundary, accents in code-switched English words, phonetic and phonological differences between Cantonese and English, no regulated grammatical structure, and lack of speech and text data make the ASR of code-mixing utterances much more than a simple integration of two monolingual speech recognition systems. On the other hand, we have little understanding of this highly dynamic language phenomenon. Unlike in monolingual speech recognition research, there are very few linguistic studies that can be referred to.

This study starts with the investigation of the linguistic properties of Cantonese-English code-mixing, which is based on a large number of real code-mixing text corpora collected from the internet and other sources. The effects of language mixing for the automatic recognition of Cantonese-English code-mixing utterances are analyzed in a systematic way. The problem of pronunciation dictionary, acoustic modeling and language modeling are investigated.

Subsequently, a large-vocabulary code-mixing speech recognition system is developed and implemented.

A data-driven computational approach is adopted to reveal significant pronunciation variations in Cantonese-English code-mixing speech. The findings are successfully applied to constructing a more relevant bilingual pronunciation dictionary and for selecting effective training materials for code-mixing ASR. For acoustic modeling, it is shown that cross-lingual acoustic models are more appropriate than language-dependent models. Various cross-lingual inventories are derived based on different combination schemes and similarity measurements. We have shown that the proposed data-driven approach based on K-L divergence and phonetic confusion matrix outperforms the IPA-based approach using merely phonetic knowledge. It is also found that initials and finals are more appropriate to be used as the basic Cantonese units than phonemes in code-mixing speech recognition applications. A text database with more than 9 million characters is compiled for language modeling of code-mixing ASR. Class-based language models with automatic clustering classes have been proven inefficient for code-mixing speech recognition. A semantics-based n-gram mapping approach is proposed to increase the counts of code-mixing n-gram at language boundaries. The language model perplexity and recognition performance has been significantly improved with the proposed semantics-based language models. The proposed code-mixing speech recognition system achieves 75.0% overall accuracy for Cantonese-English code-mixing speech, while the accuracy for Cantonese characters is 76.1% and accuracy for English lexicons is 65.5%. It also attains a reasonable character accuracy of 75.3% for monolingual Cantonese speech.

Cross-lingual speaker adaptation has also been investigated in the thesis. Speaker independent (SI) model mapping between Cantonese and English is established at different levels of acoustic units, viz phones, states, and Gaussian mixture components. A novel approach for cross-lingual speaker adaptation via Gaussian component mapping is proposed and has been proved to be effective in most speech recognition tasks.

摘　　要

語碼混合（code-mixing）現象指的是說話者在一句話中使用兩种語言或語言變體的現象。該現象在雙語社會中非常普遍。在香港，廣東話及英語的語碼混合(Cantonese-English code-mixing) 在人們的日常對話中極爲常見，其主要形式是在廣東話口語中插入英語單詞或詞組。本文主要針對廣東話及英語語碼混合的語音識別方法進行了深入的研究。

語碼混合是一種多變而複雜的語言現象。廣東話及英語語碼混合語音的自動識別是一項艱巨的任務。未知的語言邊界，帶口音的英語發音，廣東話和英語在語音學和音韻學上的差別，不規則的語法結構，以及缺少訓練數據都使語碼混合語音識別與廣東話或英語的單語種語音識別相比更爲困難。

本文首先由語言學特性的角度出發，對大量的廣東話及英語語碼混合數據進行了統計分析。並且，我們還針對這種語碼混合現象對語音識別性能所造成的影響進行了系統的分析和討論。

針對口音問題，我們探討了廣東話及英語語碼混合語音中可能出現的發音變異，並以此爲依據修改了發聲字典。在聲學模型方面，跨語種（cross-lingual）模型比單語種（monolingual）模型更適合識別語碼混合的語音。我們採用了不同測量相似度的方法估計廣東話和英語的不同發聲單元之間的聲學距離（acoustic distance）和發音距離（phonetic distance），並以此爲依據集群了針對廣東話及英語語碼混合的跨語

種音素系統（phoneme inventory）。缺少足夠的廣東話及英語語碼混合的文本數據使我們難以建立有效語言學模型。我們發現基於自動分類的分類語言學模型（class-based LM）並不能改善由於訓練文本的數據稀疏所造成的在語音識別上的缺陷。本文提出了一種基於語義的 n 元組（n-gram） 映射方法，以期有效增加廣東話及英語語碼混合的 n 元組的出現頻率。實驗結果表明該方法可以有效的改善語言學模型混淆度並提高語音識別正確率。

本文所提出的語音識別系統，對廣東話及英語語碼混合語音的識別正確率爲 75.0%，其中廣東話單字的正確率爲 76.1%，而英語單詞的正確率則爲 65.5%。實驗結果進一步表明該系統也可以成功的識別單語種廣東話口語的語句，其識別正確率爲 75.3%。

本文還對跨語種之間的說話人自適應問題（speaker adaptation）進行了研究。我們提出了一種新穎的基於高斯混合分量映射的跨語種自適應方法。實驗結果表明利用該方法，我們可以成功的利用少量的特定人的廣東話數據對該說話人進行英語的說話人自適應。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Summary

This chapter provides the background and motivation for this thesis. We start by describing a brief history of the development of automatic speech recognition (ASR), followed by review of code-mixing speech recognition. Then we focus on Cantonese-English code-mixing ASR. Challenges in code-mixing ASR are highlighted and various Cantonese and code-mixing speech recognition systems are compared. Significant degradation from Cantonese ASR to code-mixing ASR helps us to establish the motivation of our research, which is to improve the performance of Cantonese-English code-mixing LVCSR. After that, we highlight the major research directions of this thesis. The chapter concludes with the organization of the thesis.

1.1 A Brief History of Automatic Speech Recognition

Speech is the most effective and decisive method of communication between humans. In addition to human-human communication, speech communication is also preferred in human-machine interaction with the advances of computer technology. Automatic speech recognition (ASR) is one of the key technologies in speech communications. The goal of ASR is to convert an input speech waveform into its written form. ASR has many applications, such as information query or retrieval systems [1][2], booking or telephone routing systems [3][4], voice dictation or broadcast news transcription systems[5][6], speech-to-speech translation systems[7], etc.

In the last six decades, automatic speech recognition has witnessed the remarkable development from isolated speech recognition (ISR) to continuous speech recognition (CSR), from keyword spotting to large vocabulary continuous speech recognition (LVCSR), from a speaker-dependent system to speaker-independent system, and from one language to several languages. The first attempts to develop ASR systems were made in the early 1950s. Bell Laboratories built the first isolated digit recognizer for a single speaker in 1952 [8]. This system was mainly dependent on measuring spectral resonance during the vowel segment of each digit. In the 1960s and 1970s, isolated word recognition was a key focus of research. Many pattern recognition ideas and signal processing techniques were successfully applied in ASR. These techniques include dynamic programming [9], linear predictive coding (LPC) [10], etc. Dynamic time wrapping (DTW) was used as the state-of-the-art approach for ISR. In the mean time, researchers in AT&T Bell Labs began a series of experiments aimed at creating speaker-independent recognition systems.

In the 1980s, the research focus shifted from ISR to connected word recognition. More successful statistical modeling methods displaced template-based approaches in recognizing continuous speech [11] [12]. Hidden Markov modeling became widely applied in virtually every speech recognition research laboratory

from the mid-1980s. On the other hand, neural networks were commonly used in implementing speech recognition systems as well [13] [14].

From the 1980s to 1990s, the major research impetus was given to large vocabulary continuous speech recognition. The Defense Advance Research Project Agency (DARPA) supported a large research programme aimed at the development of high-accuracy continuous speech recognition systems with 1,000 words. The first speaker-independent LVCSR system SPHINX was built by CMU [15], and researchers in AT&T [16], BBN [17], IBM [18], Lincoln labs [19] and MIT [20] made major contributions to this DARPT project as well.

From the 1990s to 2000s, with the acceleration of globalization, the research interest in speech recognition which was developed originally for one language has been exported to several languages. Several multi-lingual/cross-lingual speech recognizers were successfully built [21][22][23].

Various languages may have different linguistic characteristics in terms of the sound system, prosodic and phonological features, the written form, the relation between letters and sounds, the presence or absence of a segmentation of the written text into useful units, the morphology of the languages, etc. All these factors have had a significant impact on the task of developing a recognition system for a given language. For example, in tonal languages such as Mandarin Chinese and Thai, the pitch contour or pitch level on a single syllable plays a significant role in its contrastive lexical functions. In such case, integration of tone information in speech recognition can be effective in improving recognition performance. On the other hand, some languages such as English and French have a natural segmentation of the written form into word units that can be properly used as lexical items in pronunciation dictionaries and language modeling. However, many languages like Chinese and Japanese lack a natural segmentation, which are written out without any spacing between adjacent words. Word segmentation is required in language modeling for these languages. In addition, language modeling for languages with very divergent written and spoken forms such as Chinese and Arabic are more difficult than that of languages with standard written forms such as English.

1.2 Previous Works on Code-mixing Speech Recognition

Code-mixing refers to the intra-sentential switching of two different languages in a spoken utterance. It is a common phenomenon in many bilingual societies, such as Spanish-English in United State, French-Italian in Switzerland, Mandarin-English in Taiwan, etc. Research activities on the automatic speech recognition of code-mixing speech have a relatively short history. Since mid 2000s, several speech recognition systems have been developed in Hong Kong, Taiwan, Singapore, and Mainland China. All of them focus on Mandarin-Taiwanese, Mandarin-English, and Cantonese-English code-mixing speech. Little work has been done on other combinations of languages found in code-mixing. In the following, we shall describe some representative work related to code-mixing speech recognition.

Speech Corpora

The research into code-mixing speech processing needs a large number of speech data. The development of speech corpora is therefore an important part of works. Several code-mixing speech corpora have been built for language identification (LID), language boundary detection (LBD), and automatic speech recognition (ASR) tasks.

To the author's knowledge, there are only two phonetically rich code-mixing speech corpora which were created for the development of speaker-independent LVCSR systems. [24] took special interest in Mandarin-English code-mixing speech from the South-East Asia region. This corpus (codenamed SEAME) is a 30 hours real spontaneous Mandarin-English code-mixing speech corpus recorded from Singapore and Malaysia speakers. All code-mixing utterances were recorded under interview and conversational settings. CUMIX [25] is a Cantonese-English code-mixing speech database designed for the training of Cantonese-English code-mixing acoustic models, and to evaluate the performance of the code-mixing speech recognition system. The data can also be

used to make a thorough study on LBD within code-mixing utterances. Besides, some mix-language speech data was collected for evaluation purpose. In [26], a read-style speech corpus with mixed-language was developed to evaluate the performance of the proposed LBD approaches. It contains both Mandarin-Taiwanese and Mandarin-English mixed utterances. In [27], hundreds of noisy Mandarin-English code-mixing utterances were collected under realistic conditions such as in restaurants, streets and other noisy places, which were used to evaluate the retrieval system.

Language Boundary Detection

Code-mixing speech recognition can be tackled as two monolingual recognition tasks if the input mixed-language utterance can be separated into language-homogeneous segments correctly. Language-specific phonological and acoustic properties were used as the primary cues to identify the languages. In [26], delta Bayesian information criteria (Δ BIC) was applied to detect the changing point between two languages at first. After that a statistical language ID framework, incorporating both LSA-based GMMs and VQ-based bi-gram LM, was used to determine the optimal number of language boundaries. Two different methods of LBD was evaluated on Cantonese-English code-mixing speech in [28]. LBD based on syllable bigram exploited the phonological and lexical differences between Cantonese and English. LBD based on syllable lattice made use of the intermediate result of speech recognition, which was more informative than the prior linguistic knowledge. [29] proposed a language identification method integrated multiple levels of linguistic cues. Acoustic, prosodic and phonetic features were used to distinguish Mandarin and Taiwanese. These previous studies showed the performance level of LBD was around 70%-80%.

Large-scale Speech Recognition Systems

Code-mixing speech recognition is still in its infancy. Only several LVCSR systems have been developed. In [30], automatic recognition of Mandarin-Taiwanese code-switching speech was investigated. It was found that Mandarin

and Taiwanese, both of which are Chinese dialects, share a large percentage of lexicon items. Their grammar was also assumed to be similar. A one-pass recognition algorithm was developed using a character-based search net. It was shown that the one-pass approach outperforms LBD-based multi-pass approaches. [31] was the first study on automatic recognition of Cantonese-English code-mixing speech. A two-pass cross-lingual recognition system was developed. The cross-lingual phoneme set was designed based on phonetic knowledge. For language modeling, the class-based language models were considered. The two pass search algorithm enabled flexible integration of language boundary information as one of the confidence scores, in addition to the acoustic and language model scores, for decoding the hypothesis mixed-language word string.

Other Applications

In addition to LVCSR task, code-mixing speech recognition has many other applications. In [64], a mixed-lingual keyword spotting system was developed for auto-attendant applications. The keywords to be detected could be in either English or Chinese. Code-mixing speech recognition can also be regarded as a translation problem. In such system, the embedded words are translated to the matrix language of the utterance. An appropriate lexicon is selected in the matrix language in order to maximize the language model likelihood [65]. In [27], a grammar constrained, Mandarin-English bilingual speech recognition system was developed for real world music retrieval. It enabled users to find a song by simply saying the name or title of the singer or song, which are allowed to be either monolingual or bilingual.

1.3 Motivation of our Research

Hong Kong is an international city where many people, especially the young generation, are Cantonese and English bilinguals. There has also been a trend that people tend to frequently embed English words into spoken Cantonese utterances, e.g. "我今個星期要趕三個 deadline 啊", "能夠同你 work together

我覺得好 exciting". Therefore, it would be highly desirable if we can develop an ASR system which is able to handle Cantonese-English code-mixing speech in addition to monolingual Cantonese utterances.

However, Cantonese and English are quite different languages. Cantonese is a tonal and mono-syllabic spoken dialect in the Sino-Tibetan family, while English is a stress-time language in the Indo-European family. They have different sound structures in terms of phonetics, phonology and prosody properties. On the other hand, Cantonese and English differ markedly in syntax and morphology as well. As a result, language-specific characteristics must be taken into account in developing recognition systems for Cantonese and English.

Moreover, automatic speech recognition of mixed-language utterances is much more than a simple integration of two monolingual speech recognition systems. First, there is no prior knowledge about when there is a switch of language so that we cannot determine which of the two recognizers should be used for a particular speech segment. While some automatic language identification techniques have been proposed, they are less successful for the code-mixing scenarios because the switching is at word level, and thus the language segments are of relatively short duration [26] [32]. Second, it is very often that the English words embedded into a Cantonese utterance are spoken with strong Cantonese accents, which a monolingual ASR system for standard English is unable to handle. Third, mixed-language speech adopts special grammar that cannot be inferred from monolingual speech. Language models need to re-built from mixed-language data.

Figure 1.1 compares the recognition performance of various Cantonese and code-mixing systems. It is noted that there is significant degradation from monolingual Cantonese ASR to code-mixing ASR. The best character accuracy attained for read-style standard Cantonese is 86.1%, which was reported in [33]. An algorithm of explicit tone recognition was integrated into this Cantonese LVCSR system [34]. On the other hand, as reported in [35], a recognition accuracy of 80.3% can be achieved without tonal information. Moreover, Cantonese is a spoken Chinese dialect in which the formal or standard form is significantly

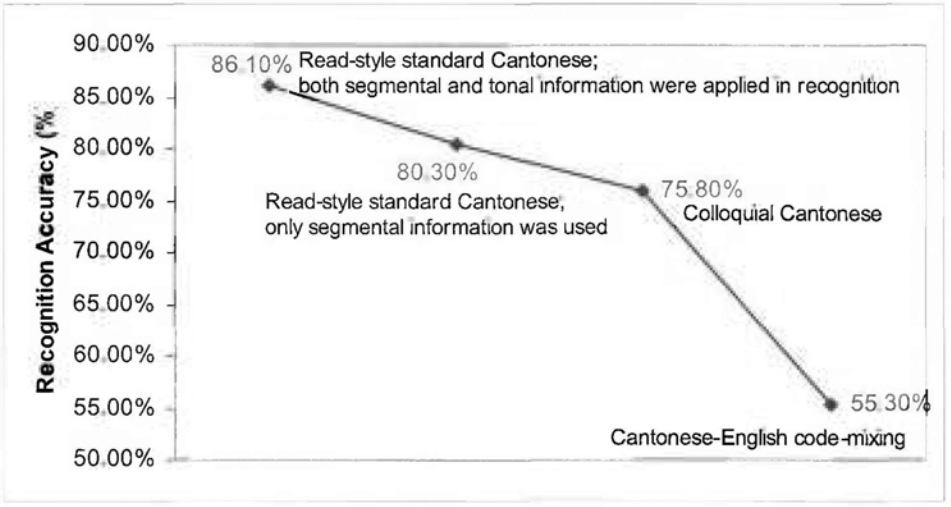


Figure 1.1: Recognition performance of different Cantonese and code-mixing speech recognition systems

different from the spoken or colloquial form. In [36], automatic recognition of spoken Cantonese was investigated. The proposed recognition system achieved the character accuracies of 75.8% for spoken Cantonese. Cantonese-English code-mixing usually occurs in casual conversational speech. In other words, English words are frequently embedded into spoken Cantonese. Recognition of Cantonese-English code-mixing speech was first studied by Chan *et al.* [31]. The recognizer attained the overall accuracy of 55.3%, while the accuracy was 56.4% and 53% for the Cantonese characters and embedded English words respectively. Compared with the monolingual results for colloquial Cantonese, more than 20% degradation can be observed. Accordingly, in this research, an LVCSR system is built aimed at achieving promising recognition accuracy for Cantonese-English code-mixing speech.

1.4 Thesis Goals and Research Focuses

The general goal of this research is to improve the performance of Cantonese-English code-mixing LVCSR. Towards this goal, our research focus is twofold. Firstly, code-mixing is a highly dynamic language phenomenon. To better

understand its unique feature, the characteristics of Cantonese-English code-mixing are investigated from a linguistic point of view in this thesis. Secondly, we plan to develop a high-performance LVCSR system which is capable to handle Cantonese-English code-mixing speech in addition to monolingual Cantonese speech. Our study covers all components of an ASR system, including acoustic models, language models and pronunciation dictionary. Subsequently, a large-vocabulary code-mixing speech recognition system is developed based on a two-pass decoding algorithm. Specifically, the following issues are considered and addressed.

Linguistic properties of Cantonese-English code-mixing

As the very beginning, we have little understanding of Cantonese-English code-mixing speech. This language phenomenon is not a simple insertion of one language into another. However, compared with many monolingual languages, there are very few linguistic studies that can be referred to. We have to understand the problems by actually working on them. In this thesis, we collect about 65,000 Cantonese-English code-mixing sentences from newspapers, magazines and online diaries at first. Linguistic study is carried out using the in-house collected data.

Effects of language mixing for code-mixing ASR

As we mentioned before, there is significant degradation from monolingual ASR to code-mixing ASR. Before making efforts to improve recognition performance, we first attempt to study how and why such degradation is caused when English is embedded in matrix Cantonese. In this research, the effects of language mixing for code-mixing ASR are investigated in a systematic way.

Pronunciation variations in code-mixing speech

It is observed that there exists many pronunciation variations in Cantonese-English code-mixing speech. Due to inconsistent pronunciation, we often get lower speech recognition accuracy on code-mixing speech than monolingual Can-

tonese/English utterances spoken by native speakers. Understanding how native Cantonese/English and Cantonese-English code-mixing speech differs in terms of pronunciations is an important first step to tackle the problem of code-mixing speech recognition. To improve the design of acoustic models and construct a more accurate bilingual pronunciation dictionary, in-depth studies on pronunciation variation in code-mixing speech are conducted in this thesis.

Cross-lingual acoustic modeling

Cantonese and English come from two different language families. They have different phonetics and phonological structures. Different phonetic units can be applied to represent Cantonese and English due to their phonological difference. It is also expected that some of the phonetic models are language-specific and the others are shared between Cantonese and English. This part of the research aims at designing an appropriate sound inventory for acoustic modeling of Cantonese-English code-mixing speech.

Language modeling for code-mixing ASR

In practice, we need to collect a large amount of Cantonese-English code-mixing text data to train statistical n-gram language models. However, because of the colloquial speaking-style and domain-specific property of code-mixing, there are practical difficulties in data collection. In this thesis, we attempt to make use of the limited amount of monolingual Cantonese and code-mixing text resources available to improve the recognition of Cantonese-English code-mixing speech.

Cross-lingual adaptation

Although both Cantonese and English are official spoken languages in Hong Kong, the usage of English is much less than Cantonese in daily communication, therefore it is much easier to collect small quantity of Cantonese speech data from a specific group of Cantonese speakers. Speaker adaptation techniques can be used to improve speech recognition performance when a small set of adaptation data from the target speaker is available. However, it is not

easy to do it across different languages, especially when the two languages are phonetically distant apart. This section of the research is focused on the use of acoustic information from an existing source language (Cantonese) to implement speaker adaptation for a new target language (English).

1.5 Thesis Outline

In the next chapter, the fundamentals of LVCSR systems are briefly reviewed. Different approaches for monolingual, multilingual/crosslingual, and code-mixing ASR are discussed.

Chapter 3 focuses on Cantonese-English code-mixing in Hong Kong. The nature of Cantonese-English code-mixing is investigated from a linguistic point of view. The difficulties and effects of language mixing for code-mixing ASR are analyzed and discussed.

Chapter 4 discusses the cross-lingual use of acoustic information for Cantonese and English. Pronunciation variation in code-mixing speech is studied and cross-lingual acoustic models are developed. In addition, a novel approach for cross-lingual adaptation via model mapping is described.

Language modeling for Cantonese-English code-mixing speech recognition is investigated in Chapter 5. In the absence of a sufficient amount of code-mixing text data, different language modeling techniques are investigated.

In Chapter 6, an LVCSR system for Cantonese-English code-mixing speech is developed and implemented. The recognition results and analysis are elaborated.

Finally, Chapter 7 summarizes the major contributions of this thesis, followed by some suggestions for future work.

□ End of chapter.

Chapter 2

LVCSR for Multilingual and Code-mixing Speech

Summary

With the globalization of today's world, one of the most important trends in present-day speech technology is the need to support multiple input languages. Research related to multilingual and cross-lingual speech has attracted much attention over the past few years. In this chapter, we introduce the state-of-the-art of multilingual and code-mixing speech recognition. The chapter starts with an overview of large vocabulary continuous speech recognition (LVCSR) systems. Feature extraction of speech signal, acoustic modeling, language modeling and decoding algorithms are described. Then the speech recognition of monolingual and multilingual speech are compared. After that we focus on code-mixing speech recognition. Two different approaches to code-mixing speech recognition are discussed. Previous studies on code-mixing speech recognition of different language combinations are reviewed. Finally, an information-theoretic similarity measurement Kullback-Leibler divergence (KLD) is introduced.

2.1 Fundamentals of LVCSR

Large vocabulary continuous speech recognition (LVCSR) systems deal with fluently spoken speech with a vocabulary of thousands of words or more [37]. The state-of-the-art approach to LVCSR is to treat the speech signal as a stochastic pattern based on statistical pattern recognition algorithm. The source-channel model of speech generation and speech recognition, shown in Figure 2.1, is introduced in this approach [38].

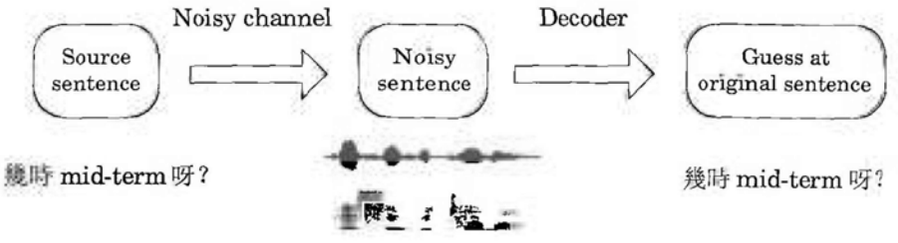


Figure 2.1: Source-channel model of speech generation and speech recognition

W denotes word sequences of the source sentence. The conversion from W to an observed speech signal S is modeled as a noisy channel on account of uncertainty in conversion. Instead of dealing with the speech signal S directly, S is firstly transformed into a sequence of acoustic feature vectors A. As a result, the problem of speech recognition is to find out the most probable word sequence $\hat{W}$ given A. This problem can be formulated as a *maximum a posterior* (MAP) decoding problem.

$$\hat{W} = \arg\max_{W} P(W|A) \tag{2.1}$$

By using the Bayes rule, the problem can be reformulated as:

$$\hat{W} = \arg\max_{W} P(A|W)P(W) \tag{2.2}$$

where $P(A|W)$ represents the conditional probability that A is produced when a particular word sequence W is being spoken. It is often referred to as an acoustic model. In typical LVCSR system, acoustic models are usually built at sub-word level. A pronunciation dictionary is used to define the ways in which the sub-word units such as phonemes can be concatenated to form words. The second

term $P(W)$ is the *a prior* probability of generating the sequences of word W, which is independent of acoustics and referred to as a language model. Language models are used to capture the regularities of the language and constrain the ways for the generation of meaningful sentences [39]. The key components in an LVCSR system based on this statistical approach are shown in Figure 2.2.

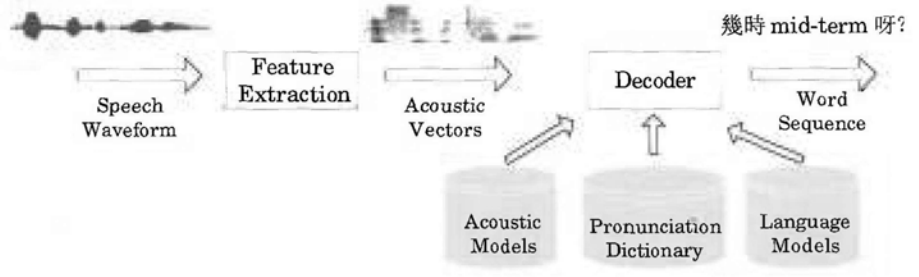


Figure 2.2: The flow diagram of a typical LVCSR system

2.1.1 Feature Extraction

To perform statistical pattern-matching, the input speech waveform needs to be converted to a sequence of acoustic feature vectors representing a short-time speech spectrum covering a period of typically 10 ms.

The Mel frequency cepstral coefficient (MFCC), linear predictive coding (LPC) [40] and perceptual linear predictive (PLP) [41] are the three most popular features used in speech recognition systems. By taking advantage of perceptional mel-scale filterbanks, MFCC outperform other acoustic features in speech recognition tasks [42]. Figure 2.3 shows the flow-diagram of the extraction of MFCC features.

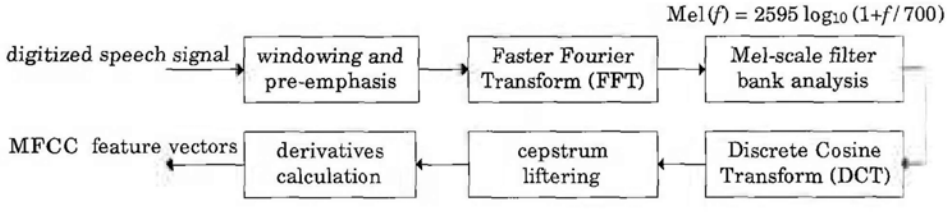


Figure 2.3: Flow-diagram of the extraction of MFCC features

2.1.2 Acoustic Model (AM)

Acoustic models are used to characterize the statistical variation of the acoustic features of a specified linguistic unit. AMs include the representation of knowledge of acoustics, phonetics, environment variability and speaker variation. Hidden markov models (HMMs) is the dominant technique in most state-of-the-art LVCSR systems for acoustic modeling [12]. An HMM is a finite state machine to generate a sequence of feature vectors, in which the actual state sequence is unknown. In acoustic models, acoustic feature vectors of a linguistic unit are observed as the output generated by an HMM and the HMM can be referred to as the template of that specific linguistic unit.

Word-level HMMs are impractical for LVCSR system and, instead, words are decomposed into sub-word units such as phones. Phone-level HMMs typically have three emitting states and a simple left-to-right topology as shown in Figure 2.4. The entry and exit states are included to concatenate models together. In addition, in view of the acoustic variations caused by contextual effect, different HMMs have to be trained for different contexts to achieve good phonetic discriminations. Triphone models are the most common context-dependent models used in LVCSR, where every phone has a distinct HMM for every unique pair of preceding and succeeding phones.

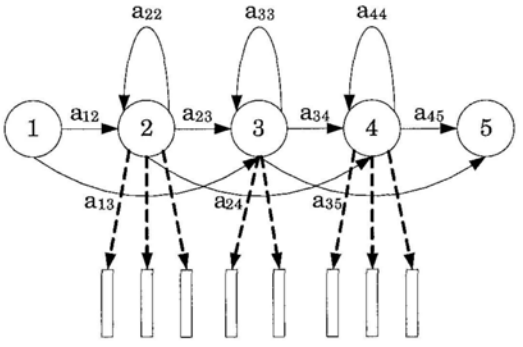


Figure 2.4: A Simple five-state left-to-right HMM

In practice, an LVCSR system typically involves hundreds of thousands of triphones. Huge amounts of parameters are needed to train when triphones are considered. It is definitely a crucial problem that too many parameters

are estimated with too little training data. To deal with this data sparseness problem, similar HMMs can be tied together. After tying, several states will share the same distributions. Hence more data are available to train a tied state and therefore give more robust estimation for the parameters of that tied state. Decision tree clustering is the most common approach for state tying, which has led to substantial improvements in recognition performance [43]. Other data-driven tying approaches have been studied in [44][45].

2.1.3 Language Model (LM)

Language models have been widely used in various natural language processing applications during the past three decades, and attempt to capture the regularities and properties of languages. In an LVCSR system, the purpose of language modeling is to improve recognition performance by making use of syntactic and lexicon information. A statistical language model assigns a probability $P(W)$ to a word sequence W. We suppose that the word sequence consists of M words, $W = w_1, w_2, ..., w_M$, so the probability $P(W)$ of observing the word sequence W can be computed:

$$P(W) = P(w_1, w_2, ..., w_M) = \prod_{i=1}^{M} P(w_i|w_1, w_2, ..., w_{i-1}) \tag{2.3}$$

However, it is impractical to predict the ith word w_i in the context of the entire history of preceding $i-1$ words. A simple but effective way is to use N-grams, in which it is assumed that w_i only depends on the preceding $N-1$ words. As a result, equation 2.3 can be approximated as:

$$P(W) \approx \prod_{i=1}^{M} P(w_i|w_{i-N+1}, ..., w_{i-1}) \tag{2.4}$$

The conditional probability in the N-gram language model can be calculated by a simple frequency count:

$$P(w_i|w_{i-N+1}, ..., w_{i-1}) = \frac{count(w_{i-N+1}, ..., w_{i-1}, w_i)}{count(w_{i-N+1}, ..., w_{i-1})} \tag{2.5}$$

Trigram (N=3) and bigram (N=2) language models dominate conventional speech recognition systems. The trigram LM is a common choice with large

training corpora with millions of words, whereas the bigram LM is often used with smaller ones.

