Pronunciation Modeling for Cantonese Speech Recognition

KAM Patgi

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy

in

Electronic Engineering

July 2003

Abstract of thesis entitled:

Pronunciation Modeling for Cantonese Speech Recognition

Submitted by **KAM Patgi**

for the degree of **Master of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong**

in **July 2003**.

The primary goal of automatic speech recognition (ASR) is to produce a textual transcription for spoken input. This can be done by establishing a mapping between the extracted acoustic features and the underlying linguistic representations. Given the high variability of human speech, each linguistic symbol may have multiple pronunciations. Pronunciation modeling for ASR is aimed at providing a mechanism by which speech recognition systems can be adapted to pronunciation variability. The objective of this thesis is to investigate various types of pronunciation variations in Cantonese speech and incorporate pronunciation model (PM) in Cantonese ASR in order to improve the performance of recognition.

In a large-vocabulary continuous speech recognition (LVCSR) system, three knowledge sources are involved: pronunciation lexicon, acoustic model and language model. A decoding algorithm is used to search for the most likely word sequence. Based on this framework, pronunciation modeling can be done by explicitly modifying these knowledge sources and/or improving the decoding technique.

There are two types of pronunciation variations, namely phone change and sound change. Phone change means that a phoneme is completely realized as another phoneme. Sound change happens when the acoustic realization is ambiguous between two phonemes.

Phone change can be handled by constructing an augmented dictionary to include alternative pronunciations at lexical level or expanding the search space to include pronunciation variants at decoding level.

Sound change can be handled by adjusting the acoustic model through Gaussian mixtures sharing and adaptation or modifying the state output probability to include variation information during the search process.

In this research, different approaches as mentioned above have been investigated for pronunciation modeling of Cantonese. The effectiveness of these approaches is evaluated with extensive experimental results. When compared with the baseline system, among the various methods, relative error reduction of 7.30%, 5.45% and 8.17% are achieved in a Cantonese LVCSR task when we apply pronunciation modeling at lexical level, acoustic model level and decoding level respectively.

摘要

自動語音識別 (ASR) 的主要目標是把人的口述語言逐詞逐句地轉換爲相應的書面語言(文字)。它能通過在所提取的聲學特徵和語言學的表示之間構建一個映射來實現。但口語有很大的可變性，每個語言學上的符號可有多個讀法。發音模型 (Pronunciation Model) 爲語音識別系統提供了一個自動適應發音變化的機制。

大辭彙量連續語音識別系統 (LVCSR) 主要包括三個知識源：發音詞典，聲學模型和語言模型以及一個用來搜索最佳詞序列的解碼器。發音模型能擴充知識源和改進解碼器的搜索技術。

發音變異可分爲兩類：音素替換 (Phone Change) 和音位變體 (Sound Change)。音素替換從一個音素完全變爲另外一個音素，音位變體是同一個音位的多個音，這些音常常介於兩個音素之間。不同層次上的發音模型能處理不同類型的發音變異。

音素替換：在詞條層次上，通過構建一個大的發音詞典來處理，這個詞典包括了所有可能發生的音素替換；或者在解碼層次上，通過擴展搜索空間來處理。

音位變體：在聲學模型層次上，通過用高斯混合分量共用及自適應的方法來重新訓練聲學模型來實現；或者通過在搜索過程中改良馬爾可夫模型的狀態輸出概率來實現。

本論文將討論上述處理發音變異的方法以及用詳細的實驗結果來比較這些方法的性能。在廣東話大詞彙量連續語音識別系統中，分別於詞條層次，聲學模型層次和解碼層次上應用發音模型，與基線試驗結果相比，相對錯誤率減少率分別爲 7.30%, 5.45% and 8.17%。

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Prof. Tan Lee for his guidance and support throughout this research work. He has contributed a lot to the ideas of this dissertation. With his supervision and advice, I learnt a lot from his wide knowledge, outstanding insight and positive and earnest working manner.

Thanks are due to Prof. P.C. Ching who has always been helpful to provide many valuable suggestions. I would also like to thank Prof. Helen Meng, Prof. Y.T. Chan, Prof. X.G. Xia for their precious advices.