The data sparsity problem is the major issue in language modeling. With limited training data, many trigrams will appear only once or twice, or even not appear. However, this does not mean that these low-count or unseen trigrams will never appear in natural speech. Various smoothing approaches have been proposed to tackle the data sparsity. Back-off is usually applied on unseen or low-count events, in which the N-gram probability is replaced by a scaled (N-1)-gram probability [46]. Another solution is to reduce counts of more frequent word sequences and therefore the resulting excess probability mass can be re-distributed amongst the less frequent word sequences [47]. This is referred to as discounting.

Perplexity can be utilized to evaluate the language models. In general, a reduction in perplexity results in improvement on speech recognition performance. The perplexity of a discrete probability distribution p is defined as:

$$Perplexity = 2^{H(p)} = 2^{-\sum_x p(x) \log_x p(x)} \tag{2.6}$$

where $H(p)$ is the entropy of the distribution. For a given N-gram language model, equation 2.6 can be expressed as:

$$Perplexity = P(w_1, ..., w_m)^{-\frac{1}{m}} = \prod_{i=1}^{M} P(w_i | w_{i-N+1}, ..., w_{i-1})^{-\frac{1}{m}} \tag{2.7}$$

2.1.4 Decoding

The purpose of decoding is to determine the optimal word sequence given a sequence of acoustic feature vectors. In LVCSR, words are usually decomposed into sub-word units such as phones, and each phone is typically modeled by HMM state sequence. As a result, the decoding is performed in a hierarchical way with three levels: state level, phone level and word level. Acoustic models, pronunciation dictionary, and language models are used to provide constraints during the decoding process.

Viterbi algorithm, as a token passing algorithm [48], is applied to search for the optimum word sequence. A token is referred to as a partial path from the

start to the current time instant through the decoding network. As the tokens propagate through the state-level network, a phone-level network is generated at first. The log-probability of the tokens is cumulated by the intra-transition probability inside the phone-level HMMs and the emission probability density of the observations given by the HMM. The pronunciation dictionary provides the transition probability from one phone to another phone, and consequently the word-level network can be generated. Finally, the optimal word sequence can be decoded by making use of the word-level transition probability given by language models.

In theory, the token-pass algorithm is guaranteed to decode the best possible pass. However, it would take too much time and space to decode. Therefore, various approaches are proposed to speed up the decoding process and reduce the search space. A tree-structured lexicon can reduce the potential search space significantly by merging the common prefixes of the words [49]. Beam search can be employed to further speed up the decoding process [50]. The pre-defined parameter beam-width, can be adjusted as a trade-off between speed and accuracy. Only the active tokens lying within the beam will be kept in memory during the decoding.

A multiple-pass decoding strategy is commonly used in Viterbi-based LVCSR systems. In the first pass, simple acoustic models and language models are employed to generate a reduced search space of the most likely hypothesis such as word-graph or N-best list. In the second pass, more accurate and complicated models can be used for re-scoring in the reduced-search space. The multiple-pass approach is useful for reducing computational effort. In addition, more complex language models or other high-level knowledge such as super-segmental information can be applied in the second-pass to improve recognition performance.

2.2 Multilingual Speech Recognition

A monolingual speech recognizer refers to a system that is designed to recognize speech from one particular language at a time. The acoustic model in this case is trained solely on data from one specific language; the pronunciation dictionary includes the phone sequence of that language only and also for the language model. The state-of-the-art in LVCSR has achieved great success over the last years for quite a number of languages.

With the globalization of today's world, more and more multilingual applications are needed. For example, an enquiry system in an international airport may have users from various countries who will speak different languages. Such system has to be able to recognize several languages. Research related to multilingual speech recognition has attracted much attention over the last decade. Many approaches are proposed, which can be summarized into three groups, i.e. simultaneous multilingual speech recognition, cross-lingual speech recognition and rapid language adaptation.

Simultaneous multilingual speech recognition systems usually have a multilingual acoustic model which consists of a collection of coexisting subsets of language dependent acoustic models. Two main strategies have been developed. The first strategy applies language identification beforehand, and then the speech recognizer of the identified language is activated to recognize the input speech utterances [21] [22]. The recognition performance in this case will be exactly the same as the monolingual system if a perfect language identifier is performed. The second strategy runs parallel recognizers simultaneously [51] [52] [53]. Such system performs an implicit language identification because the language identity can be determined according to the recognized words.

A cross-lingual speech recognizer applies a language-independent universal phone set for all languages. The crucial problem in cross-lingual speech recognition is to exploit acoustic-phonetic similarities across languages and design an appropriate phoneme inventory. Knowledge-based and data-driven methods are the two main approaches for phoneme merging. Various linguistic information can be used to assign phonemes into classes. The International Phonetic

Alphabet (IPA) and Speech Assessment Methods Phonetic Alphabet (SAMPA) are widely used as reference schemes in most knowledge-based systems [54] [23]. Data-driven approaches require a combination scheme and a similarity measure to decide which phonemes can be merged. Many systems make use of the phoneme confusion matrix and merge the most confusing phonemes into classes [55]. On the other hand, various distance measures are employed to calculate the acoustic similarity between phonemes [56].

Language adaptation and cross-lingual transfer are usually applied to port an existing recognizer to the new target language with very limited or even no training data available. The term cross-lingual transfer is usually preferred if no training data can be used for the target language. In this case, many studies show that multilingual transfer models outperform monolingual ones [57]. Language adaptation technique can be applied to adapt the pre-existing language-independent system toward a new target language, using only limited speech data from the target language [58]. It is found that language adaptation performance is highly correlated to the amount of data available for language adaptation. On the other hand, cross-lingual systems have proven to be more effective than monolingual ones [59].

2.3 Code-mixing Speech Recognition

Code-mixing refers to intra-sentential switching between two different languages or language varieties in spoken utterances. It is a common phenomenon in bilingual societies. In code-mixing, the major language is referred to as the primary language, or the matrix language, and the other language is the secondary language, or embedded language. Different combinations of languages are found in code-mixing, for example, Spanish-English in United States, German-Italian and French-Italian in Switzerland, and Hebrew-English in Israel [60]. In Taiwan, code-mixing between Chinese dialects, namely Mandarin and Taiwanese, has become common in recent years [61]. Hong Kong is an international city where many people, especially the younger generation, are Cantonese and En-

glish bilinguals. English words are frequently embedded into spoken Cantonese. In this case, Cantonese is the primary language, and English is the embedded language [62]. Table 2.1 lists some examples of code-mixing in several languages.

Table 2.1: Examples of code-mixing in several languages (the embedded language is emphasized in boldface)

Involved languages	Examples
Spanish-English	Siempre está **promising** cosas. (He is always promising things.)
French-Italian	No, parce que **hanno** donné des cours. (No, because they have taught courses.)
Arabic-French	tlabt wahdi l' **immigration**. (I asked alone for the immigration.)
Mandarin-Taiwanese	我們計劃去夜市吃 **蚵仔煎**。 (We plan to eat oyster omelet in the night market.)
Cantonese-English	份 **assignment** 今日下晝之前要交。 (We need to submit the assignment before this afternoon.)

Compared with monolingual and conventional multilingual speech recognition, code-mixing speech recognition is more challenging because of the unknown language boundaries. From the results reported in [28], there is significant degradation from monolingual ASR to code-mixing ASR. There have been two main approaches to code-mixing speech recognition [30] [28]. The first approach is similar to simultaneous multilingual speech recognition, which involves a language boundary detection (LBD) algorithm that divides the input utterance into language-homogeneous segments. The language identity of each segment is determined, and the respective monolingual speech recognizer is applied. In this approach, the performance of second-step recognition will be restricted to the performance of first-step LBD. If the performance of LBD can not be achieved 100% then it will directly degrade the ultimate recognition results for code-mixing speech. However, compared with the conventional language identification task, LBD in code-mixing is more difficult since the du-

ration of individual language segments is relatively shorter [26] [32]. The second approach aims to develop a cross-lingual speech recognition system, which can handle multiple languages in a single utterance. The acoustic models, language models, and pronunciation dictionary are designed to be multilingual and cover all languages concerned. Many previous studies showed that the latter one is more appropriate than the former due to the performance limitation of LBD.

2.4 Kullback-Leibler Divergence

Kullback-Leibler divergence (KLD) is an information-theoretic measure of (dis)similarity between two probability distributions [66]. It has been widely used in various applications. The KLD between two given distributions Q and R is defined as:

$$D_{KL}(Q||R) = \int q(x) \log \frac{q(x)}{r(x)} dx \tag{2.8}$$

where q and r denote the densities of Q and R. Then, the symmetric form of KLD between Q and R is:

$$\begin{aligned} D_{KL}(Q,R) &= D_{KL}(Q||R) + D_{KL}(R||Q) \\ &= \int q(x) \log \frac{q(x)}{r(x)} dx + \int r(x) \log \frac{r(x)}{q(x)} dx \end{aligned} \tag{2.9}$$

For two multivariate Gaussian distributions, equation 2.9 has a closed form:

$$\begin{aligned} D_{KL}(Q,R) =& \frac{1}{2} tr\{(\Sigma_r^{-1} + \Sigma_q^{-1})(\mu_r - \mu_q)(\mu_r - \mu_q)^T \\ &+ \Sigma_r \Sigma_q^{-1} + \Sigma_q \Sigma_r^{-1} - 2I\} \end{aligned} \tag{2.10}$$

where μ and Σ are corresponding mean vectors and covariance matrices, respectively.

In speech processing such as speech recognition and speech synthesis, a recurring problem is measuring the similarity of two given speech units, e.g. states, phones. Use of KLD has been discovered to be a useful measurement between hidden Markov models (HMMs) of acoustic models when the temporal structure of HMMs is aligned by dynamic programming [67] [68]. In this thesis, KLD will

be applied to investigate the acoustic similarity between Cantonese and English in different levels of acoustic units, viz phones, states and Gaussian mixture components.

□ End of chapter.

Chapter 3

Cantonese-English Code-mixing Speech Recognition

Summary

Hong Kong is a bilingual society, where Chinese and English are both official spoken languages. As a result of this bilingualism, mixing of Cantonese and English is very common in Hong Kong. This chapter begins with an introduction of the language situation in Hong Kong. Characteristics of Cantonese and English are described respectively. We shall then introduce the essence of Cantonese-English code-mixing. The nature of Cantonese-English code-mixing is investigated from a linguistic point of view first. After that major challenges in Cantonese-English code-mixing LVCSR are discussed. Finally, a series of experiments are performed to analyze the effect of language mixing on the recognition performance of Cantonese-English code-mixing utterances.

3.1 Speaking Habit in Hong Kong

Hong Kong is an officially bilingual territory. Historically, Hong Kong was a British colony from 1840 to 1997, and English was the sole official language of Hong Kong between 1883 and 1974. Nowadays, both English and Chinese are the official languages as defined in the Basic Law of Hong Kong. Unlike other racially homogeneous cities in Britain and other Western countries, more than 95% of the population in Hong Kong is Chinese [69]. Cantonese, as the mother tongue of most Hong Kong residents, is widely used in daily communications. Although English is the usual language of only 1% of people, it is still taught in schools and spoken by over 30% of the population [70]. As a major working language in Hong Kong, English is commonly used in education, commercial activities and legal matters.

As a result of the bilingualism, code-mixing between Cantonese and English is very common in Hong Kong. In this particular situation, Cantonese is the primary language, also known as the matrix language, and English is the secondary language, usually referred to as the embedded language [62]. Cantonese-English code-mixing speech normally follows the Chinese grammar and syntax, and the English word is usually used as a substitute for its Chinese equivalent, although word order sometimes may change due to the parts-of-speech (POS) of the code-switched words.

3.1.1 Characteristics of Cantonese

Cantonese is one of the major Chinese dialects spoken by tens of millions of people in southern China, Hong Kong, Macau and many overseas Chinese communities [71].

Cantonese Phonology and Phonetics

Similar to Mandarin (Putonghua), Cantonese is a tonal and monosyllabic language of the Sino-Tibetan language family. The basic unit of written Cantonese and Mandarin is the character. Each character in written Cantonese is pro-

nounced as a single syllable carrying a specific lexical tone. If the tone changes, the lexeme has a different meaning. On the other hand, Cantonese is homophonic and homographic. A character may have multiple pronunciations and different characters may share the same pronunciation as well. Table 3.1 shows some examples of homographs and homophones in Cantonese.

The general syllable structure of Cantonese is C1-V-C2, where C1 and C2 are consonants, and V is either a vowel or a diphthong. Since C1 and C2 are optional, all Cantonese syllables take the forms V, C-V, C-V-C or V-C [72]. As shown in Table 3.2, each syllable can be divided into an Initial (C) and a Final (V-C). There are more than 600 legitimate Initial-Final combinations, i.e., base syllables. There are six tones in Cantonese. If tonal difference is considered, the total number of distinct syllables is about 1,800.

Table 3.1: Examples of homographs and homophones in Cantonese

Homographs	樂	/ngaau6/, /lok6/, /ngok6/
Homophones	/ji1/	衣，依，醫，伊，漪，椅

Table 3.2: Composition of Cantonese syllables

<table>
<tr><td colspan="4">Tonal syllable (1761)</td></tr>
<tr><td colspan="3">Base syllable (625)</td><td rowspan="3">6 Tones</td></tr>
<tr><td>Initial (19)</td><td colspan="2">Final (53)</td></tr>
<tr><td>Onset (19)</td><td>Nucleus (20)</td><td>Coda (6)</td></tr>
</table>

Cantonese initials can be classified into five classes according to the manner of articulation. They are plosives, affricates, fricatives, nasals and approximates, while the former three classes are unvoiced and the other two classes are voiced. Cantonese finals can be divided into five categories: vowels, diphthongs, vowels with a stop coda, vowels with a nasal coda, and syllabic nasals [73]. Each final contains at least one vowel element except for the syllabic nasals /m/ and /ng/. Eleven different vowels are found in Cantonese. Seven of them can appear independently and the other four are usually followed by consonant codas or other vowels. Ten different diphthongs are found in Cantonese and all of them end with /i/ or /u/. There are six different codas in Cantonese. Three of them

Cantonese Initials		
Plosive		
b	p	[p a] (爸)
d	t	[t a] (打)
g	k	[k a] (加)
gw	k^w	[k^w a] (瓜)
p	p^h	[p^h a] (扒)
t	t^h	[t^h a] (他)
k	k^h	[k^h a] (卡)
kw	k^{wh}	[k^{wh} a] (誇)
Fricative		
s	s, ʃ	[s a] (沙)
		[ʃ y] (書)
f	f	[f a] (花)
h	h	[h a] (蝦)
Affricate		
z	ts, tʃ	[ts ɪ] (之)
		[tʃ y] (朱)
c	ts^h, $tʃ^h$	[ts^h ɪ] (痴)
		[$tʃ^h$ y] (處)
Approximate		
l	l	[l a] (啦)
w	w	[w a] (蛙)
j	j	[j ɐu] (憂)
Nasal		
m	m	[m a] (媽)
n	n	[n a] (拿)
ng	ŋ	[p^h a ŋ] (烹)

Cantonese Finals					
Vowel			**Vowel-Nasal**		
ɪ	i	[s i] (絲)	m	im	[t im] (點)
yu	y	[ʃ y] (書)	n	in	[p ɪn] (邊)
u	u	[f u] (夫)	ng	iŋ	[p iŋ] (兵)
e	ɛ	[s ɛ] (借)	yun	yn	[t yn] (端)
oe	œ	[h œ] (靴)	un	un	[p un] (搬)
o	ɔ	[s ɔ] (梳)	ung	ʊŋ	[s ʊŋ] (鬆)
aa	a	[s a] (沙)	eng	ɛ ŋ	[h ɛ ŋ] (輕)
			eon	ɵn	[t ɵn] (吞)
			oeng	œŋ	[l œŋ] (良)
			on	ɔn	[ɔn] (安)
Syllabic Nasal			ong	ɔŋ	[p ɔŋ] (邦)
			am	ɐm	[ɐm] (啥)
m	m	[m] (唔)	an	ɐn	[f ɐn] (婚)
ng	ŋ	[ŋ] (吳)	ang	ɐŋ	[p^h ɐŋ] (朋)
			aam	am	[s am] (衫)
			aan	an	[s an] (山)
			aang	aŋ	[s aŋ] (生)
Diphthong			**Vowel-Stop**		
ui	ui	[f ui] (灰)	ip	ip	[j ip] (業)
ei	ei	[h ei] (稀)	it	ɪt	[j ɪt] (熱)
eoi	ɵy	[s ɵy] (衰)	ik	ɪk	[j ik] (益)
oi	ɔɪ	[s ɔi] (鰓)	yut	yt	[t^h yt] (脫)
ai	ɐi	[s ɐi] (西)	ut	ut	[f ut] (闊)
aai	aɪ	[w aɪ] (威)	uk	ʊk	[f ʊk] (福)
iu	ɪu	[s ɪu] (燒)	ek	ɛ k	[p ɛ k] (畀)
ou	ou	[s ou] (鬚)	eot	ɵt	[s ɵt] (術)
au	ɐu	[s ɐu] (收)	oek	œk	[l œk] (略)
aau	au	[s au] (筲)	ot	ɔt	[h ɔt] (喝)
			ok	ɔk	[h ɔk] (殼)
			ap	ɐp	[s ɐp] (濕)
			at	ɐt	[s ɐt] (失)
			ak	ɐk	[s ɐk] (塞)
			aap	ap	[s ap] (圾)
			aat	at	[s at] (撒)
			aak	ak	[s ak] (索)

Figure 3.1: List of Cantonese initials and finals. Jyut Ping symbols, IPA symbols and examples are listed on the left, middle, and right respectively.

are nasal codas /m/, /n/, and /ng/, while the other three are stop codas /p/, /t/, and /k/, and they are are unreleased. Figure 3.1 lists 19 Cantonese initials and 53 finals. They are labelled with both Jyut Ping [74] and IPA symbols for ease of comparison. Examples are also provided for each phoneme.

Written Cantonese vs. Spoken Cantonese

Cantonese is a spoken Chinese dialect that the formal or written (standard) form is significantly different from the spoken or colloquial form. Standard Chinese is the official written language in mainland China, Taiwan and Hong Kong. Cantonese speech, when being written down, shows substantial differences from standard Chinese. The lexicons of standard Chinese and spoken Cantonese are quite different. Different words may be chosen to represent the same meaning. For example, "差人" in spoken Cantonese and "警察" in standard Chinese. Moreover, written Cantonese is neither taught in schools nor recommended for official and documentary usage, such that some of the spoken Cantonese words do not have a standard written form. For example, the lexicon which means "yesterday"can be written in four different forms — 尋日, 噚日, 琴日, 擒日.

On the other hand, there is a consensus that many daily used colloquial function words, which are not found in standard Chinese, are not allowed in standard Chinese writing. Figure 3.2 shows some of these examples. Cantonese speech, with several colloquial function words and spoken lexical items, are referred to as colloquial Cantonese, and are usually used in casual situations. When such Cantonese speech is written down, it may be unintelligible to Mandarin speakers [75]. Figure 3.3 shows examples of standard Chinese, and colloquial Cantonese. In addition, an example of Cantonese-English code-mixing is also given in the figure for comparison purposes. The three sentences have the same meaning.

Examples of colloquial Cantonese terms
佢, 嘅, 嗰, 喺, 哋, 啲, 乜, 冇, 咩, 哂, 嚟, 睇, 咁

Figure 3.2: Examples of colloquial Cantonese terms

Standard Chinese: 我做完了三份功课，终于可以睡觉。

Colloquial Cantonese: 我做晒三份功課，終於可以瞓喇。

Code-mixing: 我 **finish** 咗三份功課，終於可以瞓喇。

English: I finished three assignments, and can finally sleep.

Figure 3.3: The differences between standard Chinese, colloquial Cantonese and Cantonese-English code-mixing, for the same meaning.

3.1.2 Characteristics of English

English is the leading language in international communications. It is the third most natively spoken language in the world after Mandarin Chinese and Spanish and is spoken by about 400 million people. It is also widely learned as a second language by more than 1,000 million people and is used as an official language in many world organizations [76].

Different from Mandarin and Cantonese, English is a stress-timed language in the Indo-European family. Stressed syllables in English usually appear at a roughly steady tempo, as well as being longer and having a higher intensity and pitch, while non-stressed syllables are shortened and accommodate stressed syllables. Two English words can be distinguished by stress for example, the words ***des****ert* and *dess****ert***, and the noun/verb pairs ***rec****ord* and *rec****ord***.

The composition of syllables in English has more variations than Cantonese. Although about 80% of the syllables in English also take the C1-V-C2 structure, the remaining 20% could be C, CC, CCV, VCC, CCCV, CCCVCC, etc. [77]

There are 24 consonants, 11 monophthongs and 3 diphthongs in general American English. Figures 3.4 and 3.5 list the IPA symbols of these consonants and vowels, respectively.

3.1.3 Lazy Tongue and Phone Change

There exist many pronunciation variations in colloquial Cantonese speech, especially when the speaking rate is fast. Speakers may not follow strictly the

	Bilabial	Labiodental	Dental	Alveolar	Post-alveolar	Palatal	Velar	Labial-Velar	Glottal
Plosive	p b			t d			k g		
Affricate					tʃ dʒ				
Nasal	m			n			ŋ		
Fricative		f v	θ ð	s z	ʃ ʒ				h
Approximant				ɹ		j	w		
Lateral Approximant				l					

Figure 3.4: IPA table, English consonant (General American)

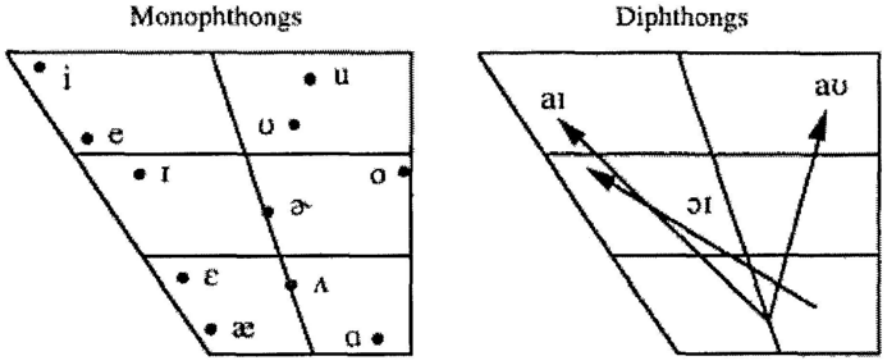


Figure 3.5: IPA table, English vowels and diphthongs (General American)

pronunciations as specified in a standard dictionary. Firstly, some sounds are difficult to pronounce and they are replaced by similar sounds which take less effort to pronounce. This phenomenon is called lazy tongue (懶音). For example, /n/ initial replaced by /l/ ("你" /**n**ei/ becomes /**l**ei/); /gw/ initial reduced to /g/ ("國" /**gw**o/ becomes /**g**o/); /ng/ initial disappeared ("我" /**ng**o/ becomes /o/). Syllable fusion is another type of pronunciation variation that usually occurs in fast speech, i.e., the initial consonant of the second syllable of a disyllabic word tends to be omitted or changed [78][79]. For instance, "今日" /gam **j**at/ may be pronounced /gam **m**at/, and "即刻" /zik **h**ak/ becomes /zik **k**ak/.

In general, English spoken by non-native speakers carries accents that are determined by their mother tongue. In Hong Kong, the mother tongue of the speaker is Cantonese. It is inevitable that their spoken English words carry a Cantonese accent to a certain extent. In many cases, the syllable structure of an

English word changes to follow the structure of legitimate Cantonese syllables [80]. Such changes usually involve phone insertions or deletions. For example, the second consonant in a CCVC syllable of English may be softened, e.g., the word "plan" is pronounced /p ae n/ instead of /p **l** ae n/ by many Cantonese speakers. A monosyllabic word with the CVCC structure may become disyllabic by inserting a vowel at the end, e.g., /f ae n **z**/ ("fans") becoming /f ae n **s i**/. It is also noted that the final stop consonant in an English word tends to be softened or dropped, e.g., /t eh s **t**/ ("test") becoming /t eh s/. This is related to the fact that the stop coda of a Cantonese syllable is unreleased [73]. In addition to phone insertion and deletion, there also exist phone changes in Cantonese-accented English. That is, an English phoneme that is not found in Cantonese is replaced by a Cantonese phoneme that people consider to sound similar. For example, /**th** r iy/ ("three") becomes /**f** r iy/ in Cantonese-accented English. Furthermore, Cantonese speakers in Hong Kong sometimes create a Cantonese pronunciation for an English word. For example, the word "file" (/f ay **l**/) is transliterated as /f ay **l o**/ (快佬 in written form). It is not a straightforward decision whether such a word should be treated as English or Cantonese. This is known as "lexical borrowing" [81].

3.2 Cantonese-English Code-mixing

According to John Gumperz [82], code-switching refers to "the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-system". In order to describe this phenomenon more precisely by the frequency of language switching, two terms, (inter-sentential) code-switching and (intra-sentential) code-mixing, are utilized. Switching above clause level is called code-switching, and switching at word level is called code-mixing. In Hong Kong, code-switching mainly occurs in the word level and switching involving linguistic units above the clause level is rare. Hence the term code-mixing is usually preferred to describe this language behaviour of Hong Kong bilinguals.

As one of the major language choices besides monolingual Cantonese and English, the appearance and frequency of Cantonese-English code-mixing mainly depends on the speaking style, the conversation domain and the language proficiency of the speakers. Firstly, code-mixing is usually regarded as informal, and therefore people seldom use it in formal speech or presentations. Besides, many elderly people usually have lower English proficiency. As a result, people tend to use monolingual Cantonese when they are speaking to their parents or their seniors. Thirdly, code-mixing is a domain-specific phenomenon. It mainly happens in domains that are related to other cultures or languages, where many of the terms are new lexicon that may not appear in Cantonese. In such cases, people need to translate it or just use the English term directly [83]. In Hong Kong, code-mixing is wide spread in five domains as shown in Table 3.3 [80].

Table 3.3: Examples of code-mixing in several domains

Domain	Examples
Computer discourse	我今朝先收到封 **email**。
Business discourse	份 **contract** 好似有少少問題。
Food discourse	點解個 **menu** 入面冇 **tiramisu** 啊？
Fashion discourse	你知唔知依家**Burberry** 做緊 **discount** 喔。
Showbiz discourse	方大同唱歌算係 **R&B** 定係 **soul music**？

3.2.1 Motivations behind Code-mixing in Hong Kong

There are some English terms which do not have Cantonese equivalents. In such case code-mixing is the only choice. However, in more cases, people are able to choose between monolingual Cantonese and Cantonese-English code-mixing. There are many discussions on the possible motivations behind code-mixing, especially from sociolinguistic and psycholinguistic perspectives. Many researchers have suggested that code-mixing is a conscious behaviour intended for certain sociolinguistic functions [84] [85] [86] [82]. Major reasons for code-mixing are summarized as follow.

Quotation

As reported in [82] [87] [88], quotation is one of the major functions in code-mixing. When we report other people's words, it seems possible that it will be reported in the original code for convenience and to avoid misunderstanding.
Example: 有個朋友問我: "What do you think?"

Emphasis or avoidance of repetition

In some cases, the embedded words in code-mixing are the translations or near equivalents of appeared lexical items in the matrix language. This kind of cross-language reiteration may be performed for (a) emphasis purposes [89]. It is also possible that the use of a translation equivalent is only for (b) avoidance of repetition.
Example (a): 好多同學放學都會去 grammar school, 補習社呀。
Example (b): 依家買 iphone 有職員優惠，好多中大 staff 都買左。

Availability and specification

Availability and specification are important functions in Cantonese-English code-mixing. Sometimes, English words and their Cantonese equivalents may not have exactly the same meaning, and therefore people may fail to find a suitable lexicon in Cantonese. As a result, English words will be applied to give more appropriate meaning when they want to specify or generalize something. For example, the English verb *book* means making a reservation without money or a deposit. Its nearest equivalent in Cantonese is 訂, which cannot specify whether a deposit is required or not. Therefore people usually prefer to use the more specific word *book* in some cases.
Example: 西班牙有間分子料理店好似要提前一年 book。

Principle of least effort

Compared with Cantonese, sometimes English expressions are shorter and more convenient. People may use the English abbreviation instead of long Cantonese phrases to reduce linguistic effort. For example, people prefer *ABC* (美國出生的

華人, American born Chinese), *FYP* (畢業設計, final year project), *GPA* (平均積點, grade point average), etc. On the other had, people usually prefer English words with less syllables instead of Cantonese terms with more syllables such as *hall* (學生宿舍), *budget* (財政預算), etc [90]. Moreover, people sometimes may use the first one or two syllables of the English words instead of the whole words. Typical examples include *reg* (register, 註冊), *con* (contact lens, 隱形眼鏡) and *pro* (professional, 專業)。

Example: 我間學校要一年級先可以住 hall。

你 MCC(矇嗰嗰) 咁，有冇帶 con 啊！

Euphemism

One of the reasons for using English terms is the preference for a euphemism. In many situations, people may use English words if they find themselves in an embarrassing position of saying such terms in Cantonese.

Example: 請問邊度有 toilet 啊?

Identity marking

Speech as one of the important markers of social identity, can be used to mark social characteristics such as social status, education status, occupation, as well as regional affiliation. Many researchers suggest that code-mixing can be used for in-group identification [91] [92] [93]. This is because people may have different code choices when relationships between participants or functions changes.

Interjection

Finally, Gumperz introduces interjection as one of the functions of code-mixing [82]. However, this function is not very common in Cantonese-English code-mixing. Only a few examples can be found.

Example: I mean, 我今晚冇時間!

3.2.2 Nature of Code-mixing English Units

Cantonese-English code-mixing is not a simple insertion of one language into another. It comes with a lot of phonological, lexical, and grammatical variations with respect to the monolingual speech spoken by native speakers. However, we have little understanding about it. To better understand this highly dynamic language phenomenon, characteristics of Cantonese-English code-mixing is investigated from a linguistic point of view in this research.

This linguistic study is based on code-mixing text data, which are collected from newspapers, magazines and online diaries from the internet. The original code-mixing text is retrieved with colloquial Cantonese terms and English words. If the collected sentence contains words which are only used in standard Chinese, they are removed. As a result, Cantonese in collected code-mixing data can be regarded as colloquial, rather than standard written Cantonese or formal Cantonese. In addition, meaningless sentences are filtered out and duplicated sentences are removed. The remaining sentences are expected to preserved the linguistic characteristics of Cantonese-English code-mixing speech. There are about 65,000 code-mixing sentences in total. They contain more than 1 million Cantonese characters and 100 thousands English words.

We first analyze the English appearance frequency in code-mixing sentences, and the results are shown in Table 3.4. It can be seen that, more than 95% of code-mixing utterances include one to two embedded English segments. In addition, it is also found that most English segments contain a single English word only, which is consistent with the observations by Tse [94]. Table 3.5 lists the number of English words in each code-mixing sentence with single English segment.

Table 3.4: Number of English segments per code-mixing sentence

No. of English segments	1	2	>= 3
No. of code-mixing sentences	53,310	10,517	1,697
Percentage (%)	81.36%	16.05%	2.59%

Table 3.5: Number of English words per code-mixing segment

No. of English words	1	2	3	>=4
No. of sentences	45,634	6,481	833	363
Percentage (%)	85.60%	12.16%	1.56%	0.68%

Code-mixing English vs. Native English

English word frequencies in Cantonese-English code-mixing and monolingual English are quiet different. In this study, the monolingual English statistics are based on the trillion-word English data from "Web 1T 5-gram Version 1", which is the largest English plain text corpus collected by Google [95] [96]. It is found that the most common 50 monolingual English words are mainly prepositions and pronouns. However, English function words are rarely found in our Cantonese-English code-mixing data. Table 3.6 shows the most common 10 words in monolingual English and code-mixing English. It can be seen that all of them are different.

Table 3.6: The most common 10 words in monolingual English and code-mixing English

Monolingual English			Code-mixing English		
the	22,914,473,646	2.24%	ok	1,725	1.82%
of	12,765,289,150	1.23%	blog	1,647	1.74%
and	12,522,922,536	1.20%	sales	1,620	1.71%
to	11,557,321,584	1.11%	post	1,291	1.36%
a	7,841,087,012	0.75%	friend	1,087	1.15%
in	7,490,628,883	0.72%	cheap	945	1.00%
for	5,357,090,483	0.51%	check	865	0.91%
is	4,551,580,934	0.44%	bb	859	0.91%
on	3,436,213,533	0.33%	book	751	0.79%
that	3,244,802,211	0.31%	set	652	0.69%

On the other hand, although there are more than 200 thousand English words in the Google corpus, only about 4,000 different English words can be

found in our Cantonese-English code-mixing text data. We also notice that most code-mixing English words are rare. Almost half of the words appear less than 4 times. However, the most 10 common words cover more than 10% of the English words in the code-mixing text corpus, the top 100 cover over 1/3 and the top 1/3 cover more than 90%. It reveals that code-mixing is focused on a small number of specific English words.

Word Classes

All English words in our code-mixing text data are labelled by 17 POS classes The distribution is given in Table 3.7. This finding is consonant with observations in previous studies by B.-H.-S. Chan [97] and David C.-S. Li [80] in that most common code-mixing English is nouns, followed by verbs and adjectives. Table 3.8 lists the most 10 common nouns, verbs, and adjectives. They account for 17%, 30%, and 30% among the English nouns, verbs and adjectives in collected code-mixing data respectively.

Table 3.7: POS distribution of English words in Cantonese-English code-mixing

N	58,031	59.43%	Num	233	0.24%
V	21,156	21.67%	V Phrase	231	0.24%
Adj	8,407	8.61%	Clause	255	0.26%
Adv	2,406	2.46%	Adj Phrase	221	0.23%
Int	2,310	2.37%	Art	143	0.15%
Prep	1,461	1.50%	Prep Phrase	78	0.08%
Conj	1,244	1.27%	Conj Phrase	44	0.05%
Pron	774	0.79%	Int Phrase	16	0.02%
N Phrase	628	0.64%			

Further analysis is performed on the code-mixing sentences with two English segments. It is found that the most common mixing type is noun-noun mixing, followed by verb-noun/noun-verb mixing and verb-verb mixing. We list the most 10 common mixing types in Table 3.9.

As the most frequently embedded word class, more than 2,000 distinct En-

Table 3.8: The common English nouns, verbs, and adjectives found in Cantonese-English code-mixing

Nouns	blog, sales, post, friend, bb, game, email, lunch, msn, printer
Verbs	check, book, set, send, test, feel, join, update, keep, cut
Adjective	cheap, high, good, full, nice, fit, free, junior, sharp, cool

Table 3.9: Common mixing type in Cantonese-English code-mixing

N_N mixing	2,876	39.13%	Adj_N mixing	250	3.40%
V_N mixing	1,216	16.55%	N_Int mixing	139	1.89%
N_V mixing	922	12.55%	Adj_Adj mixing	130	1.77%
V_V mixing	521	7.09%	V_Adj mixing	95	1.29%
N_Adj mixing	315	4.29%	Adv_N mixing	92	1.25%

glish nouns are found in our data. We observe that more than 20% of the nouns are the terminologies in computer & technology domain. Besides, names are also very common in Cantonese-English code-mixing. There are all kinds of names, such as names of people, places, films, songs, events, books, companies, brands and so on. Sometimes the names are cited in their original language because they have not been translated into Chinese. There are other cases in which people cite the original name although the translated Chinese name is available.

Compromise Forms

According to Clyne [86], compromise form refers to code-mixing syntactic units which are governed by higher-order units. It can indicate how English and Cantonese are intertwined in code-mixing. In our data, plenty of examples of compromise forms are found. They are further subdivided according to their internal structure. It is found that N-N, V-N and Adj-N compromise forms are common underlying code-mixing structures. Many examples can be found. Table 3.10 shows typical N-N, V-N, and Adj-N compromise forms, which appear dozens of times in our data.

Table 3.10: Typical N-N, V-N, and Adj-N compromise forms in Cantonese-English code-mixing

N-N	V-N	Adj-N
van仔	寫blog	紅van
sales姐	做sales	女sales
blog友	食lunch	舊post
bb衫	上youtube	新version
model公司	出trip	冇mood
usb線	唱k	其他forum
iq題	落order	cheap機
sim卡	hang機	budget有限
cd機	book位	
pe堂	gel頭	

Table 3.11: Code-mixing between English free morphemes and Cantonese bound morphemes

Cantonese aspect markers	Code-mixing examples	Cantonese classifiers	Code-mixing examples
住	keep住, hold住	個	個post,個mon(monitor)
緊	plan緊, circulate緊	班	班friend, 班fans
定	book定, set定	隊	隊band, 隊team
過	update過, confirm過	部	部printer, 部pc
埋	join埋, send埋	件	件tee, 件cake
到	feel到, check到	張	張form, 張coupon
下	search下, test下	篇	篇blog, 篇paper
完	download完, print完	條	條thread, 條banner
好	tune好, pack好	架	架van, 架piano

Cantonese is an isolating language in that words are not marked by morphology to show their role in a sentence. Many bound morphemes in English can not be found in Cantonese. As a result, mixing between English free morphemes and Cantonese bound morphemes is very common in Cantonese-English code-mixing. Table 3.11 shows the 10 most typical Cantonese aspect markers, which are usually mixed with English verbs to represent the tense. Besides, mixing between English nouns and Cantonese classifiers is also very common in Cantonese-English code-mixing, and typical examples are listed in Table 3.11.

Hong Kong Style English & Borrowing

In Cantonese-English code-mixing, some English words are highly influenced by the Cantonese language and consequently adopt Cantonese morphology and its grammar system. This ungrammatical or nonsensical English in Cantonese contexts is called Hong Kong style English. Table 3.12 includes some examples of these words:

Table 3.12: Examples of Hong Kong style English

Code-mixing	Cantonese	English meaning
cheap cheap 地	便便地	cheaply
check (一) check	查(一)查	check it
mind 唔 mind	介唔介意	mind or not?
sup 唔 support	支唔支持	support or not?
on 返 line	上返網	online again
thank 一個 you	多一個謝	thank you
so 咩 rry	對咩唔住	why apologize?

In addition, there are some new lexicons which do not appear in traditional English speech, but are unique to code-mixing. Some of these words exist in traditional English, but have a totally different meaning [98]. For example, people prefer the first one or two syllables of English words instead of the whole words in code-mixing, such as the word "bio" (biology), "arche" (archeology), "mon" (monitor), "sem" (semester), "U lib" (university library), etc.

The distinction between code-mixing and borrowing is discussed in a number of papers [99][100][101]. In general, most scholars agree that if the embedded words have been fully integrated into the phonology and morphosyntax of the matrix language, it is regarded as lexical borrowing. In Hong Kong, borrowing is commonly used by many local people. For some common borrowing lexicons, the English words can be written in Cantonese characters with similar sounds. For example, the word "order" (/ao r d er/) is transliterated as /ao r d aa/ (柯打 in written form). In this case, it is not easy to make a straightforward decision whether such a word should be treated as English or Cantonese. This may because some of the words are so common that people don't realize that they come from English. Table 3.13 lists some typical borrowing words [102].

Table 3.13: Example of lexical borrowing

Borrowed term in Cantonese	English words
巴士	bus
的士	taxi
吉他	guitar
爹地	daddy
士多	store
多士	toast

In our study, the borrowed term will be regarded as a Cantonese lexicon if it can be written in Cantonese, and the written form can be further found in the Cantonese text database. In contrast, it will be regarded as English if this lexical item does not have a well-accepted Cantonese written form.

3.2.3 Syntactic Constraints of Cantonese-English Code-mixing Sentences

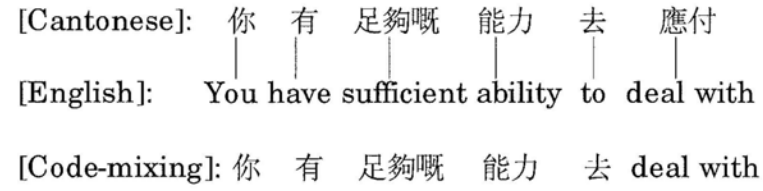
During the past three decades, many studies have been carried out on the syntactic constraints in code-mixing. The most comprehensive research is that of Snakoff & Poplack and Di Sciullo *et al.* They suggested three major syntac-

tic constrains in code-mixing, including equivalence constraint, free morpheme constraint and government constraint [103][104][105].

Equivalence Constraint

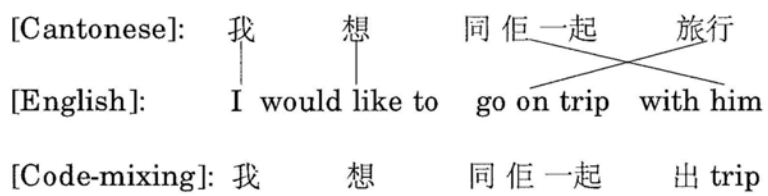
According to the equivalence constraint, code-mixing will tend to take place when surface structures of Cantonese and English map onto each other [104][106]. Many studies discuss this constraint in various code-mixing sceneries. Different observations are found in different language combinations. Clyne agreed with this constraint in his research on German-Dutch code-mixing [107], while Bokamba claimed that the equivalence constraint may fail in code-mixing between English and some African languages [108].

In our data, it is found that most of the alternation of languages occurs at points where the juxtaposition of Cantonese and English elements does not violate the syntactic rule of either language such as:

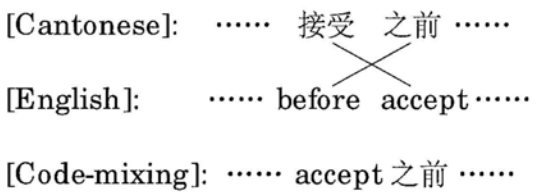


Sometimes, Cantonese-English code-mixing may still occur when the surface structure is unique to only one language. In this situation, the code-mixing sentence will follow the syntactic structure of Cantonese. Some examples are found in our data. We list typical cases as follows.

Firstly, Cantonese and English are typical languages which can follow a SVO (Subject Verb Objective) pattern. However, some Cantonese sentences may follow a SOV (Subject Objective Verb) structure which does not appear in English. In such case, code-mixing sentences will follow a SOV construction as well:



Besides, vocabulary related to time and space usually occur before the verb in Cantonese sentence. However, this type of words normally appears after the verb in English. For example:



Free Morpheme Constraint

According to Snakoff and Poplack, the principle of the free morpheme constraint is that language switching may not occur between a bound morpheme and a lexical form unless the latter has been phonologically integrated into the language of the bound morpheme. This is an arguable constraint. The effectiveness is found to vary from language to language. Although it was proven in Spanish-English code-mixing [104], Clyne and Bokamba point out that this constraint may fail for German-Dutch code-mixing and Kiswahili-English code-mixing [107][108].

In our study, it is found that the free morpheme constraint is not appropriate in Cantonese-English code-mixing. Plenty of counterexamples are found. As we discussed before, quite a few compromise forms that consist of an English verb (free morpheme) and Cantonese aspect marker (bound morpheme) are found in Cantonese-English code-mixing. Besides, other Cantonese bound morphemes such as privative marker and adverb marker are also commonly found in our code-mixing data for example: 好enjoy, 太cheap, access唔到, 唔fair, etc.

Government Constraint

The government constraint for code-mixing was formalized by Di Sciullo *et al.* in 1986 [105]. According to this constraint, the alternation of language is prohibited between two elements when there is a government or selection relation holding between them. However, in our study on Cantonese-English code-mixing, it is found that the government constraint fails badly for a large

number of cases. For instance, in most of the V-N code-mixing compromise forms (e.g. quit宿, 做facial), there is a government relationship between the verb and noun, and therefore the alternation of languages should be prohibited in such case according to the government constraint. But in fact, the V-N compromise form is one of the most typical patterns in our code-mixing data.

3.3 Difficulties in Cantonese-English Code-mixing LVCSR

For code-mixing speech recognition, the input utterance contains both Cantonese and English. They are quite different in terms of phoneme inventories and phonological structures. Besides, there exist many pronunciation variations in code-mixing speech such that the standard pronunciation dictionary will be inaccurate. In addition, new lexicons unique to code-mixing speech may not be found in monolingual dictionaries. Finally, it is practically difficult to collect sufficient code-mixing data for effective acoustic modeling and language modeling.

3.3.1 Phonetic & Phonology Differences

Cantonese and English have different phoneme inventories. Some phones appear in both languages, while the other does not. We use IPA to facilitate an intuitive comparison between Cantonese and English phonemes. Some of the phones in Cantonese and English are labelled with the same IPA symbols by phoneticians. These phones are expected to be phonetically very close to one another and they are merged into the same phone classes. Figure 3.6 shows the results. Cantonese phonemes are on the left and English phonemes are on the right. The phones that exist in both Cantonese and English are in the middle. There are 65 phones in total. Seventeen phones are shared between two languages. Twenty-two are English-specific and the remaining 26 are Cantonese-specific.

In code-mixing speech recognition, the acoustic models are expected to cover

Cantonese	Shared	English
ɐ, ɵ, ɔ, y, œ, a, ai, ɵy, ei, ɛu, au, ɔi, ou, ɐi, ui, iu, ɐu, $tʃ^h$, k^h, k^w, p^h, t^h, ts, ts^h, k^{wh}	ɪ, i, ɛ, ʊ, f, t, tʃ, n, s, ʃ, l, j, k, ŋ, w, u, p, m,h	e, æ, ɚ, o, ɑ, ʌ, b, v, θ, ð, d, z, ɹ, dʒ, ʒ, g

Figure 3.6: IPA-based phoneme inventories of Cantonese and English

all possible phones in the two languages. There are two possible approaches: (1) monolingual modeling with two separate sets of language-specific models; and (2) cross-lingual modeling with some of the phoneme models shared between the two languages. Monolingual modeling has the advantage of preserving the language-specific characteristics and is most effective for monolingual speech from native speakers. In code-mixing speech where the English words are Cantonese-accented, an English phoneme tends to resemble or even become identical to a Cantonese counterpart. In this case, we may treat them as the same phoneme and establish a cross-lingual model to represent them. As shown in Figure 3.6, Cantonese and English have a number of phonemes that are phonetically identical or similar to each other. The degree of similarity varies. In principle, cross-lingual modeling can be applied to highly similar phonemes, while language-specific models are more appropriate if the phonetic variation is relatively large. In the next chapter, we will compare the effectiveness of cross-lingual and mono-lingual acoustic modeling and try to establish an optimal phoneme set for code-mixing speech recognition.

Besides, there is also a great difference between Cantonese and English in phonological level. As we discussed before, all Cantonese syllables follow the CVC structure, while English has a more complicated structure. Consecutive consonants which are not found in Cantonese may occur in English. As a result, the consecutive consonants context-dependent models may be under-trained if the amount of Cantonese training data and English training data is different in large proportion, and therefore the overall recognition accuracy may be degraded.

3.3.2 Accent & Lexicon

The pronunciation dictionary for code-mixing speech recognition is a mixture of English and Cantonese words. Each word may correspond to multiple pronunciations, which are represented in the form of phoneme sequences. Due to the effect of the Cantonese accent, the English words in code-mixing speech are subject to severe pronunciation variations as compared to those in standard English spoken by native speakers. This may introduce degradation into speech recognition accuracy since the phone sequence is greatly different from those in the standard dictionary. As a result, it is essential to reflect such variation in the dictionary. On the other hand, as discussed in Section 3.1.3, the common pronunciation variations in colloquial Cantonese should also be included. In this research, in order to build a better code-mixing dictionary, pronunciation variation in Cantonese-English code-mixing speech will be investigated in a systematic way. The details will be described in the next chapter.

On the other hand, there are some new lexicons which do not appear in traditional English speech, but are unique to code-mixing. These new lexicons should be included in the code-mixing dictionary as well.

3.3.3 Lack of Code-mixing Data

In our application, the most common type of code-mixing is where one or more Cantonese words in the utterance are replaced by the English equivalent. The grammatical structure of code-mixing sentences is based largely on that of monolingual Cantonese. Word n-gram is by far the most commonly used technique for language modeling in LVCSR. To train a set of good n-gram models, a large number of spoken materials in computer-processable text format are needed. This presents a great challenge to our research since it is difficult in practice to find such materials for code-mixing speech. For the training of acoustic models, we need a large amount of code-mixing speech data as well. The development of speech and text corpora is therefore an important part of the work on code-mixing ASR.

3.4 Effect of Language Mixing for Cantonese-English Code-mixing ASR

While automatic speech recognition of either Cantonese or English alone has achieved a great degree of success, recognition of Cantonese-English code-mixing speech is not as trivial. The difficulties encountered in code-mixing speech recognition were discussed in the last section. Previous studies have also reported that there is significant degradation from monolingual ASR to code-mixing ASR [63]. Before making efforts to improve recognition performance, we first attempt to analyze the effect of language mixing on recognition performance of code-mixing utterances, particularly on how and why the degradation is caused specifically by code-mixing.

The analysis is mainly based on speech recognition performance. As a result, IPA-based cross-lingual acoustic models are designed to perform baseline recognition experiments. The details of the phoneme inventory can be found in Figure 3.6. The acoustic models are tri-phone HMMs trained by 29 hours of monolingual Cantonese and Cantonese-English code-mixing utterances from the CUSENT [109] and CUMIX [25] databases. The details for these corpora will be introduced in the next chapter. The acoustic feature vector consists of 12 MFCC coefficients, log energy, and their first and second-order derivatives. Each state in HMM has 16 Gaussian mixtures.

3.4.1 Code-mixing vs. Monolingual

Recognition performance on Cantonese-English code-mixing utterances is given in Table 3.14. Speech recognition experiments were also performed on monolingual Cantonese utterances and embedded English words for comparison purposes. The recognition performance is compared with syllable accuracy for Cantonese and word accuracy for English. No language model is applied, and the lexicon contains both Cantonese and English entries. The embedded English words are extracted from the code-mixing utterances. It means that the language boundary information is correct. The word accuracy of the extracted

embedded English is 76%, which is the upper bound accuracy with perfect language boundary information. Compared with the word accuracy attained for code-mixing English, more than 20% degradation is found.

The recognition accuracy attained for code-mixing Cantonese is 57.6%. Compared with the monolingual results, more than 5% degradation is observed. This may be due to the effect of embedded English. Further analysis is carried out to investigate how embedded English exerts influence on the recognition accuracy of code-mixing speech.

Table 3.14: Baseline recognition accuracy on code-mixing and monolingual speech data

Code-mixing			Monolingual	Embedded
Overall	Cantonese	English	Cantonese	English
57.3%	57.6%	54.1%	62.4%	76%

3.4.2 Effect of Embedded English

Since the accuracy of the embedded words may play an important role in code-mixing ASR, we divide the recognition results on code-mixing utterances into two sets in terms of the recognition accuracy of their English segments. As a result, the embedded English words in **Set A** were recognized correctly, while the English words in **Set B** were wrongly recognized. The speech recognition results for these two sets are given in Table 3.15. Another experiment is performed on **Set B** only. This is an oracle experiment. We assume that the embedded English words are already known before decoding. The grammar network designed for this experiment is shown in Figure 3.7. The recognition results are given in Table 3.15.

It is noticed that the recognition performance is quite different between **Set A** and **Set B**. If the embedded English words can be recognized, the recognition accuracy on code-mixing Cantonese is 63.2%, which is comparable to the monolingual results shown in Table 3.14. This means that the embedding effect does not degrade the performance of code-mixing ASR if the embedded

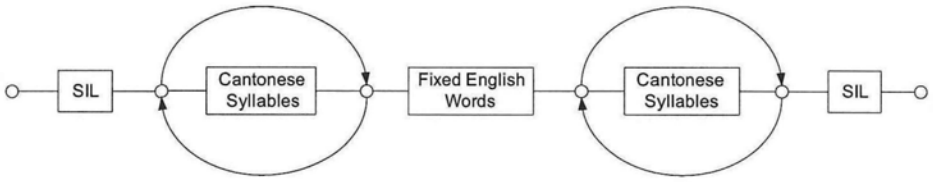


Figure 3.7: Grammar network for oracle experiment.

segments can be recognized correctly. On the other hand, if the embedded words can-not be recognized, the error on embedded language will lead to significant degradation on the matrix language.

The recognition accuracy attained by the oracle experiment is 59.2% and 94.5% for code-mixing Cantonese and English, respectively. Recognition errors still exist on code-mixing English since the oracle experiment does not specify the language boundaries. The assimilation caused by code-mixing makes some short English words very highly similar to Cantonese syllables, and therefore, some of the Cantonese syllables are wrongly decoded as the reference English, while the real English segments are recognized as Cantonese. For code-mixing Cantonese, the performance attained by the oracle experiment shows obvious improvement, although it is still lower than that of the monolingual case. This shows that the information about the embedded language is very useful.

Table 3.15: Evaluation results in terms of the recognition accuracy on embedded English

Recognition accuracy	Set A (Correctly rec. Eng)	Set B (Wrongly rec. Eng)	Set B (Oracle exp.)
Overall	66.5%	46.7%	62.4%
Cantonese syl.	63.2%	51.3%	59.2%
English word	100%	0%	94.5%

3.4.3 Error Propagation of Embedded English

Further analysis is carried out to investigate the error propagation of the embedded English.

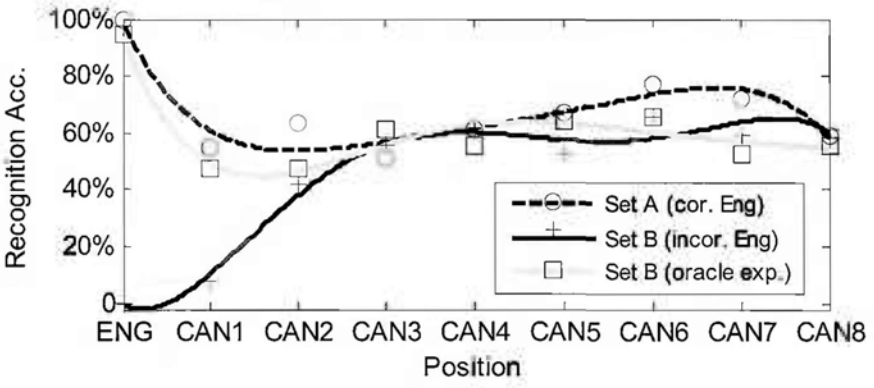
We divide the recognition results on code-mixing utterances into three cases in terms of the position of the embedded English segments. For each case, the error propagation are observed under the following conditions: (1) when English is recognized correctly (**Set A**), (2) when English is wrongly recognized (**Set B**), and (3) the oracle experiment for **Set B**. The error propagation graphs are given in Figure 3.8. Graphs (a), (b) and (c) show the error propagation when the embedded English words are at the beginning, middle and end of the code-mixing utterances, respectively.

It can be observed that the embedding effect does not induce error nearby if the embedded segments can be recognized correctly. However, if the English words are wrongly recognized, the error propagates to the surrounding Cantonese syllables. In particular, when the English words appear at the beginning/end of the utterances, the error significantly propagates to the first and second following/preceding Cantonese syllables, respectively. When the English words are in the middle of the utterances, the errors slightly spread to the first preceding Cantonese syllable, but this seriously affects the first subsequent Cantonese syllable. In the oracle experiment, embedded English will only propagate at most to one neighbouring Cantonese syllable.

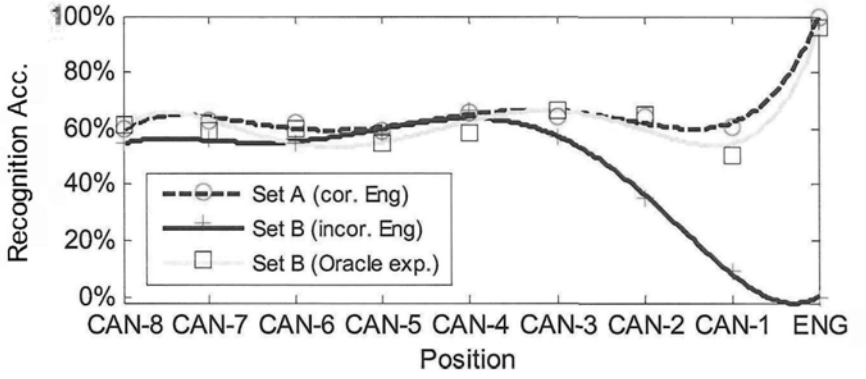
Another analysis is performed to reveal the degree of the embedded language effect with different error types in English. In this study, we classify the error on code-mixing English into three types as follows.

- Error type I: the whole English segment is recognized as another English segment.
- Error type II: the whole English segment is recognized as Cantonese syllables.
- Error type III: part of the English segment is recognized as another English segment, while the remaining part is recognized as Cantonese syllables.

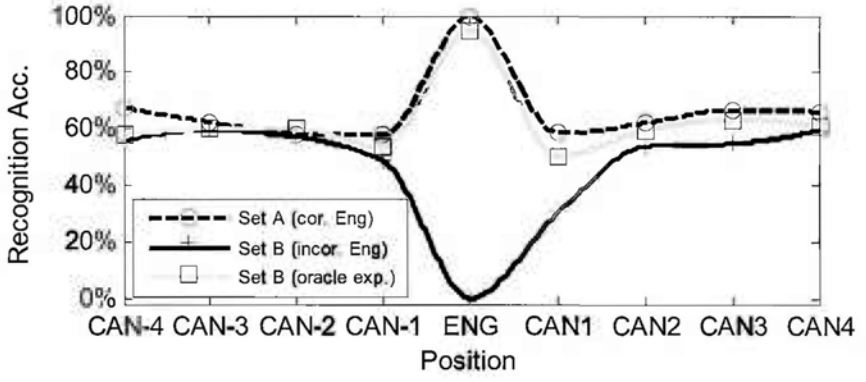
The English error type is dependent on the accuracy of language boundary information, which is implicitly detected in speech recognition because the language identity can be determined according to the recognized words. We



(a) When the English words are at the beginning



(b) When the English words are at the end



(c) When the English words are in the middle

Figure 3.8: Error propagation of embedded English

analyze the speech recognition performance in terms of the accuracy of this implicit language boundary detection (LBD). The results are shown in Table 3.16, where CM denotes the overall accuracy for code-mixing.

Table 3.16: Recognition performance with accuracy of language boundary information

Inaccuracy on LB (ms)	Set A			Set B			Set B (Oracle)		
	Ratio	Acc.(%)		Ratio	Acc.(%)		Ratio	Acc.(%)	
		CM	Can.		CM	Can.		CM	Can.
<25	57.1	66.3	62.9	5.6	56.0	60.9	44.0	64.7	61.3
25 - 75	31.2	67.1	63.8	4.7	56.1	61.2	29.9	62.2	58.5
75 - 150	7.4	68.3	65.3	3.3	45.8	51.1	11.0	61.5	58.0
>150	4.4	62.5	58.6	45.3	46.8	51.3	10.1	60.5	56.9
No overlap	0	NA	NA	40.6	44.1	48.2	5.0	52.0	56.3

To evaluate the performance of LBD, the detected boundaries of a language segment are compared to the manually aligned true boundaries and the time difference is measured as the inaccuracy of LBD. In this study, we divide the inaccuracy of LBD into several categories as shown in Table 3.16. If the difference between detected boundaries and true boundaries is smaller than 25ms (one frame), it can be regarded as a perfect detection. If the difference is smaller than 75ms, which is about the time duration of one phone, it still can be marked as a successful detection. However, if the inaccuracy of LBD is larger than 75ms, an LBD error will be recorded and the error can be further classified into 3 levels: a) the inaccuracy of LBD is smaller than 150ms, which is the typical duration of one syllable; b) the inaccuracy of LBD is larger than 150ms; and c) no overlap can be found between detected boundaries and true ones. As a result, English error type I usually happens when the language boundary can be successfully detected (inaccuracy of LBD is less than 75ms), English error type II appears along with the LBD error c, and English error type III relates to LBD errors a and b.

If English words can be correctly recognized (set A), it is not surprising

that almost 90% of these words are recognized with correct boundary information (inaccuracy of LBD $< 75ms$). It is also noticed that, if embedded English words cannot be recognized (set B), only 10% of these errors in English are language-specific errors (English error type I). In this case, the language boundary information is expected to be correct, and therefore the recognition accuracy in Cantonese is not affected by the embedded language. On the other hand, 90% of the errors in English belong to English error types II and III, which are caused by confusion between the languages (inaccuracy of LBD $> 75ms$). In this case, additional deletion and insertion error is introduced by imperfect language boundary information, and therefore, error on the embedded language will propagate.

3.4.4 Influencing Factors for Embedded English

Factor analysis is carried out to determine the influencing factors for recognition performance in code-mixing English. Studies are performed using selected factors as the following.

- Position of English words
- Phonology of English words
- Characteristics of nearby Cantonese syllables

The recognition performances with different positions are given in Table 3.17. It is shown that the English accuracy is relatively lower when English is in the middle. This may be due to the bidirectional articulation from the preceding and following Cantonese syllables.

Further analysis is performed to investigate the effects of English phonology. The study is performed in terms of syllable numbers and syllabic structure of English words. Table 3.18 shows the English accuracy with different syllable numbers per word. It indicates that the recognition accuracy increases with increase of syllables.

Table 3.17: Recognition results with different English positions

Position	Beginning	Middle	End
Overall Acc.	56.7%	57.3%	57.2%
Cantonese syl. Acc.	56.7%	57.6%	57.1%
English word Acc.	56.3%	**53.5%**	57.6%

Table 3.18: English accuracy with the number of syllables per English word

No. of Syllables	Proportion	English word Acc.
1	22.2%	37.4%
2	45.6%	53.8%
≥ 3	32.2%	64.2%

Next, analysis is performed on the syllable structure in terms of number of syllables per English word. We observe that if the embedded English is a monosyllabic word, the syllable structure indicates a significant effect on recognition accuracy. The English accuracy is found to be much lower when the C-V-C structure is retained. However, with an increase in syllable numbers, the effect of the syllable structure decreases. If the English word consists of three or more syllables, the recognition performance is not affected by its syllable structure.

Another analysis is carried out to investigate the effect of the immediately preceding/following Cantonese syllables around the embedded English. The study is performed in terms of analysis syllable correctness and the syllabic structure. It is found that there is indeed a strong correlation between English word accuracy and the correctness of the neighbouring Cantonese syllables. Table 3.19 gives the details when English is embedded in the middle of utterances. It shows that the recognition accuracy on English words can be over 80% when the surrounding Cantonese syllables can be recognized. However, the recognition performance is degradated to less than 40% if the surrounding Cantonese syllables cannot be recognized. On the other hand, no obvious observation is made in relation to the Cantonese syllable structure.

Table 3.19: English accuracy versus surrounding Cantonese correctness

Cantonese	English word Acc.	Cantonese
Correct	82.1%	Correct
Correct	41.8%	Wrong
Wrong	56.7%	Correct
Wrong	37.0%	Wrong

3.4.5 Summary

This study mainly investigated the effect of embedded language on code-mixing speech recognition in Cantonese-English code-mixing ASR. It is found that the recognition accuracy in code-mixing is significantly affected by the recognition performance of the embedded language. In particular, the embedding effect does not degrade the performance of code-mixing ASR if the embedded segments can be recognized correctly, but significant degradation is found in the matrix language if the embedded words can not be recognized. This indicates that the recognition performance on embedded language is very important, and it is believed that the improvement in the embedded language will bring improvement to the matrix language as well. Future studies to improve the performance of embedded English are therefore desirable.

□ End of chapter.

Chapter 4

Cross-lingual Use of Acoustic Information for Cantonese & English

Summary

As the key components in speech recognition, the acoustic models, pronunciation dictionary and language models need to be carefully designed for Cantonese-English code-mixing speech. This chapter discusses the cross-lingual use of acoustic sources and characteristics for Cantonese and English. It begins with an introduction of speech corpora used in this thesis. After that we present a study on pronunciation variations in Cantonese-English code-mixing speech spoken by native Cantonese speakers, followed by the implications of these observations for the design of a code-mixing speech recognition system. Then we focus on cross-lingual acoustic modeling. Various combination schemes and similarity measurements are investigated to design different cross-lingual phoneme sets. Finally, cross-lingual adaption via model mapping is discussed.

4.1 Speech Corpora

Three speech corpora are selected in this study. They are native English corpus TIMIT [110], read style Cantonese corpus CUSENT [109] and Cantonese-English code-mixing corpus CUMIX [25].

TIMIT

TIMIT is a speech corpus of American English spoken by speakers from different dialectal regions in the United States. It has been widely used for the development and evaluation of speaker independent phone recognition systems. There are 630 speakers in the corpus. Each speaker produces 10 utterances that are categorized as follows:

sa: 2 common sentences for all speakers

sx: 5 sentences selected from a pool of 450 phonetically balanced sentences

si: 3 sentences randomly selected without phonetic coverage considerations

Only **sx** and **si** utterances are used in our work. **sa** utterances are excluded because they have fixed contents and therefore may lead to undesirable bias of the acoustic models towards certain phonemes. Among the 630 speakers, 462 were designated to be the training speakers and the remaining were for testing.

CUSENT

CUSENT was developed by the DSP & Speech Technology Laboratory of the Chinese University of Hong Kong (CUHK). It is a large collection of read-style formal Cantonese sentences from local newspapers, which were designed to be phonetically rich. The construction of this corpus is intended for the development of speaker independent continuous speech recognition for Cantonese. There are 40 male and 40 female speakers, among which 68 speakers are designated for training (300 sentences per speaker) and 12 speakers are for testing (100 sentences per speaker). The total number of training and testing utterances are 20,400 and 1,200 respectively, where distinct sentences are 5,100 and 600 respectively.

CUMIX

CUMIX is a database developed specifically for Cantonese-English code-mixing speech recognition. The spoken contents in CUMIX are mainly daily conversations or jargon by university students in Hong Kong. There are three different types of utterances in CUMIX:

CM: Cantonese-English code-mixing utterances

MC: monolingual colloquial Cantonese utterances

ME: monolingual English words and phrases

It contains 16 hours of speech data from 74 speakers. The training data include utterances from 20 male and 20 female speakers. Each speaker has 200 **CM** utterances and 100 **ME** utterances. There are 14 male and 20 female speakers in the test data. Each of them has 120 **CM** utterances and 90 **MC** utterances. The details of the CUMIX corpus are given in Table 4.1.

Table 4.1: A summary of CUMIX

		Training data	Test data
		20 males, 20 females	14 males, 20 females
CM	Duration:	7.5 hours	4.25 hours
	Duration of English segments:	1.13 hours	0.57 hours
	Total no. of utterances:	8,000	3,740
	No. of unique sentences:	2,087	2,256
	No. of unique English segments:	1,047	1,069
MC	Duration:		2.75 hours
	Total no. of utterances:		3,060
	No. of unique sentences:		1,742
ME	Duration:	1.5 hours	
	Total no. of utterances:	4,000	
	No. of unique sentences:	1,000	

4.2 Pronunciation Variation Modeling for Cantonese-English Code-mixing ASR

As we discussed in the last chapter, there exists many pronunciation variations in Cantonese-English code-mixing speech, and therefore we often get lower speech recognition accuracy on code-mixing speech than monolingual speech spoken by native speakers. Understanding how native and code-mixing speech differs is an important first step to tackle the problem of code-mixing speech recognition. In the last decade, much work has been done to study pronunciation variations in Western languages in monolingual speech recognition. However, works related to code-mixing speech, particularly for Cantonese-English code-mixing speech is rare. To improve the design of acoustic models and construct an accurate bilingual pronunciation dictionary, in-depth studies on pronunciation variations in code-mixing speech are needed.

4.2.1 Phone Recognition Experiments

We are interested in how and where native and code-mixing speech differ from each other in terms of pronunciation variation. A data-driven computational approach is adopted to reveal significant pronunciation variations in Cantonese-English code-mixing speech, in addition to those variations that have been well understood in monolingual speech recognition. We assume that frequent and systematic phonetic variations can be reflected by noticeable confusion patterns in automatic speech recognition. A series of phone recognition experiments using state-of-the-art acoustic modeling techniques has been carried out.

Baseline phone recognizer

In the first step, 20,000 utterances from CUSENT and 4,000 utterances from TIMIT are used to train context-dependent baseline phone models for Cantonese and English respectively, which represent carefully articulated speech from native speakers. The acoustic feature vector consists of 39 conventional MFCCs, and each state in HMM has 16 Gaussian mixtures. For Cantonese,

73 Initials and Finals are used as the basic modeling units. For English, 39 IPA-based phone-level units are applied [111].

Then the baseline phone recognizers are evaluated with different types of data as shown in Table 4.2. In the evaluation, no phone grammar was applied in the English phone recognizer, and the Cantonese recognizer followed the Initial-Final constraint.

Table 4.2: Test data sets applied in phone recognition experiments

	Data sets	Characteristics
ENG	TIMIT	native English
	CUMIX_ME	monolingual English words with Cantonese accents
	CUMIX_CME	English extracted from code-mixing utterances
CAN	CUSENT	read-style formal Cantonese
	CUMIX_MC	monolingual colloquial Cantonese utterances
	CUMIX_CMC	Cantonese extracted from code-mixing utterances

For the embedded language English, a phone recognition experiment is first carried out on TIMIT. The observed phonetic variation patterns are regarded as the references for subsequent analysis. When speech data in the CUMIX corpus (CUMIX_ME & CUMIX_CME) are used for testing, some unseen variation patterns are expected. By comparing these patterns with the references, additional variations can be identified. Common variations found in CUMIX_ME & CUMIX_CME are probably due to the Cantonese accents of the speakers, while their differences may indicate the effect of the code-mixing scenario. For the matrix language Cantonese, the recognition results on the CUSENT test utterances are used as the benchmark. Phonetic variations in colloquial Cantonese are analyzed in the same way as for non-native English.

Phone recognition accuracy

The phone accuracies for different test data are given in Table 4.3. The benchmark accuracies (matched conditions) are 64.5% and 85.6% for native English and read-style formal Cantonese respectively, which are comparable to the state-

of-the-art results reported in previous studies [112][113]. Compared with native English, the recognition accuracy for non-native English speech in CUMIX is much lower. There is no significant difference between monolingual English words spoken by Cantonese speakers and embedded English words in code-mixing utterances. For colloquial Cantonese in CUMIX, the recognition accuracy is about 20% lower than the benchmark. This indicates a significant effect of the change in speaking style.

Table 4.3: Phone recognition results for English & Cantonese

	Data set	Acc.	Ins.	Del.
ENG	TIMIT (native)	64.50%	6.30%	6.50%
	CUMIX_ME (non-native)	35.90%	4.70%	20.60%
	CUMIX_CME (code-mixing)	33.30%	4.50%	20.70%
CAN	CUSENT (read)	85.60%	3.60%	0.10%
	CUMIX_MC (colloquial)	66.40%	4.50%	3.50%
	CUMIX_CMC (code-mixing)	63.70%	5.10%	3.60%

Analysis is also done for individual phonemes. Figure 4.1 shows a global view of the recognition results on the 39 English phonemes. Along the horizontal axis, the phonemes are arranged in the order of ascending accuracy for native English (TIMIT utterances). The recognition accuracy for non-native English in CUMIX is lower than that for native English in relation to almost all phonemes. The degree of degradation varies a lot among different phonemes. We list all English phonemes with significant degradation (~35% absolute reduction) from native to non-native English in Group A of Table 4.4. We expect that these phonemes are subject to great variation. In non-native English speech, the difference between monolingual English words and code-mixing English words applies to only a few phonemes, as shown in Group B of Table 4.4. Further discussion will be given in Section 4.2.2.

Figure 4.2 shows the recognition accuracies for individual phonemes for read-style and colloquial Cantonese speech. For most phonemes, colloquial Cantonese exhibits lower recognition accuracy than read-style formal Cantonese, but the

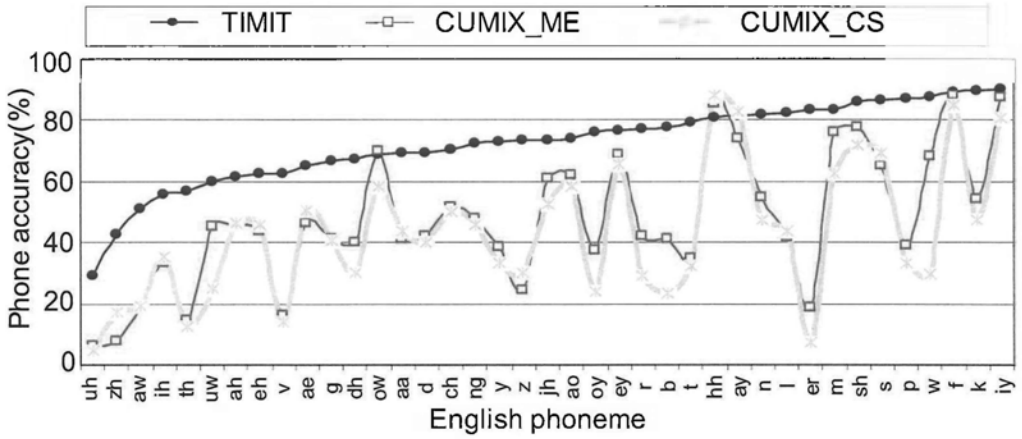


Figure 4.1: Recognition results for individual English phonemes

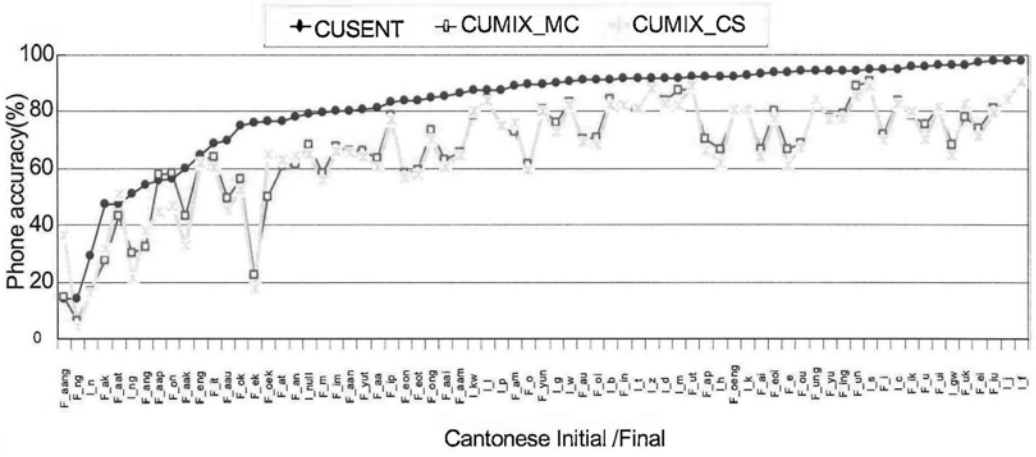


Figure 4.2: Recognition results for individual Cantonese initial/finals

gap is not as great as that between non-native and native English. The degree of degradation varies greatly among different phonemes. It is also seen that there is no noticeable difference between colloquial Cantonese in monolingual utterances and in code-mixing ones.

4.2.2 Analysis of the Confusion Matrix

A confusion matrix shows the degree of confusion between each pair of phonemes. Since there exists great variation in the frequency of occurrences of different phonemes in different databases, the number of confusing cases is normalized with respect to the phoneme frequency count. We compute the

Table 4.4: Phonemes with significantly different recognition accuracies in different types of English speech

	native vs. non-native (Group A)	monolingual vs. code-mixing (Group B)
Phonemes with significantly different accuracies	/th/, /v/, /l/, /z/, /oy/, /r/, /b/, /t/, /er/, /p/, /k/	/uw/, /oy/, /b/, /r/, /er/, /w/, /dh/

normalized confusion matrix as:

$$\hat{C}(i,j) = \frac{C(i,j)}{\sum_j C(i,j)} \tag{4.1}$$

where $C(i,j)$ denotes the number of cases that phoneme j is mis-recognized phoneme i. The value of $\hat{C}(i,j)$ is between 0 and 1.

As mentioned previously, frequent and systematic pronunciation variations can be reflected by noticeable confusion patterns in the normalized confusion matrix $\hat{C}(i,j)$. There are many pairs of phonemes that have a non-zero value of $\hat{C}(i,j)$. We assume that only those exceeding a certain threshold $T(0 < T < 1)$ are related to systematic pronunciation variations that we are interested in.

Moreover, the difference between two normalized confusion matrices for the same language can be represented by the discrepancy matrix defined below.

$$\triangle_{AB}(i,j) = \hat{C}_B(i,j) - \hat{C}_A(i,j) \tag{4.2}$$

The value of $\triangle_{AB}(i,j)$ is between -1 and 1.

Analysis of the English Confusion Matrix

Figure 4.3 shows part of the normalized confusion matrices between native and non-native English phonemes. Only major confusions ($T = 0.15$) are included for the ease of visualization. Numbers in boldface indicate highly confused phonemes, which are considered to be phonetic variations in non-native English. Similar analysis can be done also on native English. Comparison between native and non-native English leads to the following observations.

1: For native English, only two significant phonetic variations are observed, /z/ to /s/ and /zh/ to /sh/. They are also seen in non-native English, to a much more severe extent.

2: Most of the phonetic variations occurring in non-native monolingual English and code-mixing English are common. The extent of these variations is also at the same level.

3: In Table 4.4, Group B lists those phonemes on which code-mixing English is more poorly recognized than non-native monolingual English. For example, /er/ is frequently mis-recognized as /aa/ and /w/ becomes /l/ in code-mixing English but not in monolingual English. For /oy/, /dh/ and /uw/, no particular new variation pattern can be seen. In addition, the deletion of /b/ and /r/ is more noticeable in code-mixing English.

4: /uh/ is found to be very badly recognized in both non-native and native English. No specific phonetic variation pattern can be observed. It may indicate that this phoneme is not effectively modeled by the HMM.

5: Consonant phonemes /b/, /r/, /t/, /p/, /l/ and /k/ as listed in Group A of Table 4.4, are found to have low recognition accuracy, because of the high deletion rate. No specific phonetic variation can be observed.

	aa	ae	ah	ao	eh	f	iy	l	n	ow	s	sh
aa				**0.19/0.21**								
ae					**0.19/0.15**							
aw	**0.22/0.26**									**0.15/0.17**		
eh		**0.27/0.26**										
er	**0.17/**		**0.17/0.26**									
ih							**0.31/0.28**					
ng									**0.16/0.27**			
ow								**0.2/0.16**				
th						**0.54/0.51**						
uh			**0.23/0.19**					**0.22/0.17**		**0.17/0.23**		
v						**0.44/0.44**						
w								**0.32/**				
z											**0.47/0.46**	
zh												**0.5/0.52**

Figure 4.3: Normalized confusion matrix from non-native English (CUMIX_CME/CUMIX_ME)

Context-dependent Phonetic Variation in English

Among the English phonetic variations discussed above, there are three context-independent phonetic variations: /th/ to /f/, /ae/ to /eh/ and /ih/ to /iy/.

The other variations typically occur under specific contextual conditions. Such context-dependent variations are identified and listed in Figure 4.4. We divide them into 4 different levels as follows.

L1: Variation occurring only in specific lexical entries

L2: Variation depending on phoneme position in the word

L3: Variation depending on syntax, such as part-of-speech (POS)

L4: Context-independent variation

Phonetic variation patterns	Contextual rules			Example words
	L1	L2	L3	
/er/ to /ah/ or /aa/		when /er/ in the end of the word	when /er/ in the noun	tutor, paper
/oy/ to /ao/	when /oy/ followed by /n/			point, join
/v/ to /f/		when /v/ in the middle/end of the word		five, objective
/dh/ to /f/	when /th/ in word 'with'			with, within
/z/ to /s/	when /z/ followed by '/ah/ /l/'	when /z/ in the end of the word	when /z/ in plural noun	overseas, proposal
/zh/ to /sh	when /zh/ followed by '/ah/ /l/', in the suffix 'sion' or 'sure'			measure, fusion, visual
/ng/ to /n/		when /ng/ in the end of the word		marketing, admin

Figure 4.4: Major context-dependent phonetic variations in English

Analysis of the Cantonese Confusion Matrix

Normalized confusion matrices on Cantonese phonemes are analyzed and compared among different types of speech. The major variation patterns in different test sets are shown in Table 4.5. F_ denotes Final units and I_ denotes Initials. For example, F_-t represents a Final ending with coda /t/ while F_aa- represents a Final starting with vowel /aa/. Although the variation patterns of read-style formal Cantonese and colloquial Cantonese are mostly common, the extent of confusability is different. Figure 4.5 shows the number of confusing

Table 4.5: Confusion patterns in Cantonese

Variation patterns	CUSENT	CUMIX_MC	CUMIX_CMC
(F_-t, F_-k with F_-p)	✓	✓✓	✓✓
(F_aa- to F_a-)	✓	✓✓	✓✓
(F_-ng to F_-n)	✓	✓✓	✓✓
(I_n to I_l)	✓	✓✓	✓✓
(F_ng to F_m)	✓	✓✓	✓✓
(I_ng to I_null)	✓	✓✓	✓✓
(F_eon to F_oeng)		✓✓	✓✓
(I_gw to I_g)		✓✓	✓✓

pairs appearing in the discrepancy matrix with different cutting threshold T, which indicate the degree of confusion due to the colloquial speaking style. It is generally observed that the degree of variation is more severe in colloquial Cantonese than in formal Cantonese. On the other hand, the degree of colloquial is partially related to the speaking rate. For example, syllable fusion is one type of pronunciation variation that mainly occurs in fast speech. Figure 4.6 shows the average syllable duration of different types of Cantonese. It can be seen that colloquial Cantonese in CUMIX is 25% faster than formal Cantonese in CUSENT.

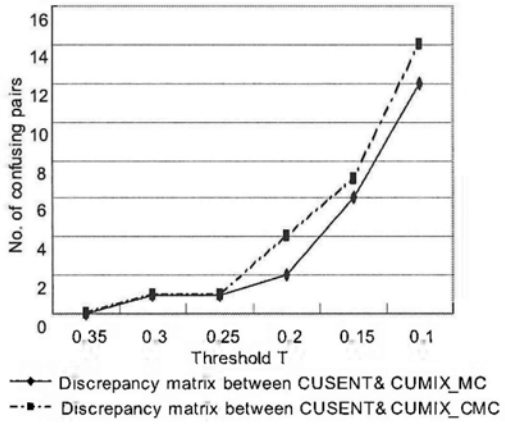


Figure 4.5: No. of confusing pairs due to the colloquial nature of Cantonese

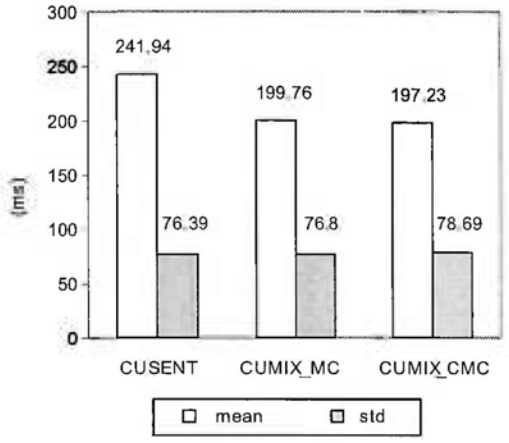


Figure 4.6: Syllable duration of different types of Cantonese

4.2.3 Implications for Code-mixing ASR

In Cantonese-English code-mixing speech, the embedded English words carry strong Cantonese accents, which are reflected by the phonetic variation patterns as discussed in Section 4.2.2. The context-dependent rules of phonetic variation in Figure 4.4 should be adopted to modify the pronunciation lexicon of code-mixing English. For example, the variation from /er/ to /aa/ at the end of a noun is very conspicuous in English spoken by Cantonese speakers. The relevant pronunciation variants should definitely be included in the pronunciation lexicon. To encounter the very likely deletion of English consonants, e.g., /l/ (as in "plan"), /k/ (as in "book"), additional lexical entries should be added to represent such variants. On the other hand, some of the variations seem to be assured. As seen in Figure 4.3, the confusions from /th/ to /f/ and from /v/ to /f/ are about 50%. In these cases, we would suggest substitution of phonemes instead of adding new pronunciations. Note that these substitutions may be context-dependent, e.g., /z/ to /s/. Based on observed variation patterns, we generate the accented dictionary in a data-driven way. The modified dictionary contains an average of 1.48 pronunciations for each lexical item.

On the other hand, the difference between monolingual non-native English and code-mixing English is marginal. There is also no significant difference between Cantonese utterances with and without code-mixing. This suggests that monolingual non-native English and monolingual colloquial Cantonese would be effective to train the acoustic models in a code-mixing speech recognition system. This will alleviate the problem of not having sufficient code-mixing data, since we believe that monolingual speech data are easier to collect than code-mixing one.

Three sets of monolingual English acoustic models are trained with different speech data, which are denoted by ME_I, ME_II, and ME_III, in Table 4.6 respectively. They are evaluated with both the standard CMU dictionary and the accented dictionary, and the recognition results are shown in Table 4.6. The testing data are English segments extracted from code-mixing utterances. ME_I attains a very low accuracy on English words. This indicates that Cantonese-

accented English is very different from the native American English found in TIMIT. ME_II improves greatly due to better matched training. Nevertheless, the recognition accuracy is still on the low side because of the involvement of TIMIT data in the training phase. ME_III gives the best recognition performance. It confirms that the acoustic model trained from the best-matched speech data gives the best performance. Moreover, noticeable improvement is observed with the accented dictionary.

Table 4.6: English word accuracy with different AMs and dictionaries

Monolingual English AMs	Training Data	Pronunciation Dictionary	English Word Acc.
ME_I	TIMIT	Standard CMU dict.	49.88(%)
	TIMIT	Accented dict.	51.32(%)
ME_II	TIMIT + CUMIX	Standard CMU dict.	70.68(%)
	TIMIT + CUMIX	Accented dict.	73.29(%)
ME_III	CUMIX	Standard CMU dict.	74.84(%)
	CUMIX	Accented dict.	**75.62**(%)

Similarly experiments are performed on colloquial Cantonese. Table 4.7 explains the three sets of monolingual Cantonese acoustic models, which are trained with read-style speech data (CUSENT), colloquial Cantonese data (CUMIX) and a pool of both, respectively. The testing data are the monolingual Cantonese (MC) utterances of CUMIX, and the modified dictionary is applied in the evaluation. Table 4.7 lists the recognition accuracy in terms of Cantonese syllables. CUSENT is a read-speech database for formal Cantonese, where many colloquial Cantonese terms or lexicons cannot be found. It is not surprising that poor performance is found in model MC_I due to the mismatched speaking style and spoken content. It is also noticed that MC_II and MC_III shows equal performance in recognizing MC speech of CUMIX, although they are trained with different sets of speech data.

Further analysis is performed on the phoneme-level (Initial/Final) and the results are given in Table 4.8. Statistical analysis of individual phonemes in-

Table 4.7: Cantonese syllable accuracy with different AMs

Monolingual Cantonese AMs	Training Data	Cantonese Syl. Acc.
MC_I	CUSENT	58%
MC_II	CUSENT + CUMIX	67.15%
MC_III	CUMIX	67.15%

dicates the performance difference between MC_II and MC_III. CUMIX is a colloquial speech corpus developed for Cantonese-English code-mixing ASR applications. The speech data in CUMIX is domain-specific due to the linguistic characteristics of code-mixing. The tri-phone coverage is more balanced in MC_II. Therefore, it is not surprising that smoother recognition results are found in MC_II. MC_III shows fairly different performance in recognizing different phonemes. It attains very high accuracy for some phonemes (related to colloquial lexicons), while attaining relatively low accuracy for others (related to standard Chinese characters). Hence, it suggests that model MC_II outperforms MC_III, especially for unseen test data.

In summary, only accented English speech data from CUMIX should be used in the acoustic modeling of Cantonese-English code-mixing speech. On the other hand, both colloquial Cantonese speech in CUMIX as well as read-style Cantonese from CUSENT should be applied in acoustic modeling in the following studies.

Table 4.8: Statistics analysis of different sets of training data and AMs

Monolingual Cantonese AMs	Triphone coverage on test speech	Cantonese Phone Acc.	Mean of 73 Phones	Std of 73 Phones
MC_I	87.60%	70.97%	70.21	19.38
MC_II	89.87%	78.08%	74.73	16.54
MC_III	97.36%	78.07%	66.85	24.33

4.3 Cross-lingual Acoustic Modeling

This part of the research aims at the development of cross-lingual acoustic modeling of Cantonese-English code-mixing speech. The crucial issue is to design an appropriate universal phoneme inventory, which should cover all phonemes of Cantonese and English.

4.3.1 Design of Cross-lingual Phoneme Inventory

As discussed, it is expected that Cantonese and English have a number of phonemes that are acoustically or phonetically identical or similar to each other. The degree of similarity varies. In principle, highly similar phonemes can be combined together. Various combination schemes and similarity measurements can be applied to decide which phonemes should be merged together.

In this thesis, professional linguistic knowledge is first implemented to cluster phonemes. We use IPA classification to facilitate an intuitive comparison between Cantonese and English phonemes, where phonemes labelled with the same IPA symbols are clustered into one class. Based on the IPA, a global unit set with 65 phones as shown in Figure 3.6 are constructed, and a conventional context-dependent tied models CL_IPA are trained based on this phoneme inventory as our benchmark. This acoustic model is applied in our pilot study to investigate the effects of language mixing for code-mixing ASR (see Section 3.4) as well.

On the other hand, various data-driven approaches are discussed to estimate which phonemes should be merged together. Acoustic and phonetic similarity between different phonemes are studied by investigating the K-L divergence and calculating the phoneme confusion matrix. Based on that, different sets of cross-lingual phoneme inventories are designed and they are evaluated in speech recognition experiments with context-independent acoustic models.

Basic Sound Units for Phoneme Clustering

As we introduced in Chapter 3, the conventional sound inventory for English is phoneme-based. It contains 39 basic English phonemes as defined in the International Phonetic Alphabet (IPA) (see Figures 3.4 and 3.5). However, different from English, different sound units (e.g. Initial/Final) can be used to represent Cantonese due to its specific phonology and phonetics. To investigate the performance of different sound representation schemes and design an appropriate inventory for the acoustic modeling of Cantonese-English code-mixing speech, three sets of sound inventories are employed for Cantonese as shown in Table 4.9. They are denoted by IF, CanP, and IPA, respectively. An IF-based Jyut Ping system has been the most widely used in monolingual Cantonese speech recognition in previous research [73]. The details of Cantonese initials and finals have been introduced in Figure 3.1. However, different from the Jyut Ping system, the IPA only assigns vowels and consonants to describe Cantonese. Vowel-nasal and vowel-stop finals are divided into vowels, nasals, and stop consonants separately. Sound units in CanP are mainly based on the IPA, except that the vowel-stop finals are kept. This is because the Cantonese-specific stop consonants /p/, /t/, and /k/ are unreleased, and their sound mainly occurs at the vowel instead of the consonant.

Table 4.9: Different sound units for Cantonese

IF	initial/final-based, 73 LSHK-based IF are involved.
CanP	mainly phone-based, 58 sound units are involved.
IPA	phone-based inventory, 43 phonemes are involved.

Phonetic Similarity based on the Confusion Matrix

Phonetic confusion indicates the degree of phonetic similarity. By inspecting language-dependent intra phonetic confusion and inter phonetic confusion between Cantonese and English phonemes in speech recognition outputs, we may understand which phonemes are highly similar and therefore can be merged together.

The procedure is as follows. Firstly, a set of language-dependent bilingual acoustic models are trained with CUSENT and CUMIX speech data. Then, phoneme-based speech recognition experiments are performed with this bilingual acoustic model. CM utterances from the development set of CUMIX are used in recognition experiments. A confusion matrix is derived from the recognition results.

As the embedded English words are Cantonese-accented, it is observed that many English phonemes are recognized as Cantonese ones, but few Cantonese phonemes are recognized as English ones. The counts in the confusion matrix determine the level of confusion between the phonemes. If the confusion from one phoneme to another exceeds 50%, it is regarded as highly similar phonetically and therefore can be clustered into one phoneme class. The observed phoneme clusters include:

- Cross-lingual clusters: e.g. English phoneme /E_er/ with Cantonese final /F_aa/
- Language-dependent clusters: e.g. English phoneme /E_f/ with English phoneme /E_v/

Acoustic Similarity based on Kullback-Leibler Divegence

Acoustic similarity between two phonemes can be indicated by the Kullback-Leibler divergence (KLD) between the corresponding HMMs. Figure 4.7 shows a schematic diagram of KLD-based phoneme clustering.

Bilingual speech of 10 male and 10 female speakers from the development set of CUMIX are involved in this study. For each speaker, single-mixture, context-independent acoustic models are trained for Cantonese and English phonemes respectively. The K-L divergence is then calculated between the HMMs of every possible pair of phonemes. As a result, each English phoneme has a mapped Cantonese phoneme, in the minimum-KLD sense. Particularly, for each speaker i, we can find a mapped Cantonese phoneme p_{ec_i} for each English phoneme p_{e_i}. The corresponding K-L score is denoted as $kls_{p_{ec_i}}$, which is calculated as

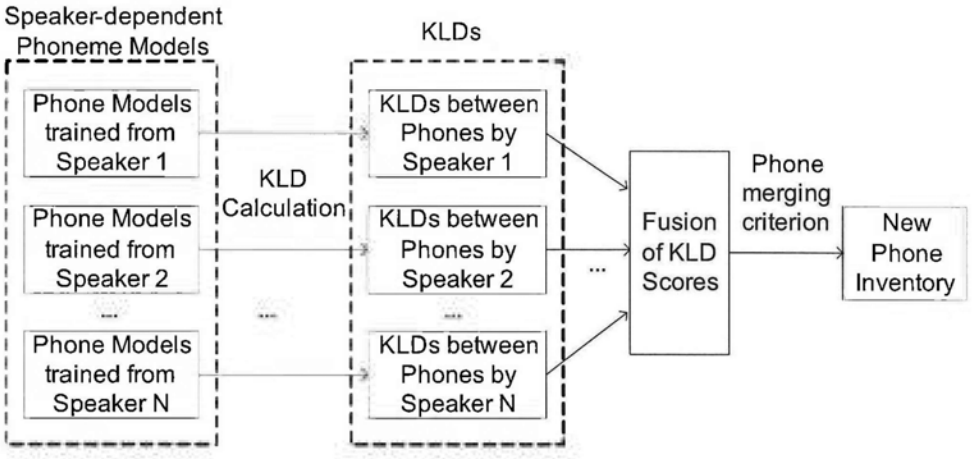


Figure 4.7: KLD-based phoneme combination

the reciprocal of their K-L divergence. In order to remove speaker-dependent characteristics, fusion is performed among all speakers as:

$$KLS_{p_{ec}} = \sum_i kls_{p_{ec_i}} \tag{4.3}$$

Where KLS is the K-L score matrix. Figure 4.8 shows part of the score matrix between English consonants and Cantonese initials. Its clear from the score matrix that each English phoneme may map to one or more Cantonese phonemes with different scores due to the speaker variability. Therefore, for each English phoneme p_e, we need to search for its speaker-independent closest Cantonese phoneme p^*_{ec} from Cantonese sound inventory C with the highest score $kls_{p^*_{ec}}$ as:

$$p^*_{ec} = \arg\max_{p_{ec} \in C} \sum_i kls_{p_{ec_i}} \tag{4.4}$$

On the other hand, we calculate the language-dependent K-L divergence between different English phonemes. For each English phoneme, we search for the closest English phoneme as well in the same way. Figure 4.9 illustrates the intra acoustic similarity of English and inter similarity between English and Cantonese. For each English phonemes, we list the closest English phonemes with the highest K-L score, and the closest Cantonese phonemes in IF and CanP-based sound inventories respectively.

	I_b	I_c	I_d	I_f	I_g	I_h	I_j	I_k	I_l	I_m	I_n	I_ng	I_null	I_p	I_s	I_t	I_w	I_z
E_b	2.35	0	0.07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E_ch	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E_d	0.35	0	0.52	0	0.54	0	0	0	0	0	0	0	0	0	0	0.05	0	0
E_dh	0	0	0.03	0.01	0	0	0	0	0.02	0	0	0	0	0	0	0	0	0
E_f	0	0	0	1.28	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0.07
E_g	0	0	0	0	1.48	0	0	0	0	0	0	0	0	0	0	0	0	0
E_hh	0	0	0	0	0	0.59	0.03	0	0	0	0	0	0	0	0	0	0	0
E_k	0.1	0.09	0	0	0.34	0	0	0.58	0	0	0	0	0	0	0	0.06	0	0
E_l	0	0	0	0	0.02	0	0	0	0.19	0	0	0.07	0.26	0	0	0	0	0
E_m	0	0	0	0	0	0	0	0	0.06	0.61	0.03	0	0.03	0	0	0	0	0
E_n	0	0	0	0	0	0.02	0	0	0.1	0.02	0.05	0.02	0.28	0	0	0	0	0
E_ng	0	0	0	0	0.02	0	0	0.02	0	0	0	0.03	0.28	0	0	0.01	0	0
E_p	0.65	0	0.07	0	0	0	0	0.13	0	0	0	0	0	0.12	0	0.37	0	0
E_r	0	0	0	0	0	0	0	0	0.55	0	0.06	0	0.03	0	0	0	0	0
E_s	0	0.19	0	0	0	0	0	0	0	0	0	0	0	0	1.42	0	0	0
E_sh	0	0.52	0	0	0	0	0	0	0	0	0	0	0	0	0.47	0	0	0
E_t	0.09	0.64	0.06	0	0.06	0	0	0.08	0	0	0	0	0	0	0	0.15	0	0
E_th	0	0.13	0	0.11	0	0	0	0.04	0	0	0	0	0	0	0.17	0	0	0
E_v	0.18	0.12	0	0.47	0.12	0	0	0	0	0	0	0	0	0	0.07	0	0	0
E_w	0	0	0	0	0	0	0.03	0	0.29	0	0	0	0	0	0	0	0.55	0
E_y	0	0	0	0	0	0	0.92	0	0	0	0	0	0	0	0	0	0	0
E_z	0	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0.72	0	0	0
E_zh	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0.16	0	0	0

Figure 4.8: Example of a K-L score matrix

In this study, if the score derived from its closest Cantonese phoneme is obviously larger than the score from its closest English phoneme (score difference > 0.2), this English phoneme is regarded as being closer to Cantonese phoneme than other English phonemes. As a result, less than half of English phonemes are found to be closer to Cantonese. Among these English phonemes, there are some more-to-one mapping, e.g., both /E_ey/ and /E_ih/ are found closer to Cantonese finals /F_ei/ than other English phonemes. Therefore, the English phoneme and its closest Cantonese phonemes will be clustered together only if the Cantonese phoneme is also found to be closer to this English phoneme than other English phonemes. On the other hand, if the K-L score between two English phonemes is very high, they will be merged together as well, such as /E_z/ and /E_s/. Details of merged phonemes are highlighted in Figure 4.9.

Selection of Cross-lingual Phoneme Inventory

Based on the acoustic and phonetic similarity investigated before, several cross-lingual phoneme inventories can be designed with different phoneme combination schemes described as follows:

a) PCM: phoneme combination based on the phonetic similarity observed in

English Phone	Optimal English phone & KLD	Optimal Cantonese phone & KLD	
		IF Based	CanP Based
E_b	0.92(E_p)	2.348(I_b)	2.348(CanP_b)
E_ch	0.38(E_s)	0.903(I_c)	0.903(CanP_c)
E_d	0.74(E_t)	0.538(I_g)	0.538(CanP_g)
E_dh	0.02(E_d)	0.033(I_d)	0.144(CanP_d)
E_f	0.92(E_v)	1.283(I_f)	1.283(CanP_f)
E_g	0.59(E_d)	1.478(I_g)	1.424(CanP_g)
E_hh	0.13(E_f)	0.593(I_h)	0.594(CanP_h)
E_jh	0.37(E_sh)	0.939(I_z)	0.939(CanP_z)
E_k	1.19(E_t)	0.577(I_k)	0.577(CanP_k)
E_l	0.54(E_r)	0.255(I_null)	1.069(CanP_ng)
E_m	1.58(E_n)	0.610(I_m)	1.968(CanP_m)
E_n	1.14(E_n)	0.282(I_null)	1.442(CanP_n)
E_ng	1.01(E_n)	0.282(I_null)	0.434(CanP_ng)
E_p	1.52(E_t)	0.648(I_b)	0.648(CanP_b)
E_r	1.13(E_l)	0.553(I_l)	0.437(CanP_ng)
E_s	1.83(E_z)	1.423(I_s)	1.423(CanP_s)
E_sh	0.80(E_s)	0.523(I_c)	0.523(CanP_c)
E_t	1.42(E_p)	0.635(I_c)	0.635(CanP_c)
E_th	0.32(E_f)	0.171(I_s)	0.171(CanP_s)
E_v	0.51(E_f)	0.472(I_f)	0.473(CanP_f)
E_w	0.26(E_r)	0.550(I_w)	0.55(CanP_w)
E_y	0.32(E_w)	0.917(I_j)	0.917(CanP_j)
E_z	2.01(E_s)	0.721(I_s)	0.721(CanP_s)
E_zh	0.40(E_sh)	0.161(I_s)	0.161(CanP_s)
E_aa	0.932(E_ao)	0.234(F_o)	0.921(CanP_o)
E_ae	1.867(E_eh)	0.318(F_aai)	0.593(CanP_e)
E_ah	0.940(E_er)	0.598(F_o)	1.1(CanP_o)
E_ao	1.010(E_aa)	0.177(F_o)	0.721(CanP_o)
E_aw	0.189(E_aa)	0.274(F_aau)	0.463(CanP_aa)
E_ay	0.403(E_ae)	1.276(F_aai)	1.466(CanP_aai)
E_eh	1.810(E_ae)	0.361(F_e)	1.279(CanP_e)
E_er	1.118(E_ah)	0.520(F_aa)	0.568(CanP_aa)
E_ey	1.000(E_ih)	1.835(F_ei)	2.49(CanP_ei)
E_ih	0.660(E_ey)	0.871(F_ei)	1.157(CanP_ei)
E_iy	0.878(E_ih)	2.229(F_i)	2.654(CanP_i)
E_ow	0.770(E_ah)	1.277(F_ou)	2.144(CanP_ou)
E_oy	0.081(E_ao)	0.045(F_oi)	0.097(CanP_o)
E_uh	0.139(E_ah)	0.179(F_ou)	0.384(CanP_o)
E_uw	0.456(E_er)	0.262(F_ou)	0.406(CanP_ou)

Figure 4.9: KLD based acoustic similarity

the phonetic confusion matrix

b) KLD: phoneme combination based on the acoustic similarity investigated in the KLD-based approach

c) KLD ∪ PCM: intersection of PCM and KLD-based phoneme combination results

d) KLD ∩ PCM: union of PCM and KLD-based phoneme combination results

Table 4.10 lists the details of merged phonemes in different combination schemes. As a result, various cross-lingual acoustic models can be trained with different sound inventories.

Table 4.11 explains different sets of acoustic models. Model IPA is a purely knowledge-based acoustic model based on IPA phonetic inventories for native American English and Cantonese. Four sets of data-driven cross-lingual acoustic models are trained for IF-based (KLD_IF, PCM_IF, UN_IF, IN_IF) and CanP-based (KLD_CanP, PCM_CanP, UN_CanP, IN_CanP) sound inventories, respectively. The prefix "UN_" represents union and "IN_" represents intersection of PCM and KLD-based phoneme combination. LD_IF and LD_CanP are language-dependent models, in which Cantonese and English phonemes are separated despite the fact that some of them are phonetically similar.

To evaluate the effectiveness of different sets of acoustic models and select the most appropriate phoneme inventory for code-mixing speech recognition, syllable/word recognition experiments are performed. All phoneme models are context-independent monophone HMMs trained with CUSENT and CUMIX. The acoustic feature vectors have 39 components: 13 MFCC and their first and second-order time derivatives. Each phone model consists of three or five emitting states, each of which is represented by 16 Gaussian mixture components.

The test data are the CM test utterances of CUMIX and recognition performance is measured in terms of syllable accuracy for Cantonese and word accuracy for English. The test results are also given in Table 4.11. The recognition performance of LD_IF and LD_CanP, particularly on embedded English words, are on the low side because of the relatively limited amount of training

Table 4.10: Merged phonemes in different combination schemes

IF-based System			
PCM	KLD	KLD ∪ PCM	KLD ∩ PCM
E_ay, F_aai	E_ay, F_aai	E_ay, F_aai	E_ay, F_aai
E_er, F_aa	E_iy, F_i	E_iy, F_i	E_ey, F_ei
E_ey, F_ei	E_ey, F_ei	E_ey, F_ei	E_ow, F_ou
E_ow, F_ou	E_ow, F_ou	E_ow, F_ou	E_z, I_s
E_uw, F_iu	E_b, I_b	E_b, I_b	
E_oy, F_oi	E_ch, I_c	E_ch, I_c	
E_l, I_l	E_g, I_g	E_g, I_g	
E_v, I_f	E_f, I_f	E_f, E_v, I_f	
E_z, I_s	E_z, E_s, I_s	E_z, E_s, I_s	
E_t, I_t	E_hh, I_h	E_hh, I_h	
	E_w, I_w	E_w, I_w	
	E_y, I_j	E_y, I_j	
	E_jh, I_z	E_jh, I_z	
		E_uw, F_iu	
		E_oy, F_oi	
		E_er, F_aa	
		E_t, I_t	
		E_l, I_l	

CanP-based System			
PCM	KLD	KLD ∪ PCM	KLD ∩ PCM
E_ay, CanP_aai	E_ay, CanP_aai	E_ay, CanP_aai	E_ay, CanP_aai
E_er, CanP_aa	E_iy, CanP_i	E_er, CanP_aa	E_ey, CanP_ei
E_ey, CanP_ei	E_ey, CanP_ei	E_ey, CanP_ei	E_f, CanP_f
E_ow, CanP_ou	E_ow, CanP_ou	E_ow, CanP_ou	E_iy, CanP_i
E_uw, CanP_iu	E_b, CanP_b	E_uw, CanP_iu	E_n, CanP_n
E_l, CanP_l	E_ch, CanP_c	E_l, CanP_l	E_ow, CanP_ou
E_t, CanP_t	E_g, CanP_g	E_t, CanP_t	E_z, CanP_s
E_f, E_v, CanP_f	E_f, CanP_f	E_f, E_th, E_v, CanP_f	
E_z, CanP_s	E_z, E_s, CanP_s	E_z, E_s, CanP_s	
E_oy, CanP_oi	E_hh, CanP_h	E_oy, CanP_oi	
E_ah, CanP_a	E_w, CanP_w	E_ah, CanP_a	
E_iy, CanP_i	E_y, CanP_j	E_iy, CanP_i	
E_d, CanP_d	E_jh, CanP_z	E_d, CanP_d	
E_n, CanP_n	E_l, CanP_ng	E_n, CanP_n	
E_ng, CanP_ng	E_m, CanP_m	E_ng, CanP_ng	
	E_n, CanP_n	E_m, CanP_m	
		E_b, CanP_b	
		E_ch, CanP_c	
		E_g, CanP_g	
		E_hh, CanP_h	
		E_w, CanP_w	
		E_y, CanP_j	
		E_jh, CanP_z	

Table 4.11: Effectiveness of different phoneme inventories

Monophone AMs	No. of Phones	Shared Phones	Recognition Accuracy on Code-mixing (%)		
			CM Overall	Can. Syllable	Eng. Word
IPA	65	17	38.75	38.86	37.79
LD_IF	112	0	42.85	44.12	31.98
KLD_IF	98	14	44.75	44.74	44.79
PCM_IF	102	10	45.33	45.35	45.17
UN_IF	92	20	44.91	45.27	41.86
IN_IF	108	4	46.29	45.94	49.24
LD_CanP	97	0	36.86	37.30	33.22
KLD_CanP	80	17	38.62	37.93	44.41
PCM_CanP	81	16	38.59	37.99	43.65
UN_CanP	71	26	38.24	37.78	42.18
IN_CanP	90	7	39.74	38.46	50.60

data and the language-dependent nature of the models. The English words in CUMIX carry Cantonese accents, such that some of the English phoneme models are very close to certain Cantonese phoneme models. In other words, similar acoustic features are captured by two different models. Hence, the confusion of English words with Cantonese syllables tends to increase. The English words are easily misrecognized as Cantonese syllables. At the same time, some cantonese syllables may be recognized as English words as well. This also explains why the performance of LD_IF and LD_CanP in recognizing Cantonese syllables slightly declines.

The IPA-based model shows the lowest performance in recognizing code-mixing speech in all cross-lingual acoustic models. Although it maintains a reasonable performance for Cantonese, it attains a very low accuracy of 37.79% for English. This confirms that Cantonese-accented English phonemes are different from the native American English phonemes defined in the IPA.

Data-driven based cross-lingual acoustic models can capture pronunciation variation more effectively. All data-driven models improve greatly in recog-

nizing English words. KLD-based and PCM-based models show comparable recognition results although almost half of the shared phonemes are different between them. UN-based models use a larger number of shared phonemes (union of KLD-based and PCM-based merged phonemes) between Cantonese and English. However, their recognition accuracies on embedded English are on the low side compared with either KLD-based or PCM-based models. IN-based models attain the best recognition performance for both code-mixing Cantonese and embedded English. Only a few shared phonemes which are found highly acoustically and also phonetically similar are implemented in IN-based models. This indicates that very few Cantonese-accented English phonemes tend to really highly resemble or even become identical to their Cantonese counterparts. It also suggests that cross-lingual models should be applied to highly acoustically and also phonetically similar phonemes, while language-specific models would be more appropriate if either the phonetic or acoustic variation is relatively large.

All IF-based cross-lingual acoustic models outperform CanP-based models in code-mixing speech recognition. The recognition accuracies on code-mixing Cantonese attained by IF-based models show above 7% consistent improvement compared with CanP-based models. This is because the initial-final scheme is Cantonese-specific. It can better preserve the language-specific characteristics. On the other hand, IF-based and CanP-based models show similar performance with regard to recognizing code-mixing English. This suggests that initials and finals are more suitable as the basic Cantonese units than the phonemes in Cantonese-English code-mixing speech recognition applications.

On the whole, IN_IF is the most appropriate sound inventory for acoustic modeling of Cantonese-English code-mixing speech. It attains the best recognition accuracy of 45.94% for Cantonese, and at the same time, it gives a satisfactory result of 49.24% for embedded English. Therefore, context-dependent triphone models will be trained with IN_IF in the next step.

4.3.2 Development of Context-dependent Acoustic Models

To represent the different contextual effects, context-dependent triphone models are further developed based on the pre-defined cross-lingual phoneme inventory **IN_IF**. Constrained by limited training data, we need to tie (cluster) models of rich contexts into generalized ones for predicting unseen contexts in test utterances. A decision tree-based clustering approach is used to decide which states to tie. To capture the co-articulation effects in code-mixing speech, states with different contexts of different languages are allowed to be tied together. We construct questions to jointly tie states mainly based on their manners and places of articulation. Findings in previous studies on phonetic and acoustic similarity are also implemented. Different types of questions designed for decision tree-based clustering include:

Mono-lingual Questions

a) language-dependent questions: e.g. "L_E_Nasal", does the left context belong to an English nasal, such as /E_m/, /E_n/ or /E_ng/?
b) language-specific questions: e.g. R_Voiced_Stop, does the right context belong to an voiced stop, such as /E_b/, /E_d/, or /E_g/? All the attributes of the question only exist in English, i.e. there are no Cantonese voiced stops.

Cross-lingual Questions

a) general language-independent questions: e.g. R_CL_Affricate, does the right context belong to an affricate, such as /E_ch/, /E_jh/, /I_z/, or /I_c/?
b) particular language-independent questions: e.g. "L_CL_Phone_Class4", does the left context belong to phone class 4, such as /E_f/, /E_v/, or /I_f/? All attributes of the question are phonetically or acoustically similar phonemes observed in sound inventory **UN_IF**.

Training of Context-dependent Models

As a result, two cross-lingual (CL) context-dependent (CD) tie models are trained as shown in Table 4.12. Only mono-lingual questions are used in growing trees in model CL_A, while both mono-lingual and cross-lingual questions are implemented to train model CL_B. IPA-based CD models are also trained with cross-lingual decision trees as the benchmark.

Table 4.12: Number of used questions and tied states in different CD acoustic models

Model	No. of used questions	No. of tied states
CL_IPA	292	40,503
CL_A	458	63,526
CL_B	510	71,190

The effectiveness of CL_IPA, CL_A, and CL_B are evaluated by syllable/word recognition experiments. No language model is applied. The test data include the CM and the MC test utterances of CUMIX. The grammar network used for recognizing CM utterances is illustrated in Figure 4.10. For MC utterances, the recognition network is simplified into a syllable loop.

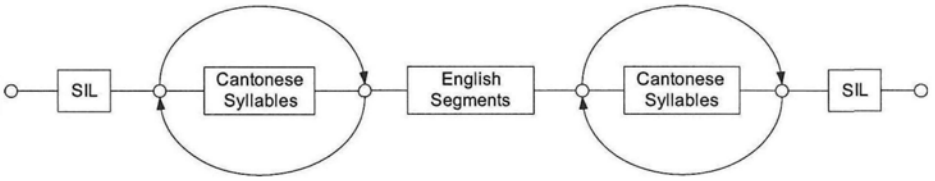


Figure 4.10: Grammar network for syllable/word recognition of code-mixing speech

The recognition performance is measured in terms of syllable accuracy for Cantonese and word accuracy for English. The test results are given in Table 4.13. It is not surprising that poor recognition results are found in model CL_IPA. Significant improvement can be observed in models CL_A and CL_B due to the better designed cross-lingual phoneme inventory. CL_B outperforms CL_A in recognizing code-mixing speech. This is because CL_B applies pre-set

cross-lingual phonetic questions in addition. Hence, the speech context at the language boundaries can be used more efficiently. In other words, more speech data are available to train embedded English words in code-mixing. Further improvement can be expected for embedded English words and neighbouring Cantonese syllables in consequence.

Table 4.13: Syllable/word accuracy of the three context-dependent acoustic models

Triphone AMs	Cantonese-English code-mixing			Monolingual Speech	
	CM Overall	Can. Syllable	Eng. Word	Cantonese	English
IPA	57.3%	57.6%	54.1%	62.4%	72.6%
CL_A	62.0%	62.0%	61.8%	65.9%	74.5%
CL_B	62.1%	62.1%	62.6%	65.6%	75.2%

There is a 3-5% accuracy degradation from monolingual Cantonese to code-mixing Cantonese. This is because the grammar network used for monolingual Cantonese utterances does not include an English segment, and therefore there should be no recognition error caused by confusion between similar Cantonese syllables and English words.

In order to get the upper bound accuracy of English words, another experiment is performed on the code-switch words only and the dictionary only includes English words. The test data are the embedded English segments extracted from the code-mixing utterances which means that the language boundary information is correct. CL_B also attains the best recognition accuracy of 75.2% in this recognition task.

4.4 Towards Cross-lingual Adaptation

The major cost factor for developing speech recognition systems for new languages or speakers is the large amount of transcribed speech data that is required for the training of accurate acoustic models. However, cross-lingual adaptation makes it possible to make use of speech resources available in one or more lan-

guages for the recognition of another target language. This allows fast and low-cost implementation of speech recognizers, which is especially useful for minority languages or dialects, in which data resources available are very limited or even not existent [114].

As we discussed earlier, Hong Kong is a bilingual society, where Chinese and English are the official spoken languages. As a major working language in Hong Kong, English is widely used in commercial activities and legal matters. The usage of English, however, is much less than Cantonese in general conversational communication. In most cases, it is much easier to collect a small amount of Cantonese speech data from a specific Cantonese speaker for speaker adaptation purposes. Of course, it is also time and labour saving if we can perform adaptation on more than one language by using only monolingual speech data captured from a desired speaker.

This part of research aims at making use of acoustic information extracted from an existing source language (Cantonese) to implement speaker adaptation for a new target language (English). It is assumed that English adaption data is not available for the target speaker. Speaker-independent (SI) and language-dependent acoustic models are trained for Cantonese and English respectively in the first step. Based on SI acoustic models, model mappings between Cantonese and English acoustic units can be established. For each English unit, we expect to find its closest Cantonese units, and vice versa. With model mapping, speaker adaptation of English models can be implemented by using Cantonese adaptation data from the target speaker.

4.4.1 Model Mapping between Cantonese & English

One of the major problems in this study is the model mapping between different languages. The mapping can be established at different acoustic levels, such as words, syllables, phones and others [115]. As shown in Figure 4.11, we investigated the use of phones, states, and Gaussian mixture components for such purpose.

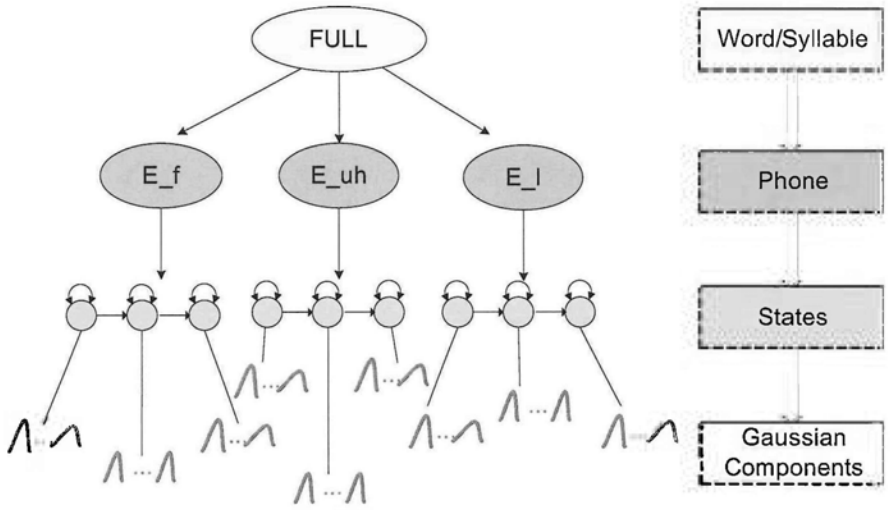


Figure 4.11: Mapping at different levels of acoustic units

Phone Mapping

Phones are most widely used as the basic acoustic unit for model mapping [116]. The phone mapping table can be manually generated based on linguistic knowledge or automatically derived in a data-driven manner [117].

As an international standard of representing speech sounds of any spoken language, the International Phonetic Alphabet (IPA) is used to create the mapping table. It classifies phones in terms of place and manner of articulation. Phones of different languages labelled by the same IPA symbol are considered as the same phone. However, as discussed in Section 3.3, the phone sets of Cantonese and English are significantly different. Only 17 phones can be shared according to their IPA symbols, 22 English-specific phones and 26 Cantonese-specific phones remain distinctively different. Similarity between these remaining phones can nevertheless be measured by their acoustic distributions. In this study, Kullback-Leibler Divergence (KLD) is used. The phone mapping for the language-specific phones is created by Eq.(4.5). Each phone is modelled by speaker-independent, context-independent HMMs with single Gaussian distribution.

$$\hat{P}^c = \arg\min_{P^c} D_{KL}(P_i^c, P_j^e) \tag{4.5}$$

where, P_i^c is a phone in the Cantonese phone set P^c, P_i^e is a phone in the English

phone set, and D_{KL} is the K-L divergence between two phones.

State-level Mapping

Since Cantonese and English are two languages highly uncorrelated phonetically, there might not exist much similarity in their phonemic counterparts. However, since speech production is constrained by limited movement of articulators, it might be possible to find some similar acoustic units at a refined, sub-phone level. Diphthongs may be rendered by several monophthongs. Furthermore, allophones, which are highly context dependent, provide more chances for phone sharing between different languages. As a result, a tied, context-dependent state-level mapping is investigated between Cantonese and English. First, we build two speaker-independent, language-specific decision trees for Cantonese and English respectively. Each leaf node in the decision tree represents a tied state, modeled by a Gaussian distribution. For each English tied states, a corresponding Cantonese tied state can be found, in the minimum KLD sense. The directional mapping from English to Cantonese can be in the form of one-to-many mapping. Different leaf nodes in the English tree may map to the same leaf node in the Cantonese tree.

In order to achieve satisfactory recognition performance, multiple Gaussian mixture components are typically used. Similarity among states may change along with the mixture splitting. Therefore, two model mapping schemes are studied, namely the **State Mapping** and **CalState Mapping**.

In **State Mapping**, mapping is estimated based on the tied states with single Gaussian models, and the mapping does not change with the incrementation of mixture components. In **CalState Mapping**, the model distribution of each state is recalculated by Eq.(4.6) in multiple mixture cases:

$$CalS_i = \sum_{k=1}^{K} w_{ik} S_{ik} \tag{4.6}$$

where w_{ik} is the mixture weight of the kth mixture of state S_i, S_{ik} is the output distribution of the kth mixture in state S_i and $CalS_i$ is the recalculated distribution of state S_i. Then, the mapping is established based on the distribution

$CalS_i$. **CalState Mapping** is created repeatedly with the increase of mixtures, until the desired number of mixture components is reached.

Mapping among Gaussian Mixture Components

In multiple-mixture HMMs, Gaussian mixture components are the smallest elements. Furthermore, in speaker adaptation based on MLLR or MAP algorithms, adaptation is usually applied to individual mixture components in the model set. In this thesis, Gaussian component mapping, denoted by **GauMix Mapping**, is investigated between Cantonese and English. For each mixture component in English, the corresponding one in Cantonese is found by minimizing the K-L divergence:

$$\hat{M}^c = \arg\min_{M^c} D_{KL}(M_i^c, M_j^e) \tag{4.7}$$

where M_i^c is a mixture component in Cantonese model set M^c, M_j^e is a mixture component in English models, and D_{KL} is the K-L divergence between two mixture components. Similar to **CalState Mapping**, **GauMix Mapping** needs to be re-estimated as the number of mixtures increases.

4.4.2 Cross-lingual Adaptation via Mapping

In the last two decades, many speaker adaptation techniques have been successfully applied [118][119]. HMM-based adaptation using MLLR or MAP techniques can be used to improve recognition performance using a small amount of adaptation data from target speakers [119]. However, it is not easy to do it across different languages, especially when two languages are phonetically distant.

In this study, cross-lingual speaker adaptation is implemented via the model mapping strategy as established earlier. With model mapping, speaker adaptation on English models can be implemented by using Cantonese adaptation data from the target speaker. Figure 4.12 gives an example of cross-lingual adaptation with **GauMix Mapping**. We want to adapt English SI models with Cantonese adaptation data from speaker A. Since the SI mapping has been created, HMM parameters for each mixture component of the English

models can be replaced by the Cantonese counterpart. When standard intra-language adaptation is implemented on Cantonese SI models with Cantonese adaptation speech from speaker A, we obtain the respective Cantonese speaker adapted (SA) models. Then, English SA models can be retrieved by Cantonese SA models via model mapping.

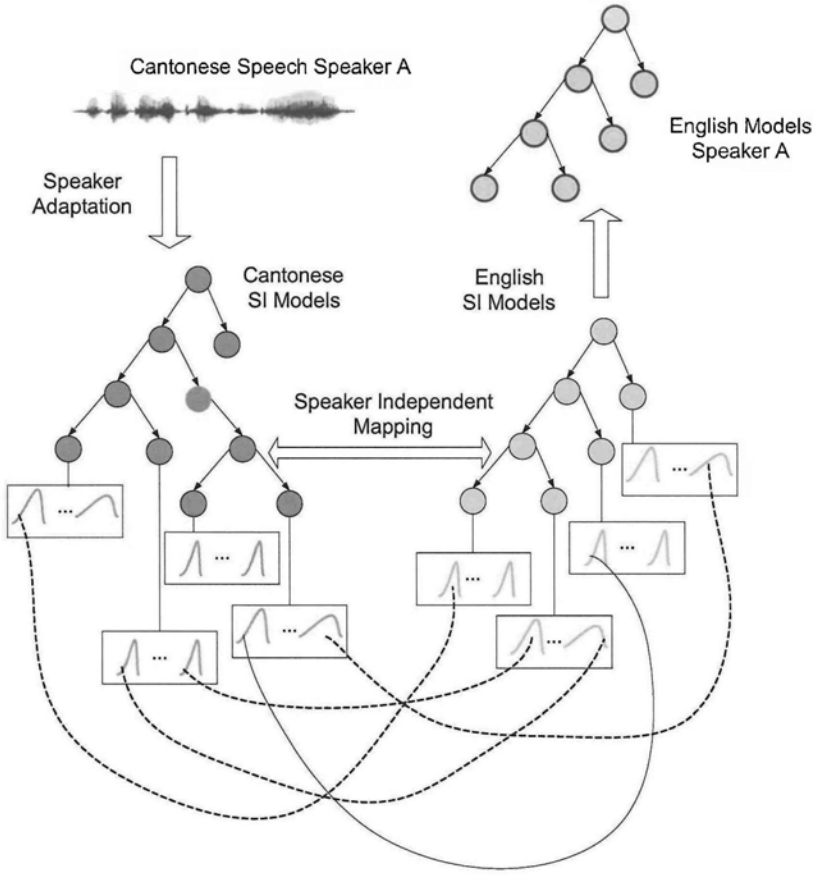


Figure 4.12: Cross-lingual speaker adaptation via Gaussian mixture component mapping

4.4.3 Experimental Setup

SI Acoustic Models

The speaker-independent, language-dependent acoustic models for model mapping are trained with the utterances from 60 speakers in the CUMIX corpus.

Each speaker provides 10 minutes of Cantonese and 4 minutes of English speech data. The acoustic feature vectors consists of 12 MFCC coefficients, log energy, and their first and second-order derivatives. The models are trained from context-independent monophone HMMs to context-dependent triphone HMMs, from single Gaussian to 8 mixture components. It is expected that these language-dependent acoustic models can preserve the characteristics of Cantonese-accented English and colloquial Cantonese respectively due to the nature of the training data.

Speaker Adaptation Setup

A cross-lingual adaptation experiment is performed on 14 speakers from the CUMIX corpus, different from the 60 speakers participating in SI acoustic modeling. For each speaker, 4 minutes of Cantonese adaptation speech data are available. We used MLLR followed by MAP adaptation techniques in this task. The adapted models are evaluated in speech recognition experiments. The test speech are English words and phrases extracted from the code-mixing utterances (CM). The vocabulary size is 1.2k, and no language model is applied.

4.4.4 Results & Discussions

The performance of the proposed cross-lingual speaker adaptation system is mainly determined by two factors: model mapping performance and speaker adaptation efficacy.

Model Mapping Effectiveness

Model mapping performance is measured in terms of mapping effectiveness, which is evaluated by recognition performance with mapped acoustic models. Mapped SI models can be retrieved by Cantonese SI models via mapping. The recognition results of different mapped SI models are shown in Figure 4.13. Results from English SI models are also given as a reference. The performance degradation from English acoustic models to mapped models is defined as the

mapping loss in this study. The lower the mapping loss, the higher the mapping effectiveness.

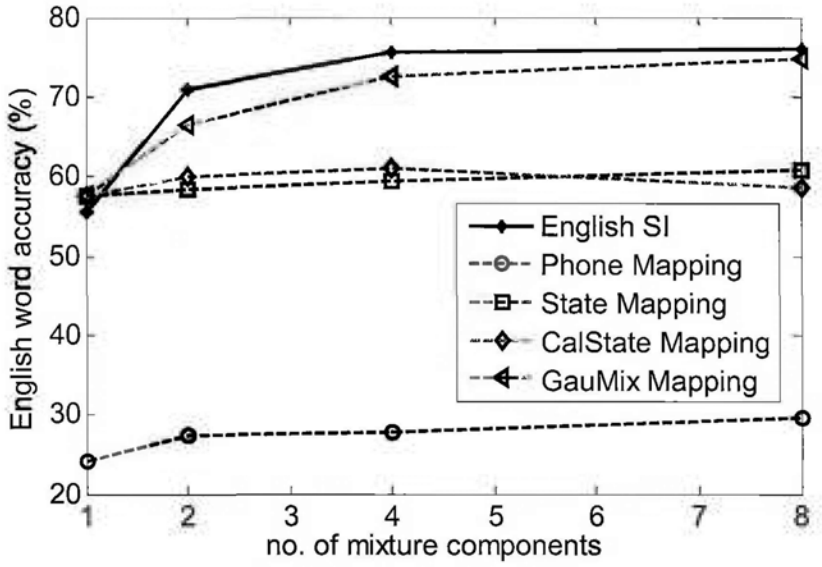


Figure 4.13: Recognition results from different SI models

From the recognition results, it is clear that model mapping effectiveness increases with the refinement of mapping units. Due to the significant phonetic difference between Cantonese and English, it is not surprising that poor recognition results are found in the **Phone Mapping** approach. State-level mapping systems improve greatly the recognition of English words due to better matched model mapping. **CalState Mapping** and **State Mapping** schemes show similar performance. However, the performance variation from single mixture to multiple mixture is not obvious. On the other hand, **English SI** models give improved performance with increasing number of mixtures. The mapping loss found in state-level mapping increases with the mixture component splitting, while model mapping effectiveness of state-level mapping is still on the low side. The recognition performances of **GauMix Mapping** models however improves with increasing mixture number, and insignificant mapping loss is found between **English SI** model and **GauMix Mapping** ones. That means the acoustic space represented by gaussian components from different languages is actually close to each other.

Further analysis is done by comparing the effective Cantonese states found in different model mapping strategies. There are 3,248 Cantonese candidate states and 969 English states available in SI acoustic models for mapping creation. Table 4.14 compares the effective Cantonese states found in **CalState Mapping** and **State Mapping**. As the number of mixtures increase, effective Cantonese states in **CalState Mapping** and **State Mapping** become more and more dissimilar. However, the total number of effective Cantonese states is almost unchanged with mixture splitting. This may be the reason why the recognition performance of the state-level mapped models is always the same even under different mixture component cases. The mapping details for **GauMix Mapping** are given in Table 4.15. It is clear that the number of effective Cantonese states increases with mixture splitting. The resolution of **GauMix Mapping** is higher than state-level mapping. Even for fairly distant states, we may still find similar mixture components among them. This also explains why **GauMix Mapping** outperforms the state-level mapping approach.

Table 4.14: Cantonese states found in **CalState Mapping**.

	1 mix	2 mix	4 mix	8 mix
used states	566	571	590	591
shared states with **State Mapping**	566	409	356	272

Table 4.15: Cantonese states found in **GauMix Mapping**.

	1 mix	2 mix	4 mix	8 mix
used mixtures	566	1,125	2,266	4,455
used states	566	947	1,491	2,076

Cross-language Speaker Adaptation Results

The cross-lingual speaker adaptation results are summarized in Table 4.16. If the mapping loss exceeds the speaker adaptation improvement, the performance obtained with cross-lingual speaker adaptation is even worse than the monolingual SI recognizer, such as the speaker adaptation results found with **State**

Mapping. Mapping between Gaussian mixture components has been proved effective in speech recognition task in earlier studies. Via **GauMix Mapping**, the adaptation models give an average 78.70% word accuracy for English over all speakers. A relative 10.13% reduction in word error rate (WER) is achieved, compared with 76.30% word accuracy obtained with English SI models. Furthermore, it is found that the effectiveness of speaker adaptation is highly correlated to mapping effectiveness. If mapping loss is ignored, the improvement from adaptation is very limited in **State Mapping**. However, the adaptation leads to a relative 13.67% error reduction via **GauMix Mapping**, compared with the mapped SI models. The degree of improvement from speaker adaptation increases with improved model mapping effectiveness.

Table 4.16: Cross-lingual adaptation results for individual speakers (% word accuracy); 4 minutes of Cantonese adaptation speech are used.

		State Mapping		**GauMix Mapping**	
Speaker	SI	Mapped	Adapted	Mapped	Adapted
spkr1	73.42	58.23	56.96	73.42	79.75
spkr2	72.00	56.00	54.67	70.67	76.00
spkr3	64.38	56.16	60.27	69.86	68.49
spkr4	79.27	62.20	65.85	74.39	84.15
spkr5	83.95	67.90	72.84	85.19	83.95
spkr6	89.61	68.83	67.53	83.12	89.61
spkr7	78.31	55.42	46.99	81.93	78.31
spkr8	79.45	64.38	67.12	79.45	83.56
spkr9	77.50	62.50	70.00	80.00	82.50
spkr10	80.28	63.38	69.01	77.46	81.69
spkr11	67.95	55.13	53.85	60.26	65.38
spkr12	74.65	50.70	50.70	76.06	81.69
spkr13	70.00	51.25	56.25	67.50	68.75
spkr14	76.62	64.94	64.94	76.62	77.92
Ave	**76.30**	**59.81**	**61.20**	**75.46**	**78.70**

Speaker adaptation experiments have also been performed on an intra-language basis for reference purpose. By applying the same adaptation data, the relative error reduction for Cantonese is 19.21% on average. The experi-

mental results show that our approach for cross-lingual speaker adaptation is promising.

Further analysis is carried out by investigating adaptation using different amounts of data. For each speaker, there is only 4 minutes of adaptation speech data. We divide it into 3 adaptation subsets, which contain 2, 3 and 4 minutes of speech respectively. The adaptation results are shown in Figure 4.14. It is clear that there is noticeable improvement with increasing adaptation data. It is believed that further improvement can be achieved if there is more adaptation data available. In addition, if a speaker-dependent Cantonese recognizer exists for a particular speaker, the corresponding speaker-dependent English recognizer can be implemented via the proposed **GauMix Mapping** in a fast and low cost way.

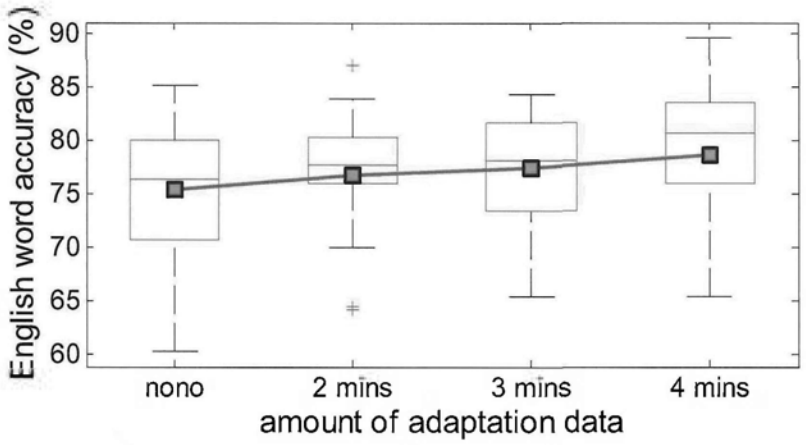


Figure 4.14: Boxplots for cross-lingual speaker adaptation results with different amounts of adaptation data, pooling all target speakers

□ End of chapter.

Chapter 5

Language Modeling for Cantonese-English Code-mixing

Summary

This chapter addresses the problem of language modeling for LVCSR of Cantonese-English code-mixing utterances spoken in daily communications. We start by collecting training text from the internet. Both monolingual colloquial Cantonese text and code-mixing text are collected for code-mixing language modeling. In the absence of sufficient amounts of code-mixing text data, different language modeling techniques are investigated. Class-based language models are developed with automatically generated classes. A semantics-based n-gram mapping scheme is developed to increase the counts of code-mixing n-grams at language boundaries. The Cantonese-to-English translation dictionary and semantics-based classes are developed as the reference for n-gram mapping. As a result, various semantics-based language models are trained. All models are evaluated in terms of perplexity.

5.1 Text Data for Cantonese-English Code-mixing Language Modeling

5.1.1 Data Collection

In general, language modeling requires a large amount of training text data. There are practical difficulties in collecting a large amount of text materials to facilitate statistical language modeling for Cantonese-English code-mixing speech. Cantonese is a spoken dialect, while the published media are dominated by standard Chinese text. On the other hand, colloquial Cantonese is neither taught in schools nor recommended for official and documentation usage. Only a limited amount of colloquial Cantonese text data can be found in certain columns of local newspapers and magazines, advertisements, and online articles [120].

In our study, the training text data for code-mixing language models are mainly collected from three major sources, namely newspapers, magazines and online diaries. Preliminary manual inspection is performed to identify the sections or columns that are highly likely to contain code-mixing text. Some Cantonese function characters that are frequently used in spoken Cantonese but rarely used in standard Chinese are used to query colloquial Cantonese data. We list these characters in Figure 5.1. In addition, some English words which frequently appear in Cantonese-English code-mixing are selected as the keywords to gather code-mixing data.

Colloquial Cantonese Function Words
啦, 喇, 啊, 呢, 咁, 咗, 嗰, 乜, 仲, 係, 嘅, 佢, 呃, 黎, 冇, 埋, 咩, 吓, 呀, 唔, 囉, 係, 喺, 晒, 黎, 嚟, 俾, 野, 嘢, 喎, 吓, 啲, 地, 咪, 啲, 緊, 攪, 掂

Figure 5.1: Keywords for collecting colloquial Cantonese data

During the data collection process, it is observed that the colloquial Cantonese terms and standard Chinese terms may mix together in written text, and these types of text are not suitable for code-mixing language modeling. As a

result, post-filtering is needed to remove these types of sentences. Figure 5.2 shows a list of colloquial Cantonese terms and standard Chinese terms selected for post-filtering. Collected sentences are filtered by colloquial Cantonese at first. A sentence is retained if it contains any two of the colloquial terms in Figure 5.2. After that, the remaining sentences are filtered by standard Chinese terms, in which sentences containing any one of the standard Chinese terms are further removed.

Colloquial Cantonese Terms	啦, 喇, 啊, 呢, 咁, 勁, 架, 左, 咗, 個, 嗰, 乜, 仲, 係, 既, 嘅, 佢, 呃, 黎, 有, 冇, 無, 埋, 咩, 吓, 邊, 呀, 唔, 囉, 係, 喺, 哂, 晒, 黎, 嚟, 來, 比, 俾, 睇, 得, 返, 搵, 野, 嘢, 未, 喎, 好, 至, 吓, 下, 先, 話, 依, 住, 幾, 哋, 地, 到, 度, 阿, 再, 番, 咪, 而, 諗, 夠, 意, 攞, 多, 少, 又, 同, 啲, 過, 緊, 丫, 岩, 攪, 掂, 我, 你, 果, 都, 喔, 边
Standard Chinese Terms	的, 們, 那, 這, 哪, 在, 不, 是, 了, 些, 他, 她

Figure 5.2: Selected colloquial Cantonese terms and standard Chinese terms for text filtering

In this thesis, a text corpus containing about 9 million Chinese characters and 300k English words is collected. It can be divided into two subsets. The first set is considered as monolingual colloquial Cantonese text. It contains 6 million Chinese characters. The second set of text data is real Cantonese-English code-mixing data, which contains 3 million Chinese characters and 300k English words. All text data in the corpus are segmented with a bilingual lexicon by the maximum matching method. The lexicon contains 16k words, comprising 12k Cantonese items and 4k English items. After segmentation, this corpus is ready for code-mixing language modeling.

5.1.2 Data Sparsity Problem

As we described above, due to the difficulty of data collection, only a small amount of colloquial Cantonese and code-mixing text data (~10 millions) are collected for code-mixing language modeling in this study. Compared with other

conventional monolingual text corpora such as WSJ corpus, Google Web 1T 5-gram corpus and CUSENT text corpus, the size of collected code-mixing text corpus is quite small [121] [96] [34]. Furthermore, the domain-specific property in code-mixing makes the data sparsity problem more serious.

Lack of sufficient amounts of text data lead to some problems for code-mixing language modeling. Not all the embedded English word in the speech data are found in the training text, creating Out-of-Vocabulary (OOV) words. Due to the data sparsity, many word sequences appearing in test speech may not be observed in training text data, especially in the language boundary (LB) context. Table 5.1 shows the n-gram coverage of the training corpus. It is clear that the data unseen problem is serious in the context of language boundary, particularly in the high order n-gram case. To deal with the problem of inadequate code-mixing training data, different language modeling techniques including class-based LM and semantics-based LM are investigated in this thesis. The English words are handled differently in these language models, and the details are discussed in the following sections.

Table 5.1: N-gram coverage of the Cantonese-English code-mixing training text

	1-gram	2-gram	3-gram
Cantonese context	99.77%	89.61%	57.90%
English or LB context	86.80%	34.81%	8.64%

5.2 Class-based Language Models

Class-based language modeling is one of the state-of-the-art approaches to handle the data sparsity problem. Word entries with similar meaning or syntactic function can be clustered into the same classes, either manually or by data-driven methods [122][123], such that the probabilities of class sequences are estimated instead of probabilities of word sequences in class-based LMs. In general, the number of classes is much lower than the number of words; therefore fewer parameters can be estimated more precisely from limited training data.

5.2.1 Automatic Clustering of Cantonese and English Words

Automatic clustering is the most widely used approach to construct a class map which defines which words are in each class. Different algorithms have been proposed to derive a class map in the last two decades [123][124]. In this study, an incremental greedy merging algorithm is used [123]. It starts with one class each for the C most frequent words and then adds one word at a time, where C denotes the target number of classes. Classes are optimized in terms of perplexity based on bi-gram statistics.

In class-based code-mixing language models, Cantonese and English words are not distinguished in the clustering process. A class can contain both Chinese and English words. As a result, three types of word classes can be found in class-based models. Two of them are monolingual classes which contain either Cantonese terms or English words. The third type of class is mixed-language class, which contains both Cantonese and English words in the same class. Figure 5.3 shows some examples of classes that we found particularly interesting. Some classes group together words having similar meaning, such as 每天, 每日. Other classes contain words that are syntactically similar, such as 反而, *thus*. It is believed that these "meaningful classes" may help the data sparsity problem by estimating the unseen code-mixing N-gram with the seen monolingual N-grams.

Categories	Eamples
Toponym	大嶼山/曼谷/馬尼拉/東京/上水/馬鞍山/廣州/海洋公園
Everyday	每天/每日
Sequence	較早前/日前
Activity	茗/酒會/講座/歡送會/典禮/研討會/埶/儀式
Action	restore/download/divert/debug/quote/refund/apply/capture/save/minor/repair/upload/copy
IT devices	等離子電視/mac/palm/cassette/mailing/walkman/compute/thinkpad/wii/iphone/inkjet/desktop/dc/pda
Adversative conjunction	反而/thus
Understanding	understand/了解/尊重/信任/appreciate/respect
Place	餐廳/露天/rooms/dream house/藥房
Occupation	大學生/廠商/運動員/captain/侍應

Figure 5.3: Some examples of word classes

5.2.2 Training of Class-based Code-mixing LMs

Class-based code-mixing language models are trained according to the class map from the automatic clustering process. Table 5.2 explains different sets of class-based 3-gram language models. The difference among them is that they are trained with different numbers of classes. It can be observed that there are very few monolingual English classes and most English words are clustered with Cantonese lexical items. With an increase of the target number of classes C predefined in clustering, the increment of monolingual Cantonese classes is higher than that of mixed-language classes. Hence, the percentage of mixed-language classes decreases with an increase in the total number of classes.

Table 5.2: Different class-based language models

Class LM	no. of class	% of mixed-language class	% of mono. Can. class	% of mono Eng. class
C-250 LM	250	86.4	13.6	0
C-500 LM	500	74.6	25.4	0
C-1000 LM	1,000	60.9	39.1	0
C-1500 LM	1,500	49.3	50.3	0.4
C-2000 LM	2,000	43.7	56.4	0.9
C-2500 LM	2,500	37.8	60.8	1.4
C-3000 LM	3,000	33.2	64.5	2.3

It is inappropriate to assume that an utterance must contain English words. The performance of class-based LMs is evaluated for both monolingual Cantonese and code-mixing speech. The test data include the MC and CM test utterances of CUMIX. Character perplexity is used as a performance index for monolingual Cantonese. For code-mixing speech, the perplexity is measured in terms of Character for Cantonese and word for English. Perplexity of different class-based LMs for monolingual Cantonese and code-mixing data are shown in Figure 5.4. The results of word-based 3-gram LM are also included as a reference.

The results in Figure 5.4 support that class-based language models can help

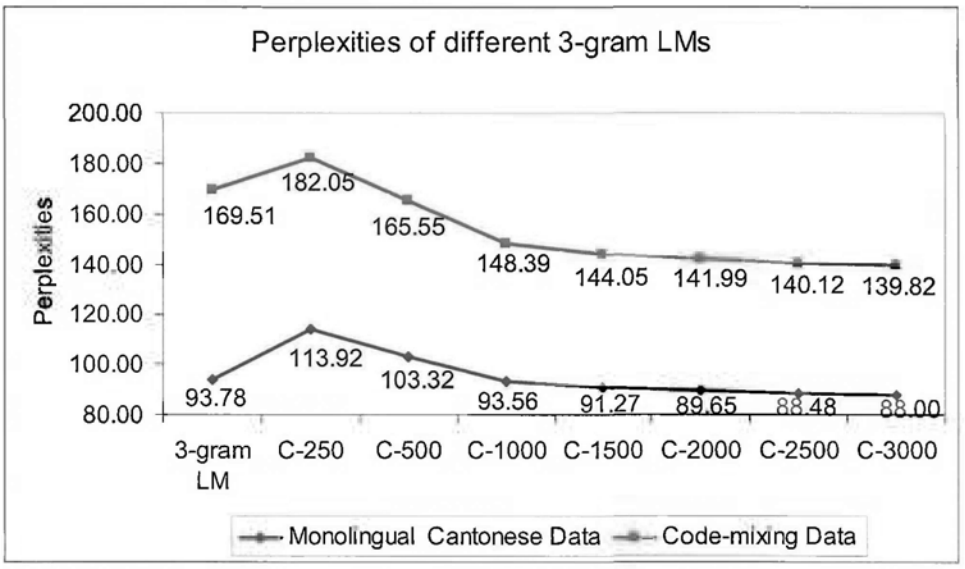


Figure 5.4: Perplexities of different language models for monolingual Cantonese and code-mixing data

to improve the perplexity, especially when the number of classes is large enough (more than 1000). However, if the number of classes is relatively small, such as 250 classes, the perplexities of class-based **C-250 LM** even increase for both monolingual Cantonese and code-mixing data, when compared with that of conventional word-based **3-gram LM**.

For different class-based language models, the perplexity reduces with an increase in the number of classes, and the perplexity reduction becomes smaller and smaller with increasing classes. When the number of classes is larger than 1,500, only slight improvement can be achieved when the number of classes is increased. The trends of perplexity reduction of different class-based LMs are the same for monolingual Cantonese and code-mixing data, although the class-based LMs show more significant perplexity improvement on code-mixing data than monolingual Cantonese data.

As Cantonese terms are dominant in code-mixing utterances, the perplexity measurement may be overwhelmed by Cantonese. Hence, perplexity only gives limited information on the performance of language models on code-mixing utterances. The performance of language models will also be evaluated in the LVCSR task in the next chapter.

5.3 Semantics-based Language Models

As we introduced in Section 2.1.3, statistical language models are widely used to capture the regularities of the language and provide linguistic constraints on the speech recognition outputs. For example, we can easily observe that P(你,好,嗎) > P(泥,郝,媽) in most conventional Chinese 3-gram LMs because word sequence *(你 好 嗎)* is more reasonable than *(泥 郝 媽)* in Chinese.

In the absence of sufficient amounts of code-mixing text data, it would be very helpful if we can reliably estimate the unseen code-mixing n-grams with the seen monolingual or code-mixing ones. As can be seen in Figure 5.5, 下午茶 is the Cantonese translation of *afternoon tea*, such that the word sequence *(食 下午茶)* and *(食 afternoon tea)* may capture similar linguistic properties. Hence, it is feasible to predict a code-mixing bi-gram *(食 afternoon tea)* with a monolingual bi-gram *(食 下午茶)*. On the other hand, a reasonable estimation of P(係, CC) can be obtained with P(係, NA) because words NA and CC are syntactically similar in these utterances (NA and CC are the two colleges of the Chinese university of Hong Kong).

In this thesis, an n-gram mapping approach is developed. Translation-based and semantics-based mapping are applied to increase the counts of code-mixing n-grams at language boundaries, such that it can better estimate the probability of low-frequency and unseen mixed-language n-gram events. In translation-based mapping schemes, the Cantonese-to-English translation dictionary is adopted to transcribe monolingual Cantonese n-grams to mixed-language n-grams. In semantics-based mapping schemes, n-gram mapping is based on the meaning and syntactic function of the English words in the lexicon.

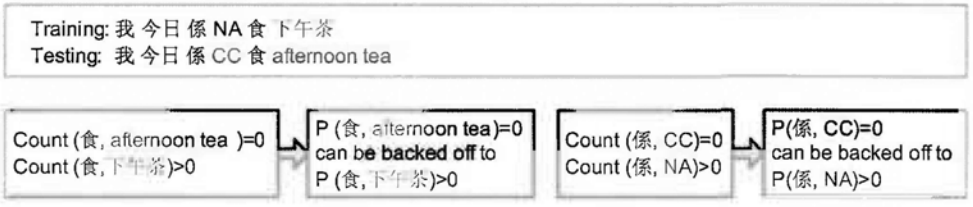


Figure 5.5: An example of a reasonable estimation of code-mixing unseen n-grams with seen n-grams

The details of the n-gram mapping approach is shown in Figure 5.6. Firstly, word n-grams are generated based on the text data in the training corpus. The Cantonese-to-English translation dictionary or semantics-based classes are applied as the reference for n-gram mapping. For example, if a Cantonese-to-English translation dictionary is used, the original monolingual Cantonese n-grams can be transcribed to mixed-language n-grams. After that, the original n-grams and the mapped n-grams are merged together and statistical N-gram language models will be trained with merged n-grams. In semantics-based mapping schemes, additional mixed-language n-grams can be estimated with the seed n-grams at language boundaries, if embedded English words in new n-grams and seed n-grams belong to the same semantics-classes. As a result, more mixed-language n-grams can be used for model training after mapping, and therefore the proposed language models are expected to better estimate the probability of low frequency and unseen mixed-language n-gram events.

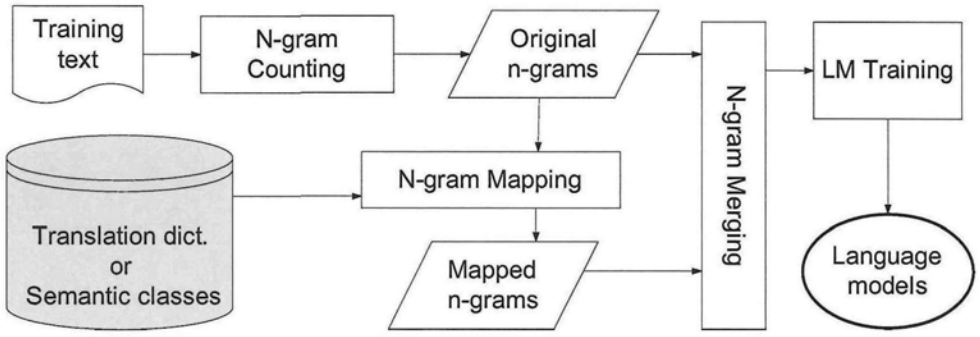


Figure 5.6: Block diagram for semantics-based LM via n-gram mapping

5.3.1 Translation-based Mapping

The effectiveness of translation-based mapping is mainly based on the Cantonese-to-English dictionary. As we mentioned before, there are obvious lexicon differences between Cantonese lexical items and standard Chinese lexicon entries. The existing Chinese-to-English dictionary is based on standard Chinese, which is not applicable in this task. Therefore, a Cantonese-to-English dictionary is developed for this study. Each English word may map to one or

more colloquial Cantonese terms. The established dictionary covers about 95% of the embedded English words in the lexicon, since not all embedded English words have Cantonese equivalents.

5.3.2 Semantics-based Mapping

There is consensus that some words are similar to other words in terms of their meaning and syntactic function. If we can successfully assign words to meaningful classes, it may be possible to make more reasonable predictions for unseen or low frequency n-grams by assuming that they are similar to other n-grams that we have seen.

In this study, the embedded English words in the lexicon are clustered into small semantic classes as shown in Figure 5.7. The clustering mainly accords with to the meaning of the words. The part-of-speech (POS) and syntactic function of the words are also considered. WordNet, a lexical database for English, is used as the main reference to determine the meaning of English words [125]. Five major rules used in clustering are shown in Figure 5.7. As a result, about 200 semantic classes are derived.

Major Semantics Classification Rules:

(1) Synonym rule: English words with identical or very similar meanings are grouped together.
Example: frustrated, disappointed

(2) Antonym rule: English words with opposite meaning are assigned to the same semantic class.
Example: upload, download

(3) Coordinate rule: English words sharing the same hypernym may be clustered together.
Example: breakfast, lunch, dinner, buffet, tea, afternoon tea

(4) Function rule: English words with similar syntactic or linguistic function may be considered the same classes (name of peoples, brand names, etc).

(5) Morphology rule: We group together words having the same morphological stem.
Example: understand, understood, understanding

Figure 5.7: Five major rules for semantics-based clustering

5.3.3 Result Analysis & Discussion

Table 5.3 explains the four sets of 3-gram language models. A conventional word tri-gram language model, denoted by **3-gram LM**, is trained as the benchmark. Translation-based mapping and semantics-based mapping are applied to train models **TL_LM** and **SM_LM** respectively. In addition, semantics-based n-gram mapping is performed after translation-based mapping for further improvement. The resulting language model is referred to as **TLSM_LM**. As can be seen in the table, the percentage of code-mixing 3-grams is significantly increased in various kinds of semantic-based language models. This is because the proposed mapping scheme is designed to increase the number of mixed-language n-grams only, while the number of monolingual Cantonese n-grams is kept the same before and after the mapping.

Table 5.3: Four language models developed for Cantonese-English code-mixing LVCSR

Language models	% of Cantonese 3-grams	% of code-mixing 3-grams
3-gram LM	92.77%	7.23%
TL_LM	84.23%	15.77%
SM_LM	84.72%	15.28%
TLSM_LM	72.82%	27.18%

Perplexity is utilized to evaluate the language models. Similar to the evaluation of class-based LM, the test data include the monolingual Cantonese (MC) utterances and Cantonese-English code-mixing (CM) utterances of CUMIX. Table 5.4 gives a summary of the perplexities of four language models described above. The perplexity is measured in terms of character perplexity for Cantonese and word perplexity for English. Tri-gram probabilities on hypothesis pure Cantonese word sequences and mixed-language word sequences are also given.

The character perplexities of four language models on CUMIX monolingual Cantonese utterances are close to each other. Conventional word-based **3-gram LM** attains the best character perplexities of 93.8. Slight increases

Table 5.4: Perplexities of four language models

Language models	ppl - MC text	ppl - CM text	log probability of hypothesis sequences	
			Pure Cantonese	Mixed-language
3-gram LM	93.8	169.8	-6.56	-13.39
TL_LM	96.3	137.4	-6.60	-11.92
SM_LM	96.6	154.2	-6.60	-12.42
TLSM_LM	101.2	132.5	-6.66	-11.25

in perplexity are found in **TL_LM** and **SM_LM**. The perplexity is further increased to 101.2 when **TLSM_LM** is used. The increase in the perplexity is expected since the percentage of monolingual Cantonese n-grams is reduced in various semantics-based LMs as shown in Table 5.3. The lower the percentage of Cantonese n-grams in language models, the higher the character perplexity in Cantonese data. In other words, the probabilities of Cantonese n-grams are deduced in three semantics-based models. This is also consistent with the derived log probabilities of hypothesis Cantonese n-grams reported in Table 5.4.

In the code-mixing case, the results in Table 5.4 support the fact that the n-gram mapping scheme is a promising approach for language modeling of Cantonese-English code-mixing speech. Different degrees of perplexity reduction can be observed in various semantics-based LMs. The perplexity of **TL_LM** is reduced by 23.6% relatively to 137.4, when compared with that of benchmark **3-gram LM**. Further perplexity reduction to 132.5 is attained in **TLSM_LM**, when semantics-based mapping is performed after translation-based mapping in language modeling. Figure 5.8 gives a simple example to represent the advantage of **TL_LM** and **TLSM_LM**. In general, a reduction in perplexity results in improvement in speech recognition performance. **TLSM_LM** is expected to attain the best performance in LVCSR task.

It is believed that the perplexity improvement is mainly due to the increase of code-mixing n-grams via the mapping process. The higher the percentage of mixed-language n-grams reported in Table 5.3, the higher the log probabilities on hypothesis mixed-language n-grams. It is also noticed that the increased

Figure 5.8: A demonstration of advantages of **TL_LM** and **TLSM_LM**

log probability on mixed-language word sequences significantly overwhelm the reduced probability on pure Cantonese word sequences, in all semantics-based language models. This also explains why the performance in terms of perplexity of semantics-based LMs on entire code-mixing utterance improves.

Table 5.5 lists the mixed-language n-gram converge of different language models. It reveals the association of language models and the test data. The higher the coverage, the better matched n-gram data for model training. Our experimental results confirm that the language model perplexity decreases with increased coverage of mixed-language context. The language model trained from the best-matched n-grams gives the best perplexity.

Further discussion is carried out to analyze the effectiveness of different mapping schemes. The proposed mapping approach is designed to increase the counts of mixed-language n-grams. However, not all of the increased mixed-language n-grams will appear in the test data. If the increased n-grams do not capture the linguistic properties of Cantonese-English code-mixing, the increase in mixed-language n-grams does not improve the n-gram converge. If the proposed n-gram mapping can successfully generate unseen n-grams, increased n-gram coverage can be observed in the code-mixing context. The results reported in Table 5.5 support the fact that the proposed n-gram mapping is a likely approach for increasing meaningful code-mixing n-grams. The code-mixing n-gram converge is significantly improved in various semantics-based LMs. In addition, the comparison between the percentage of code-mixing n-grams shown in Table 5.3 and the n-gram word coverage reported in Table 5.5 can indicate the efficiency of increasing reasonable mixed-language n-grams of different mapping

schemes. The comparison suggests that translation-based mapping is more efficient than others. In **TL_LM**, 8.54% increased code-mixing 3-grams make 5.74% coverage improvement. Besides the observed converge improvement of unseen code-mixing n-grams, the increased n-grams may also increase the counts of some seen code-mixing n-grams.

Table 5.5: N-gram word sequence coverage of the code-mixing context

	3-gram LM	TL_LM	SM_LM	TLSM_LM
1-gram	86.80%	95.51%	92.23%	97.47%
2-grams	34.81%	50.00%	44.82%	59.43%
3-grams	8.64%	14.38%	12.04%	18.83%

Another detailed analysis focuses on translation-based mapping schemes. Table 5.6 explains the three sets of language models developed with different subsets of our Cantonese-to-English translation dictionary. Different from model **TL_LM**, which uses all possible translation pairs in the dictionary, only parts of entries in the translation dictionary related to English OOVs or low count English terms are selected in the development of these three sets of LMs. Perplexities of different LMs for code-mixing data are also given in Table 5.6.

Table 5.6: Three sets of translation-based LMs developed with different parts of the Cantonese-to-English dictionary

LMs	Selected English terms	% of selected English terms in the whole dict	ppl
TL_LM_s1	English OOVs	$\sim$ 10%	163.2
TL_LM_s2	word occurrence $<$ 3	$\sim$ 20%	157.0
TL_LM_s3	word occurrence $<$ 10	$\sim$ 40%	152.3

The performance of different translation-based LMs for code-mixing data improves with increasing size of translation dictionary, as we expect. It is clear that there is an obvious perplexity reduction from 169.8 to 163.2, even if translation-based mapping is only considered for English OOVs. In addition, more than 20% of English words appear less than 3 times in the training text.

The perplexity is further reduced to 157.0 where these lexical items are also considered. Besides, it is observed that almost 40% of English words occur less than 10 times in the training text, and these words are usually considered as low count terms for statistical language models. The perplexity is sequentially reduced to 152.3 if all of these low count English terms are tackled.

In summary, the proposed n-gram mapping approach is proven to be effective to increase the mixed-language n-grams and reduce the language model perplexity. The performance of different language models will be further evaluated in the real LVCSR tasks in next chapter.

□ End of chapter.

Chapter 6

Cantonese-English Code-mixing Recognition Performance Evaluation

Summary

The implementation of a large vocabulary continuous speech recognition system for Cantonese-English code-mixing speech is described in this chapter. The three acoustic models and four language models discussed before are evaluated in the LVCSR tasks. The recognition system with cross-lingual AM **CL_B** and semantics-based LM **TLSM_LM** attains the best recognition performance. It achieves the best overall accuracy of 75% for code-mixing speech. The corresponding character accuracy for Cantonese and word accuracy for English are 76.1% and 65.5%, respectively. This system also achieves similar character accuracy of 75.3% for monolingual colloquial Cantonese utterances. The results confirm that our proposed LVCSR system can successfully recognize Cantonese-English code-mixing utterances in addition to monolingual Cantonese speech.

6.1 Large Vocabulary Code-mixing Speech Recognition System

A large vocabulary code-mixing speech recognition system is developed as shown in Figure 6.1. It consists several core components such as cross-lingual acoustic models, bilingual pronunciation dictionary, and statistical language models as described in previous chapters. The input utterance could be either code-mixing speech with one or more English words, or monolingual Cantonese speech. The decoding algorithm is implemented with the HTK Toolkits [126]. It consists of two passes as described below.

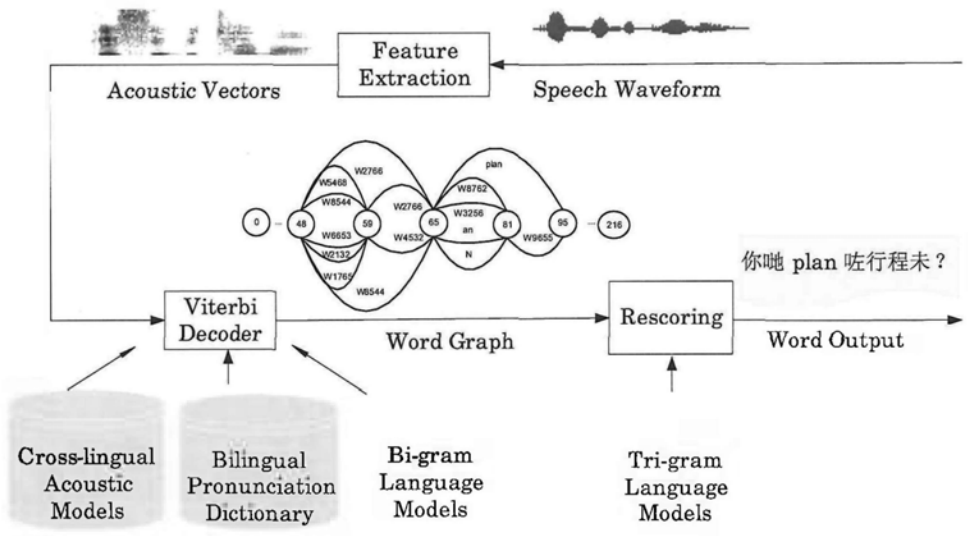


Figure 6.1: Flow diagram of the LVCSR system used in experiments

In the first pass, the cross-lingual acoustic models, bilingual pronunciation dictionary and bi-gram language models are used to generate word lattices. The first-pass decoding is based on the token-passing algorithm. Each token refers to the partial hypothesis starting from the first frame of the utterance. At each time step, a feature vector is taken up and the existing tokens are extended through the HMM states in the recognition network. If there are many competing tokens at a network node, only the best 4 tokens will be kept and the others discarded. In this way, a mixed-language character/word graph is generated as a compact representation of multiple hypotheses. The basic elements of the

lattice are nodes and arcs. Each of this will represent a hypothesized Cantonese lexical item or a hypothesized English word/phrase. It also records the acoustic likelihood, the language likelihood, the start time and the end time of hypothesized words. In the second pass, the character/word lattices are re-scored with tri-gram language models to produce the most probable output word sequence.

6.2 LVCSR Results

The performance of the code-mixing large vocabulary speech recognition system in Figure 6.1 is evaluated using the CM test utterances of CUMIX. The performance is measured by character accuracy for the Cantonese part and word accuracy for the embedded English segments. The LVCSR results are listed in Table 6.1. The accented dictionary generated in Section 4.2 is applied in all recognition experiments. Three cross-lingual acoustic models and four language models discussed before are evaluated in this LVCSR task.

For the acoustic models, it is clear that **CL_B** outperforms models **CL_A** and **CL_IPA** in all LVCSR tasks, when different language models are involved. This is in line with the syllable/word recognition results reported in Table 4.13.

By comparing Table 5.4 and Table 6.1, it is shown that the language models with lower perplexity achieve better recognition performance, except for the class-based model **C-3000 LM**. **TLSM_LM** has the lowest perplexity and achieves the best recognition accuracies in all recognition experiments. However, the recognition performance of **C-3000 LM** is disappointing. Although the perplexity of **C-3000 LM** as shown in Figure 5.4 is significantly lower than that of conventional **3-gram LM**, it can not achieve better recognition accuracy than the baseline **3-gram LM**.

Acoustic model **CL_IPA** and language model **3-gram LM** are implemented as benchmarks in our study. The LVCSR system with **CL_IPA** and **3-gram LM** is regarded as the baseline recognizer. Recognition improvements are expected if better acoustic models or language models are applied. We list the improvement of best acoustic models and language models compared with the

benchmark **CL_IPA** and **3-gram LM** at the end of each column and row of Table 6.1, respectively. The element in the bottom right corner indicates the improvement from the best LVCSR recognizer compared with the baseline. It is found that accuracy improvements attained from AMs vary from 0.5% to 1.3%, when different language models are applied in LVCSR experiments. It can be observed that higher improvement is attained with better language models. The highest accuracy improvement of 1.3% is achieved with the best language models **TLSM_LM**. The same trend is also found in language models. Accuracy improvements attained by LMs vary from 3.2% to 4.0%, the best accuracy improvement of 4.0% being achieved when the best acoustic model **CL_B** is employed. It is also noticed that the improvement attended by language models are always more obvious than that of acoustic models. Figure 6.2 lists some examples of the recognition outputs. It is clear that some English words and Cantonese characters can be corrected by better acoustic models. Further improvement can be achieved with better language models.

Table 6.1: Overall accuracies of Cantonese-English Code-mixing LVCSR

	3-gram LM	c-3000 LM	TL_LM	TLSM_LM	
CL_IPA	70.5%	70.0%	73.3%	73.7%	(+3.2%)
CL_A	70.5%	70.5%	73.6%	74.3%	(+3.8%)
CL_B	71.0%	71.0%	74.4%	75.0%	(+4.0%)
	(+0.5%)	(+0.5%)	(+1.1%)	(+1.3%)	(+1.5%)

In addition to the overall performance for code-mixing speech, the character accuracies for Cantonese parts and word accuracies for embedded English segments are given in Table 6.2 and Table 6.3, respectively. Results without language models of syllable/word recognition experiments given in Table 4.13 are also reported here for further discussion. The results show that the overall recognition accuracy of code-mixing speech is dominated by the recognition accuracy of Cantonese terms.

Sent. 1	CL_IPA + 3-gram LM: 一 定 有 得 睇 大 漠 CL_B + 3-gram LM: 一 定 有 得 睇 demo CL_B + TLSM_LM: 一 定 有 得 睇 demo **Reference: 一 陣 有 得 睇 demo**
Sent. 2	CL_IPA + 3-gram LM: 雖 然 明 知 做 阿 四 但 我 都 會 成 怒 CL_B + 3-gram LM: 雖 然 明 知 做 阿 四 但 我 都 唔 會 say no CL_B + TLSM_LM: 雖 然 明 知 做 阿 四 但 我 都 唔 會 say no **Reference: 雖 然 明 知 做 阿 四 但 我 都 唔 會 say no**
Sent. 3	CL_IPA + 3-gram LM: 每 個 人 都 會 遊 客 time CL_B + 3-gram LM: 每 個 人 都 會 遊 客 time CL_B + TLSM_LM: 每 個 人 都 會 有 hard time **Reference: 每 個 人 都 會 有 hard time**

Figure 6.2: Examples of recognition results

In general, statistical n-gram language models usually improve recognition performance by incorporating linguistic knowledge. It is clear from Table 6.2 that the Cantonese character accuracies are evidently higher than syllable accuracies without integrating language models. This confirms that the developed language models are effective in recognizing Cantonese, in which more than 10% of wrongly recognized Cantonese syllables can be corrected with language models to provide correct character outputs.

Table 6.2: LVCSR accuracies for code-mixing Cantonese characters

	without LM	3-gram LM	c-3000 LM	TL_LM	TLSM_LM
CL_IPA	57.6%	73.1%	72.5%	74.7%	74.9%
CL_A	62.0%	73.1%	73.1%	74.9%	75.5%
CL_B	62.1%	73.5%	73.6%	75.7%	76.1%

Table 6.3: LVCSR accuracies for embedded English words

	without LM	3-gram LM	c-3000 LM	TL_LM	TLSM_LM
CL_IPA	54.1%	48.6%	49.2%	61.7%	63.9%
CL_A	61.8%	48.6%	48.4%	62.5%	64.4%
CL_B	62.6%	50.5%	49.2%	63.4%	65.5%

It is also noticed that the character accuracies of four language models for code-mixing Cantonese are comparable to each other. Around a 2% improvement can be observed in **TL_LM** compared with baseline **3-gram LM** and class-based **c-3000 LM** in terms of character accuracy. **TLSM_LM** slightly outperforms **TL_LM** and attains the best recognition accuracy. This is because the proposed semantics-based language models aim at providing better estimation of mixed-language n-grams for code-mixing speech recognition, and therefore the improvement is only shown for Cantonese characters at language boundaries, which account for a small part of entire Cantonese speech in code-mixing utterances.

In contrast to the observations made for Cantonese characters, English word accuracies are not at the same level when different language models are used. Firstly, **3-gram LM** is found to be inefficient in recognizing embedded English words in code-mixing speech recognition. Compared with purely acoustic decoding results without a language model, the word accuracy of English in code-mixing speech, however, is decreased obviously. This is because the training data for language modeling contain much more Cantonese characters than English words, and therefore the **3-gram LM** is overtrained for Cantonese. Many English words which can be correctly decoded by acoustic models will be wrongly recognized as Cantonese characters after involving langauge models. Class-based model **C-3000 LM** shows a similar performance to that of **3-gram LM**. It means that the automatic classification of Cantonese terms and English words is unprofitable for code-mixing ASR.

TL_LM improves greatly in recognizing English words due to better matched mixed-language n-gram data available for language modeling. Further improvement is achieved by combining the semantics-based mapping. **TLSM_LM** attains the best recognition accuracy for the embedded English words. It is noted that many wrongly recognized English words can be corrected by semantics-based language models. This proved that our proposed semantics-based n-gram mapping approach is promising for language modeling of Cantonese-English code-mixing speech.

On the whole, an LVCSR system with acoustic model **CL_B** and language model **TLSM_LM** attains the best overall accuracy of 75%, as expected. Compared with the baseline recognizer, the recognition accuracy is improved by 4.5%. It also attains the best performance in recognizing Cantonese terms as well as English words. The best character accuracy and word accuracy are 76.1% and 65.5% respectively. Compared with the benchmark results, 3% and 16.9% accuracy improvement are achieved for Cantonese characters and English words, respectively. The experimental results confirm that the proposed LVCSR system can recognize Cantonese-English code-mixing utterances satisfactorily.

6.3 Analysis & Discussion

6.3.1 Lattice Error Rate

The recognition performance in terms of the lattice error rate (LER) is also analyzed in this study. The LER is the oracle word error rate of the most correct path through the lattice. It is computed by aligning the referenced word sequence with the word graph to find the path with the least number of word errors. If the correct path exists in the word graph, the lattice error rate for that utterance is 0%.

Table 6.4 lists the details of the LERs by using different acoustic and language models during the first-pass search. Similar as 1-best results reported in Table 6.1, the improvements in terms of LER of best acoustic models and language models compared with the benchmark models are listed at the end of each column and row, respectively. By comparing Table 6.1 and Table 6.4, it is shown that the developed data-driven acoustic models and semantics-based language models not only result in better 1-best recognition outputs, but also produce lattices with lower oracle word error rates. The error reductions in LERs are less significant than that of 1-best WREs. Using **CL_B + TLSM_LM** during decoding produces the best lattice with a 1.2% (absolute) lower oracle WER compared to the baseline lattice produced by benchmark **CL_IPA + 3-gram LM**.

Table 6.4: Lattice error rates by using different acoustic and language models during decoding

	3-gram LM	c-3000 LM	TL_LM	TLSM_LM	
CL_IPA	6.1%	6.5%	5.4%	5.2%	(-0.9%)
CL_A	6.0%	6.6%	5.3%	5.2%	(-0.8%)
CL_B	5.7%	6.3%	5.0%	4.9%	(-0.8%)
	(-0.4%)	(-0.2%)	(-0.4%)	(-0.3%)	(-1.2%)

On the other hand, LER indicates the lower bound of the word error rate that is attainable by 2-pass word-graph re-scoring. It is clear that there is a significant gap between the lattice oracle WERs and the 1-best WERs. This indicates that a lower WER is obtainable if a better language model can be applied in the second pass re-scoring. In addition, lattice re-scoring can integrate high-level sources (e.g. duration and F0 information) which may not be easily incorporated in the first decoding pass to further improve the recognition accuracy. Therefore, we suggest future works to explore offline second pass lattice re-scoring.

6.3.2 Error Composition of Embedded English

Further analysis of recognition performance is done by analyzing the error composition of embedded English words, as shown in Figure 6.3. Besides the overall word error rate (WER), it also shows the details of insertion (INS), deletion (DEL) and substitution (SUB) rate of wrongly recognized English words when different language models are applied. **CL_B** is used as the acoustic model in this study.

Firstly, it is found that the insertion error is very low. This means that very few Cantonese characters are misrecognized as English words. On the other hand, a small amount of English words are recognized as other English terms. This type of substitution error is mainly caused by incorrect language boundaries; thus the hypothesis English word and the reference English word have no or just very little overlap in time duration. For example, the word *arrive*

is mistakenly recognized as *wife*, and *resolution* becomes *solution*. Finally, it is clear that recognition errors on English words are largely due to the deletion errors. In such cases, English words are recognized as Cantonese characters.

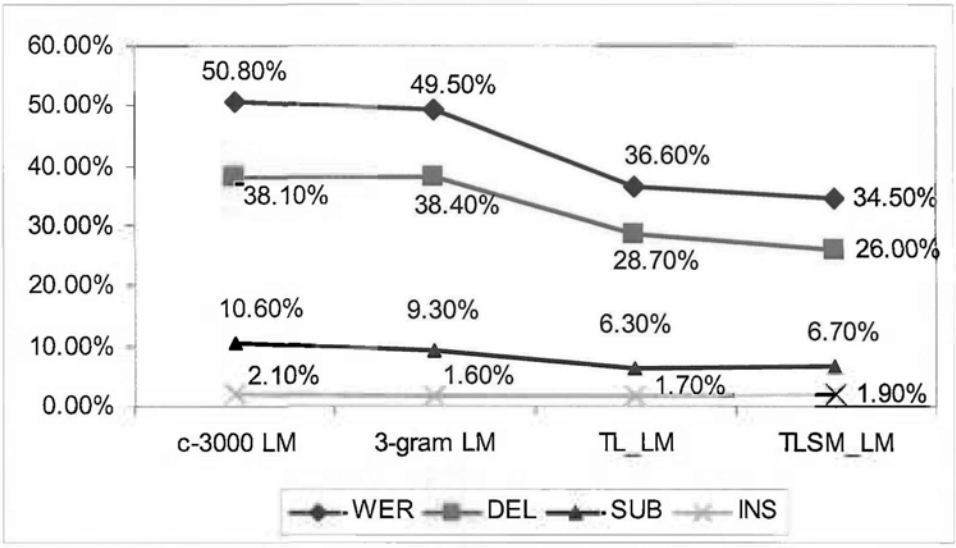


Figure 6.3: Error distribution for embedded English words

It is also noted that the gap between word error rate and deletion rate is about the same for different language models. It seems that the improvement of recognition performance for English words is mainly due to the reduction of deletion. Nevertheless, the lowest word deletion rate of 26.0% is still on the high side. It indicates that the improved code-mixing language model **TLSM_LM** still has bias for Cantonese. Further improvement for English is expected if there is more mixed-language n-grams available for model training.

6.3.3 Scale Factors of Language Models

As we introduced at the beginning of this chapter, tri-gram language is used in the lattice re-scoring in the second pass. Different scale factors can be selected to integrate language models. Figure 6.4 shows the recognition performance on code-mixing speech as a function of various scale factors s of different language models. The best acoustic model **CL_B** is used in this study. Graphs (a), (b), and (c) show the overall accuracy, Cantonese character accuracy and En-

glish word accuracy, respectively. It is clear that performance differences among different language models enlarge with the increased scale factor s. The appropriate value of s for code-mixing speech recognition is in the range of 12-15, which is consistent with the common setting in monolingual LVCSR systems.

6.3.4 System Performance for Monolingual Cantonese

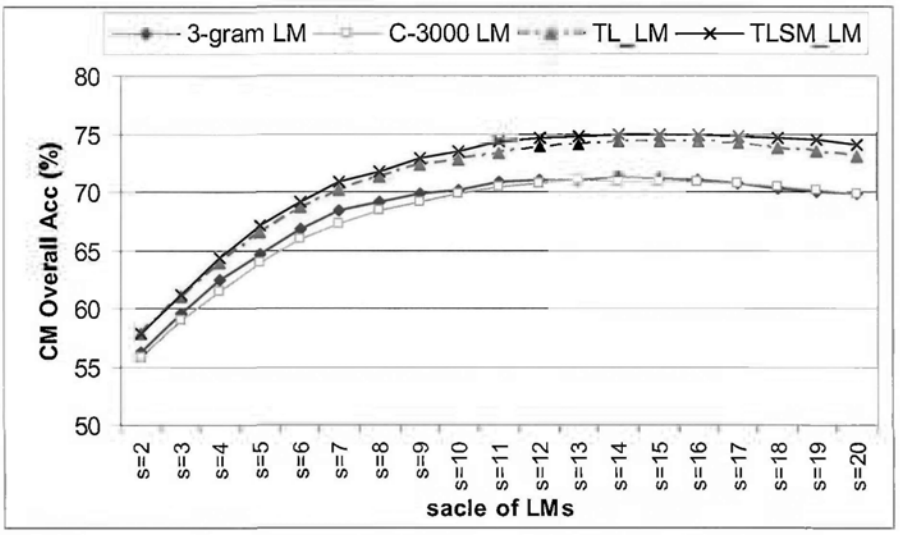
According to previous discussions, a better LVCSR system for Cantonese-English code-mixing speech can be built with acoustic model **CL_B** and language model **TLSM_LM**. Monolingual colloquial Cantonese utterances (MC) of CUMIX are also used to evaluate this LVCSR system.

Two recognition experiments are carried out with different lexicons in this evaluation. In the first experiment, the same bilingual pronunciation dictionary used in code-mixing speech recognition is employed. In such case, no constraint is applied and the recognition output can be either code-mixing or monolingual Cantonese text sequences. In the second experiment, a monolingual Cantonese pronunciation dictionary is applied. This assumes that the input speech must be monolingual Cantonese utterances and therefore there should be no recognition error caused by the confusion between similar Cantonese and English lexical items. We list the recognition results in terms of Cantonese character accuracy in Table 6.5.

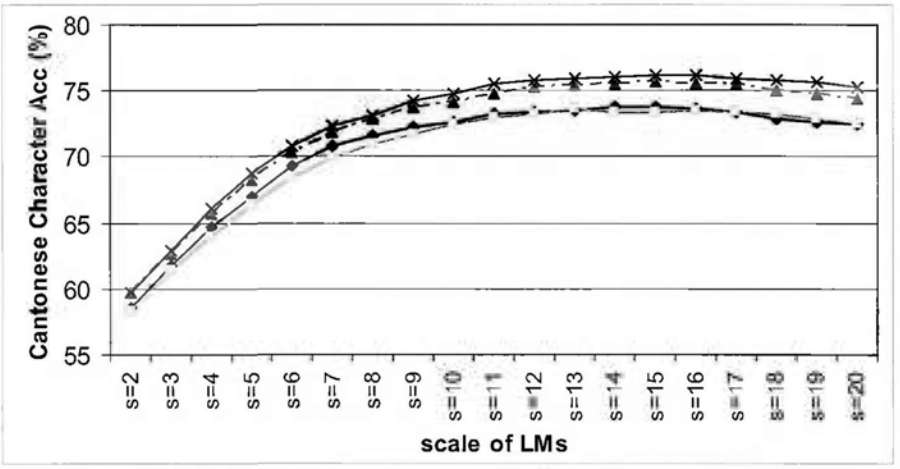
Table 6.5: Recognition accuracy on MC test utterances

	with bilingual dict.	with Cantonese dict.
Cantonese character Acc.	75.3%	75.4%

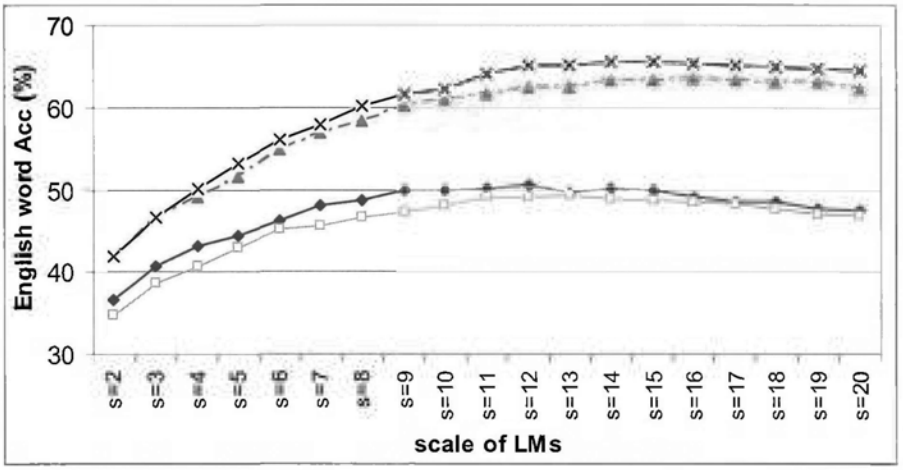
The table shows that our proposed code-mixing LVCSR system attains the character accuracy of 75.3% for monolingual Cantonese speech. If the monolingual property is known before decoding, the recognition accuracy can be increased from 75.3% to 75.4%. This means that only 0.1% of Cantonese characters are mistakenly recognized as English words in the first experiment. It is also noticed that the proposed recognition system attains a similar performance for MC and CM utterances. The results suggest that we can build a univer-



(a) Overall accuracy for code-mixing speech



(b) Character accuracy for code-mixing Cantonese



(c) Word accuracy for embedded English

Figure 6.4: LVCSR performance on code-mixing speech against scale factors s of various language models used in lattice re-scoring

sal recognition system which is able to handle code-mixing and monolingual Cantonese speech together.

6.3.5 Discussion of Code-mixing Recognition Errors

In this section, we attempt to analyze the code-mixing recognition results at sentence level. It is found that the proposed LVCSR system can properly recognize 21.8% of code-mixing utterances. In other words, nearly 78% of code-mixing utterances are mistakenly recognized with either incorrect English words or Cantonese characters. We are interested in further examining utterances with wrongly recognized English words or Cantonese characters in language boundaries. The recognition outputs of these utterances are observed and examples of typical errors are given in Figure 6.5.

Error (a) usually appears along with imperfect acoustic models. As can be seen in example sentences 1 and 2, the embedded English words are recognized as other English words with similar pronunciation. Nevertheless, the recognition outputs are still meaningful code-mixing sentences. This type of errors can be corrected with better acoustic models.

Errors (b) and (c) represent errors on English words and Cantonese characters at language boundaries because of insufficient code-mixing language models, respectively. This type of error can be tackled if language models can be trained with more code-mixing text data.

The occurrence of error (d) has nothing to do with code-mixing. It will be appeared in monolingual speech recognition of Cantonese accented English as well. This error may not be recovered by improved acoustic models and tri-gram language models trained with more code-mixing data. For example, the pronunciation of *offer* and *author* are fairly similar in Cantonese-accented English. On the other hand, both of the references and recognition outputs of sentence 9 are in line with tri-gram language models. However, it is found that the recognition results are unreasonable at sentence level. As a matter of fact, this type of errors may be handled by high-order n-gram or skip/distance language models. Integration of linguistic constraint may be also helpful.

Error (a)	Sent 1	Rec: 我 重 有 其 他 chance 呢 Ref: 我 重 有 其 他 choice 咩
	Sent 2	Rec: steven 同 佢 分 析 咗 好 多 但 最 後 都 係 唔 明 白 Ref: even 同 佢 分 析 咗 好 多 但 最 後 都 係 唔 明 白
Error (b)	Sent 3	Rec: 我 畫 完 幅 所 謂 嘅 咩 就 收 工 Ref: 我 畫 完 幅 所 謂 嘅 map 就 收 工
	Sent 4	Rec: 扶 助 米 埔 過 第 二 個 囉 Ref: full 咗 咪 報 過 第 二 班 囉
	Sent 5	Rec: 跟 住 落 咗 中 環 比 錢 confirm 機 票 同 酒 店 Ref: 跟 住 落 咗 中 環 比 錢 九 龍 機 票 同 酒 店
Error (c)	Sent 6	Rec: in general 來 港 大 學 生 係 應 該 醒 目 啲 Ref: in general 黎 講 大 學 生 係 應 該 醒 目 啲
	Sent 7	Rec: 會 唔 會 令 個 professor 劉 德 華 印 象 架 Ref: 會 唔 會 令 個 professor 留 低 壞 印 象 架
	Sent 8	Rec: 我 地 主 要 受 local 學 生 Ref: 我 地 主 要 收 local 學 生
Error (d)	Sent 9	Rec: 個 offer 好 好 人 重 幫 我 簽 咗 個 明 天 Ref: 個 author 好 好 人 重 幫 我 簽 咗 個 名 添
Error (e)	Sent 10	Rec: 比 人 一 個 好 旺 嘅 感 覺 Ref: 比 人 一 個 好 warm 嘅 感 覺
	Sent 11	Rec: 每 個 地 區 都 有 自 己 獨 特 嘅 球 場 Ref: 每 個 地 區 都 有 自 己 獨 特 嘅 culture
Error (f)	Sent 12	Rec: 張 banner 又 冇 限 大 細 Ref: 張 banner 有 冇 限 大 細
	Sent 13	Rec: 如 果 有 另 一 個 angle 睇 呢 Ref: 如 果 由 另 一 個 angle 睇 呢
Error (g)	Sent 14	Rec: 呢 件 衫 係 我 自 己 decide 喇 Ref: 呢 件 衫 係 我 自 己 design 架
	Sent 15	Rec: 你 好 快 咁 高 呼 一 次 喇 Ref: 你 好 快 咁 go through 一 次 喇
Error (h)	Sent 16	Rec: guest house 會 唔 會 好 似 酒 店 咁 貴 價 Ref: guest house 會 唔 會 好 似 酒 店 咁 貴 架
	Sent 17	Rec: 呢 啲 嘢 understood 架 喇 Ref: 依 啲 嘢 understood 架 喇

Figure 6.5: Examples of different code-mixing recognition errors

Similar to error (d), conventional acoustic models and statistical language models may not help to reduce errors (e) and (f). However, as indicated in example sentences 10-13, it is believed that suprasegmental information should be useful to amend mistakenly recognized English words and Cantonese characters. Integration of F0 and duration information in recognition system is highly desirable in future works.

It would be very difficult to deal with error (g). It is clear that the semantics of recognition outputs of example sentences 14 and 15 are reasonable. Moreover, wrongly recognized words and the references capture very similar phonetic and suprasegmental information (e.g. *decide* and *design*, 高呼 and *go through*).

Error (h) can be neglected. As shown in example sentences 16 and 17, the mistakenly recognized Cantonese characters would not affect the meaning of the sentences.

□ End of chapter.

Chapter 7

Conclusions and Suggestions for Future Work

7.1 Conclusions

Cantonese-English code-mixing is a common speaking phenomenon of people residing in Hong Kong. For most local residents, Cantonese is their primary language, also known as the matrix language, while English is the secondary language, also known as the embedded language.

In this thesis, we focus on the development of a high-performance Cantonese-English LVCSR system. We begin with some preliminary studies to investigate the linguistic properties of Cantonese-English code-mixing and to analyze the effect of language mixing for ASR performance. After that we have investigated how to improve the performance of Cantonese-English code-mixing LVCSR. The study covers all components of the ASR system, including acoustic models, language models and pronunciation dictionary. The major conclusions are given below.

Linguistic Properties of Code-mixing

In order to better understand this highly dynamic language phenomenon, we first attempt to investigate the linguistic properties of code-mixing. The study is based on a large number of real code-mixing text corpora collected from

the internet and other sources. Our study reveals clearly that code-mixing is not a simple insertion of one language into another. It comes with a lot of phonological, lexical and grammatical variations with respect to monolingual speech spoken by native speakers.

Effects of Language Mixing for Code-mixing ASR

Although automatic speech recognition of either Cantonese or English alone has achieved a great degree of success, significant degradation in recognition accuracy is noted for Cantonese-English code-mixing ASR. By examining the recognition results of Cantonese-English code-mixing speech, we notice that the recognition accuracy of the embedded language plays a significant role in relation to the overall performance. In particular, significant performance degradation is found in the matrix language if the embedded words cannot be recognized correctly. We also study the error propagation effect of the embedded English. The results show that the error found in a particular embedded English word may propagate to two neighbouring Cantonese syllables. This indicates that the recognition performance on embedded language is very important, and it is believed that the enhancement in the recognition of embedded language will bring improvement to the matrix language as well.

Pronunciation variations

Pronunciation variations in Cantonese-English code-mixing speech are investigated. The analysis is performed by comparing the confusion matrices obtained from speech recognition experiments with different types of monolingual and code-mixing speech data. It is found that English words spoken by Cantonese speakers, whether or not in a code-mixing utterance, carry strong Cantonese accents. There is also no significant difference between Cantonese utterances with and without code-mixing. Based on the analysis of the confusion matrices, a number of context-independent and context-dependent phonetic variation patterns are established and we modify the pronunciation dictionary according to the observed variations. The experimental results show that noticeable recog-

nition improvement is attained with the modified pronunciation dictionary. On the other hand, in order to select effective training materials, various sets of acoustic models are trained with different speech data. The recognition results confirm that by using accented English words extracted from CUMIX in the acoustic modeling of Cantonese-English code-mixing speech will lead to a satisfactory outcome. Both colloquial Cantonese speech in CUMIX as well as read-style Cantonese from CUSENT could however be applied in the acoustic modeling for code-mixing ASR.

Acoustic Modeling

We have shown that cross-lingual acoustic models are more appropriate than language-dependent models. To design a cross-lingual phoneme set, we need to measure the similarity between phonemes of the two languages. Various cross-lingual inventories are derived based on different combination schemes and similarity measurements. It is shown that the proposed data-driven based approach outperforms the IPA-based approach using merely phonetic knowledge. It is also found that initials and finals are more appropriate as the basic Cantonese units than phonemes in code-mixing speech recognition applications. The IF-based cross-lingual AMs show consistent recognition improvement compared with phoneme-based models. The proposed cross-lingual models attains the best overall syllable/word accuracy of 62.1%. The overall recognition performance is improved by nearly 5% as compared with the IPA-based models. In particular, the accuracy of recognizing embedded English words increases from 54.1% to 62.6%.

Language Modeling

To deal with the problem of inadequate code-mixing training data, different language modeling techniques are investigated. Class-based language models trained with automatic clustering classes show significant reduction in perplexity. However, the recognition performance of developed class-based models is disappointing as the models cannot improve the recognition accuracy in real

LVCSR tasks. The proposed semantics-based n-gram mapping approach is proved to be very effective for language modeling of code-mixing ASR. The percentage of code-mixing 3-grams increases from 7.2% to 27.2% via translation-based and semantics-based mapping. It is also noticed that 10.7% of English OOVs, 24.6% and 10.2% of unseen code-mixing 2-grams and 3-grams can be observed after mapping. The advantage of the n-gram mapping method can be further confirmed in LVCSR tasks. With proposed semantics-based language models, the recognition accuracy on embedded English words of code-mixing speech can be increased by more than 15%, compared with the conventional statistical tri-gram LM. A 2% improvement on Cantonese characters can be observed as well, and the improvement is mainly presented on boundary Cantonese characters nearby English words. The success of semantics-based language models suggest that well-founded mapping is an effective approach to deal with data sparseness. It is worth to further investigate how English and Cantonese are intertwined in code-mixing and apply them in mapping.

Speaker adaptation

Speaker adaptation techniques can be used to improve speech recognition performance if a small amount of adaptation data from the target speaker is available. However, it is not easy to do so when two or more languages are involved, especially when the two languages are phonetically distant. Cross-lingual speaker adaptation via model mapping is investigated in this study. It focuses on the use of acoustic information from an existing source language (Cantonese) to implement speaker adaptation for a new target language (English). SI model mapping between Cantonese and English is established for different acoustic units. Phones, states and Gaussian mixture components are used as the mapping units respectively. With model mapping, cross-lingual speaker adaptation can be performed. The performance of the proposed cross-lingual speaker adaptation system is determined by the effectiveness of both model mapping and speaker adaptation. Experimental results show that model mapping effectiveness increases with refinement of mapping units, and the better the model mapping,

the more effective the speaker adaptation. Mapping between Gaussian mixture components is proved to be successful for various speech recognition tasks. A relative error reduction of 10.12% for English words is achieved by using a small amount of (4 minutes) Cantonese adaptation data, compared with the SI English recognizer. The experimental results show that our approach for cross-lingual speaker adaptation is promising. In addition, if a speaker-dependent Cantonese recognizer exists for a particular speaker, the corresponding speaker-dependent English recognizer can be implemented via the proposed GauMix Mapping in a fast and low cost way.

LVCSR System

A complete large vocabulary continuous speech recognition system for Cantonese-English code-mixing speech has been developed and implemented. The best overall recognition accuracy for code-mixing test utterances of CUMIX is about 75%. In particular, it achieves an accuracy of 76.1% and 65.5% for Cantonese characters and English words respectively. The proposed code-mixing LVCSR system yields significant improvement as compared with the baseline system. We also find that this performance level is noticeably higher than the previously reported methods, which attained an overall accuracy of 55.3% for the same sets of Cantonese-English code-mixing test utterances. In addition, the proposed code-mixing LVCSR system can successfully recognize monolingual Cantonese speech as well. The recognition accuracy for Cantonese characters in monolingual utterances is 75.3%, which is comparable to the results of previous studies (character accuracy of 75.8%) on colloquial Cantonese LVCSR (for the same database).

7.2 Summary of Contributions

The major contributions of this thesis are summarized hereunder:

- A text corpus of 65,000 real code-mixing sentences is collected from contents obtained via the internet. Domains of code-mixing sentences and

POS of embedded English words are manually labelled. The characteristics of Cantonese-English code-mixing are investigated from a corpus linguistic point of view.

- The effects of language mixing for automatic recognition of Cantonese-English code-mixing utterances are analyzed in a systematic way.

- A data-driven computational approach is adopted to reveal significant pronunciation variation in Cantonese-English code-mixing speech, in addition to those variations that have been well understood in monolingual speech recognition. The findings are successfully implemented to construct a more appropriate bilingual pronunciation dictionary and select effective training materials for code-mixing ASR.

- Various similarity measurements are applied to investigate the acoustic and phonetic similarity between different phonemes of Cantonese and English. Based on that, different sets of cross-lingual phoneme inventories are designed and evaluated in speech recognition experiments.

- A text database with more than 9 million characters are compiled for language modeling of code-mixing ASR. Class-based language models with automatic clustering classes have been proven inefficient for code-mixing speech recognition. A semantics-based n-gram mapping approach is proposed to increase the counts of code-mixing n-gram at language boundaries. The language model perplexity and recognition performance is significantly improved by the proposed semantics-based language models.

- Speaker independent model mapping between Cantonese and English is established at different levels of acoustic units, viz phones, states and Gaussian mixture components. A novel approach for cross-lingual speaker adaptation via Gaussian component mapping is proposed and is proven effective in speech recognition tasks.

- A large vocabulary continuous code-mixing speech recognition system is built. This system can successfully recognize Cantonese-English code-

mixing and monolingual Cantonese speech spoken in daily conversations. The overall recognition accuracy for code-mixing speech increases from (previously reported) 55.3% to 75.0%.

7.3 Suggestions for Future Work

Although our proposed code-mixing LVCSR system significantly outperforms previously reported code-mixing recognizers and the baseline system trained with conventional methods, the accuracy on embedded English words is still much lower than that of Cantonese characters. We should continue to improve the performance of code-mixing LVCSR, particular for English words.

In this study, less than 3 hours of English speech data are available for code-mixing acoustic modeling. To improve recognition performance, more Cantonese-accented English speech should be collected and developed in future research. In this project, we have collected over 15 hours of real spontaneous speech, which has not been fully used yet. This set of spontaneous speech should be exploited in future work.

We find that more than 20% of English words are mistakenly recognized as Cantonese characters in our code-mixing recognition experiments. However, less than 1% of Cantonese characters are recognized as English words. This means that our proposed semantics-based language models are still biased towards Cantonese. Therefore, we need to collect more code-mixing text data for language modeling. Moreover, advanced language model techniques such as skip or long-distance language modeling may be considered for further research.

From the analysis of code-mixing LVCSR results, we find that high-level knowledge such as suprasegmental information may be helpful to correct the wrongly recognized English words and Cantonese characters. It is worth to integrate such information into the code-mixing LVCSR and see whether even better performance could be achieved.

□ End of chapter.

Bibliography

[1] H. M. Meng, S. Lee, and C. Wai, "Cu forex: A bilingual spoken dialog system for foreign exchange enquiries," in *Proc. ICASSP*, 2000, pp. 1229–1232.

[2] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," in *Proc. IEEE*, vol. 88, no. 8, 2000, pp. 1338–1353.

[3] Peabody, Mass, and Ghent, "Scansoft powers premier travel inn uk speech-based booking system," 2005.

[4] V. C. Matula, "Improved lsi-based natural language call routing using speech recognition confidence scores," Technical Report of Avaya Labs Research (ALR-2004-023), Tech. Rep., 2004.

[5] "Philips dictation systems, http://www.dictation.philips.com."

[6] P. Woodland, D. P. M. Gales, and S. Young, "Broadcast news transcription using htk," in *Proc. ICASSP*, vol. 2, 1997, pp. 719–722.

[7] W. W. Verbmobil, *Foundations of speech-to-speech translation.* Springer, 2000.

[8] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[9] H. Sakoe and S. Chiba, "Dynamic programming algoritm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[10] F. Itakura, "Minimum prediction residual applied to speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[11] J. Ferguson, *Hidden Markov Models for Speech.* IDA, Priceton, 1980.

[12] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. IEEE*, vol. 77, 1989, pp. 257–286.

[13] R. P. Lippman, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, vol. 4, no. 2, pp. 4–22, 1987.

[14] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, pp. 393–404, 1989.

[15] K. F. Lee, H. W. Hon, and D. R. Reddy, "An overview of the sphinx speech recognition system," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 38, pp. 600–610, 1990.

[16] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language*, vol. 4, pp. 127–165, 1990.

[17] Y. L. Chow, M. O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, S. Makhoul, S. Roucos, and R. M. Schwartz, "Bbylos: The bbn continuous speech recognition system," in *Proc. ICASSP*, 1987, pp. 89–92.

[18] F. Jelinek, *Statistical Methods for Speech Recognition.* MIT Press, 1997.

[19] D. B. Paul, "The lincoln robust continuous speech recognizer," in *Proc. ICASSP*, 1989, pp. 449–452.

[20] V. Zue, J. Glass, M. Phillips, and S. SeneR, "The mit summit speech recognition system: A progress report," in *Proc. DARPA Speech and Natural Language Workshop*, 1989, pp. 179–189.

[21] E. Wong and S. Sridharan, "Three approaches to multilingual phone recognition," in *Proc. ICASSP*, vol. 1, 2003, pp. 44–47.

[22] Z. Wang, U. Topkara, T. Schultz, and A. Waibel, "Towards universal speech recognition," in *Proc. ICMI*, 2002, pp. 247–252.

[23] U. Uebler, "Multilingual speech recognition in seven languages," *Speech Communication*, vol. 35, pp. 53–69, 2001.

[24] D.-C. Lyu, T.-P. Tan, E. S. Chug, and H. Li, "Seame: a mandarin-english code-switching speech corpus in south-east asia," in *Proc. Interspeech*, 2010, pp. 1986–1989.

[25] J. Y. C. Chan, T. Lee, and P. C. Ching, "Development of a cantonese-english code-mixing speech corpus," in *Proc. EUROSPEECH*, 2005, pp. 1533–1536.

[26] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and C.-Y. Lin, "Automatic segmentation and identification of mixed-language speech using delta-bic and lsa-based gmms," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 1, pp. 266–276, 2006.

[27] Q. Zhang, J. Pan, and Y. Yan, "Mandarin-english bilingual speech recognition for real world music retrieval," in *Proc. ICASSP*, 2008, pp. 4253–4256.

[28] Y. C. Chan, P. C. Ching, T. Lee, and H. CAO, "Automatic speech recognition of cantonese-english code-mixing utterances," in *Proc. INTERSPEECH*, 2006, pp. 113–116.

[29] D.-C. Lyu and R.-Y. Lyu, "Language identification on code-switching utterances using multiple cues," in *Proc. INTERSPEECH*, 2008, pp. 711–714.

[30] D.-C. Lyu, R.-Y. Lyu, Y. chin Chiang, and C.-N. Hsu, "Speech recognition on code-switching among the chinese dialects," in *Proc. ICASSP*, 2006, pp. 1105–1108.

[31] Y. C. Chan, "Automatic speech recognition of cantonese-english code-mixing utterances," Master's thesis, The Chinese University of Hong Kong, 2005.

[32] Y. C. Chan, P. C. Ching, T. Lee, and H. Meng, "Detection of language boundary in code-switching utterances by bi-phone probabilities," in *Proc. ISCSLP*, 2004, pp. 293–296.

[33] T. Lee, W. Lau, Y. W. Wong, and P. C. Ching, "Using tone information in cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1, pp. 83–102, 2002.

[34] Y. Qian, "Use of tone information in cantonese lvcsr based on generalized character posterior probability decoding," Ph.D. dissertation, The Chinese University of Hong Kong, 2005.

[35] W. N. Choi, "An efficient decoding method for continuous speech recgonition based on a tree-structured lexicon," Master's thesis, The Chinese University of Hong Kong, 2001.

[36] Y. T. Yeug, "Language modeling for speech recognition of spoken cantonese," Master's thesis, The Chinese University of Hong Kong, 2009.

[37] J. L. Gauvain and L. Lamel, "Large-vocabulary continuous speech recognition: advances and applications," in *Proc. IEEE*, vol. 88, no. 8, August 2000, pp. 1181–1200.

[38] C. H. Lee, F. K. Soong, and K. K. Paliwal, *Automatic Speech and Speaker Recognition: Advanced Topics.* Kluwer Academic Publishers, 1996.

[39] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?" in *Proc. IEEE*, vol. 88, no. 8, August 2000, pp. 1270–1278.

[40] J. W. Picoone, "Signal modeling techniques in speech recognition," in *Proc. IEEE*, vol. 81, no. 9, September 1993, pp. 1215–1247.

[41] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[42] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Signal Processing*, vol. 28, pp. 357–366, 1980.

[43] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. Human Language Technology Workshop*, vol. 37, 1994, pp. 307–312.

[44] M. Y. Hwang and X. Huang, "Shared distribution hidden markov models for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 414–420, 1993.

[45] S. J. Young, P. C. Woodland, and W. J. Byrne, "State clustering in hmm-based continuous speech recognition," *Computer Speech and Language*, vol. 8, no. 4, pp. 369–384, 1994.

[46] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modeling," *Computer Speech and Language*, vol. 8, no. 1, pp. 1–38, 1994.

[47] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 35, no. 3, pp. 400–401, 1987.

[48] S. J. Young, N. H. Russell, and J. H. S. Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," *CMU Tech Report*, 1989.

[49] S. Ortmanns and H. Ney, "Experimental analysis of the search space for 20000 word speech recognition," in *Proc. EUROSPEECH*, 1995, pp. 901–904.

[50] R. Haeb-Umbach and H. Ney, "Improvements in beam search for 10000-word continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 353–356, 1994.

[51] T. Nagarajan and H. A. Murthy, "Language identification using parallel syllable-like unit recognition," in *Proc. ICASSP*, vol. 1, 2004, pp. 401–404.

[52] K. V. S. Jayram, V. Ramasubramanian, and T. V. Sreenivas, "Language identification using parallel sub-word recognition," in *Proc. ICASSP*, vol. 1, 2003, pp. 32–35.

[53] T. Nagarajan and H. A. Murthy, "Language identification using parallel syllable-like unit recognition," in *Proc. ICASSP*, vol. 1, 2004, pp. 401–404.

[54] J. Köhler, "Multilingual phone models for vocabulary-independent speech recognition tasks," *Speech Communication*, vol. 35, pp. 21–30, 2001.

[55] B. Imperl and B. Horvat, "The clustering algorithm for the definition of multilingual set of context dependent speech models," in *Proc. EUROSPEECH*, 1999, pp. 887–890.

[56] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998, pp. 259–262.

[57] ——, "Adaptation of pronunciation dictionaries for recognition of unseen languages," in *Proc. SPIIRAS International Workshop on Speech and Computer*, 1998, pp. 207–210.

[58] J. Köhler, "Language adaptation of multilingual phone models for vocabulary-independent speech recognition tasks," in *Proc. ICASSP*, 1998, pp. 417–420.

[59] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICSLP*, 1998, pp. 1819–1822.

[60] P. Auer, *Code-Switching in Conversation: Language, Interaction and Identity.* Routledge, London, 1998.

[61] C. M. Chan, "Two types of code-switching in taiwan," in *Proc. Sociolinguistics Symposium*, 2004.

[62] H. Halmari, *Government and Code-switching: Explaining American Finnish.* Benjamins, Amsterdam, 1997.

[63] Y. C. Chan, H. Cao, P. C. Ching, and T. Lee, "Automatic recognition of cantonese-english code-mixing speech," *Computational Linguistics and Chinese Language Processing*, vol. 14, no. 3, pp. 281–304, 2009.

[64] S.-R. You, S.-C. Chien, C.-H. Hsu, K.-S. Chen, J.-J. Tu, J.-S. Lin, and S.-C. Chang, "Chinese-english mixed-lingual keyword spotting," in *Proc. ISCSLP*, 2004, pp. 237–240.

[65] P. Fung, X. Liu, and C. S. Cheung, "Mixed language query disambiguation," in *Proc. ACL*, 1999, pp. 333–340.

[66] T. Hastie, "A close look at deviance," *The American Statistics*, vol. 41, pp. 16–20, 1987.

[67] T. A. Myrvoll and F. K. Soong, "Optimal clustering of multivariate normal distributions using divergence and its application to hmm adaptation," in *Proc. ICASSP*, vol. 1, 2003, pp. 552–555.

[68] Y. Zhao, C. Zhang, F. K. Soong, M. Chu, and X. Xiao, "Measuring attribute dissimilarity with hmm kl divergence for speech synthesis," in *Proc. ISCA Speech Synthesis Workshop*, 2007, pp. 206–210.

[69] "Hong kong statistics - population and vital events." Hong Kong Census and Statistics Department, 2007.

[70] J. Gibbons, *Code-mixing and Code Choice: A Hong Kong Case Study.* Multilingual Matters, 1987.

[71] S. R. Ramsey, *The Languages of China.* Princeton University Press, 1987.

[72] S. Matthews and V. Yip, *Cantonese: A Comprehensive Grammar.* London, Routledge, 1994.

[73] P. C. Ching, T. Lee, W. K. Lo, and H. M. Meng, "Cantonese speech recognition and synthesis," *Advances in Chinese spoken language processing*, pp. 365–386, 2006.

[74] LSHK, *Hong Kong Jyut Ping Characters Table.* Linguistic Society of Hong Kong Press, 1997.

[75] R. Bauer, "Written cantonese in hong kong," *Cahiers de Linguistique Asie Orientale*, vol. 17, no. 2, pp. 245–293, 1988.

[76] Ethnologue, 1999.

[77] M. Wester, "Syllable classification using articulatory-acoustic features," in *Proc. Eurospeech*, 2003, pp. 233–236.

[78] P. Kam, "Pronunciation modeling for cantonese speech recognition," Master's thesis, The Chinese University of Hong Kong, 2003.

[79] W. Y. Wong, "Syllable fusion and speech rate in hong kong cantonese," in *Proc. Speech Prosody*, 2004, pp. 255–258.

[80] D. C. S. Li, *Issues in Bilingualism and Biculturalism: a Hong Kong Case Study.* Peter Lang Publishing, 1996.

[81] H. S. Chan, "Code-mixing in hong kong cantonese-english bilinguals: Constraints and processes," Master's thesis, The Chinese University of Hong Kong, 1992.

[82] J. Gumperz, *Discourse strategies.* Cambridge Uiversity Press, 1982.

[83] D. C. S. Li, "Cantonese-english code-switching research in hong kong: a y2k review," *World Englishes*, vol. 19, no. 3, pp. 305–322, 2000.

[84] N. K. Kamwangamalu and C. L. Lee, "Chinese-english code-mixing: A case of matrix language assignment," *World Englishes*, vol. 10, no. 3, pp. 247–261, 1991.

[85] N. K. Kamwangamalu, "Mixers and mixing: English across cultures," *World Englishes*, vol. 11, no. 2/3, pp. 173–181, 1992.

[86] M. Clyne, *Community language: The Australian experience.* Cambridge Uiversity Press, 1991.

[87] M. W. J. Tay, "Code switching and code mixing as a communicative strategy in multilingual discourse," *World Englishes*, vol. 8, no. 3, pp. 407–417, 1989.

[88] L. R. Cheng and K. Bulter, "Code-switching: a natural phenomenon vs language 'deficiency'," *World Englishes*, vol. 8, no. 3, pp. 293–309, 1989.

[89] J. R. Rayfield, *The language of a bilingual community.* The Hague: Mouton, 1970.

[90] D. C. S. Li, *Why do Hongkonger code-mixing? A linguistic perspective.* City university of Hong Kong, 1994.

[91] M. Chan and H. Kwok, *A Study of Lexical Borrowing from English in Hong Kong Chinese.* University of Hong Kong, 1982.

[92] Y. S. Cheung, "The use of english and chinese languages in hong kong," *Language learning & Communication*, vol. 3, no. 3, pp. 243–414, 1984.

[93] K. K. Luke and J. C. Richards, "English in hong kong : Function and status," *English Worldwide*, vol. 3, no. 1, pp. 47–63, 1982.

[94] A. Tse, "Some observations on code-switching between cantonese and english in hong kong," *Languages and Linguistics*, vol. 4, pp. 101–108, 1992.

[95] P. Norvig, "Natural language corpus data," *Beautiful Data*, 2009.

[96] "Google web 1t 5-gram corpus," LDC.

[97] B. H. S. Chan, "In search of the constrains and processes of code-mixing in hong kong cantonese-english bilingualim," City Polytechnic of Hong Kong, Research Report 33, 1993.

[98] J. Simpson and E. Weiner, *Oxford English Dictionary*, 3rd ed. Clarendon Press, 1989.

[99] S. N. Sridhar and K. K. Sridhar, "The syntax and phycholinguistics of bilingual code mixing," *Canadian Journal of Phychology*, vol. 34, no. 4, pp. 407–416, 1980.

[100] H. F. Schatz, "Code switching or borrowing? english elements in the dutch of dutch-american immigrants," *ITL (Review of Applied Linguistics)*, vol. 83/84, pp. 125–162, 1989.

[101] S. Poplack, S. Wheeler, and A. Westwood, "Distinguishing language contact phenomenon: Evidence from finnish-english bilingulism," *World Englishes*, vol. 8, no. 3, 1989.

[102] M. Chan and H. Kwok, *A Study of Lexical Borrowing from English in Hong Kong Chinese.* University of Hong Kong, 1990.

[103] D. Sankoff and S. Poplack, "A formal grammar for code-switching," *International Journal of Human Communication*, vol. 14, no. 1, pp. 3–45, 1981.

[104] S. Poplack, *"Sometimes I'll start a sentence in Spanish y termino en espanol": toward a typology of code-switching.* Cambridge: Cambridge University Press, 1982, ch. Spanish in the United State, pp. 230–263.

[105] A. M. DiSiciullo, P. Muysken, and R. Singh, "Government and code-mixing," *Journal of Linguistics*, vol. 22, pp. 1–24, 1986.

[106] E. Woolford, "Bilingual code-switching and syntactic theory," *Linguistic Inquiry*, vol. 14, pp. 520–536, 1983.

[107] M. Clyne, "Constraints on code switching: How universal are they?" *Linguistics*, vol. 25, pp. 739–764, 1987.

[108] E. G. Bokamba, "Are there syntactic constraints on code-mixing?" *World Englishes*, vol. 8, no. 3, 1988.

[109] W. K. Lo, T. Lee, and P. C. Ching, "Development of cantonese spoken language corpora for speech applications," in *Proc. ISCSLP*, 1998, pp. 102–107.

[110] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom," NIST Speech Disc1-1.1, NISTIR 4930, 1993.

[111] C.-H. Lee, L. R. Rabiner, R. Pieraccinit, and J. G. Wilpon, "Acoustic modeling of subword units for large vocabulary speaker independent speech recognition," in *Proc. DARPA Speech & Natural Language Workshop*, 1989, pp. 280–291.

[112] W. K. Lo, H. M. Meng, and P. C. Ching, "Sub-syllabic acoustic modeling across chinese dialects," in *Proc. ISCSLP*, 2000, pp. 97–100.

[113] S. J. Young, "The general use of tying in phoneme-based hmm speech recognizers," in *Proc. ICASSP*, 1992, pp. 569–572.

[114] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *Proc. ICASSP*, 2005, pp. 1–4.

[115] S. Maskey, L. Tomokiyo, and A. Black, "Bootstrapping phonetic lexicons for new languages," in *Proc. Interspeech*, 2004, pp. 69–72.

[116] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. ICSLP*, 1996, pp. 2195–2198.

[117] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, 2001.

[118] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol. 88, no. 8, 2000.

[119] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer speech and language*, pp. 171–185, 1995.

[120] D. B. Snow, *Cantonese as Written Language: the growth of a written Chinese vernacular.* Hong Kong University Press, 2004.

[121] D. B. Paul and J. M. Baker, "The design for the wall street. journal-based csr corpus," in *Proc. DARPA Speech and Language Workshop*, 1992, pp. 357–362.

[122] S. Martin, J. Liermann, and H. Ney, "Algorithms for bigram and trigram word clustering," *Speech Communication*, vol. 24, no. 1, 1998.

[123] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, 1992.

[124] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical langugae modelling," in *Proc. EUROSPEECH*, 1993, pp. 973–976.

[125] C. Fellbaum, *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[126] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book (for HTK Version 3.4).* Cambridge University, 2009.