Sincere thanks are given to Dr. F. Soong who deals in truth and in sincerity with all his students. I benefited much from his constructive suggestions, sharing of his experience and his encouragement.

I would like to thank all the colleagues and friends in DSP Group who helped me in many different ways. To name only some of them: L.Y. Ngan, C.H. Yau, K.Y. Kwan, W.N. Choi, Y.J. Li, Y. Qian, C. Yang, B.R. Chen, S.W. Lee, Y.Y. Tam. The technical assistance from Dr. W.K. Lo is much appreciated.

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

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Speech communication is the dominant mode of human interaction and information exchange. Most conventional computer operating systems and applications depend on keyboard and mouse as user input. Spoken language technology is developed to make computer systems more user-friendly as the computer will have the fundamental human abilities to speak, listen, understand and learn.

Automatic speech recognition (ASR) has become a hot topic of research for many years. Statistical methods of converting spoken utterances into meaningful text have been extensively investigated in recent years. However, due to the large variability of speech, the performance of ASR systems is considered inadequate in many aspects. Researches are undergoing to take into account the variability of speech in the ASR process. One of the directions is to provide a mechanism by which speech recognition systems can be adapted to pronunciation variations.

The research described in this thesis is aimed at handling different types of pronunciation variations in continuous Cantonese speech by incorporating pronunciation model into the ASR system. The ultimate goal is to improve the accuracy of recognition.

1.1 Automatic Speech Recognition

Given an input speech utterance, ASR is to produce a highly probable hypothesis of the underlying word sequence being spoken. This can be done by establishing a mapping between properly extracted acoustic features and linguistic representations. The research on ASR started over 50 years ago. Many different approaches in ASR

evolved in these years. Recently a statistical approach is commonly adopted [1]-[4]. In this approach, as shown in Figure 1.1, speech recognition is accomplished with three knowledge sources, namely the pronunciation lexicon, the acoustic model (AM) and the language model (LM). They together define a search space from which the most likely sentence(s) or word string(s) can be determined with an efficient search algorithm [4].

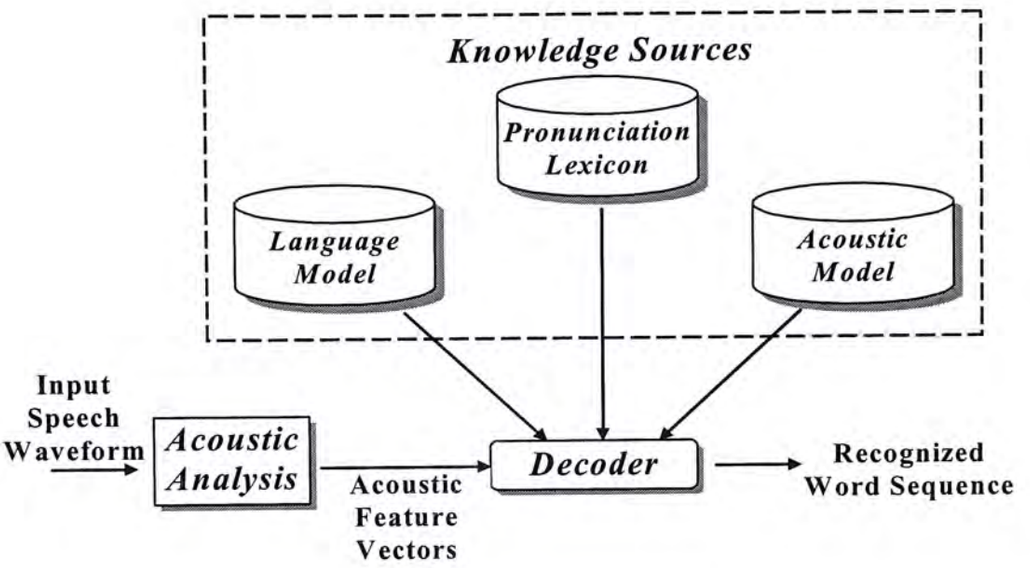


Figure 1.1: Block diagram of an ASR system

The acoustic model characterizes statistical variation of the acoustic properties of sound units. It produces the probability for the input acoustic features being observed. The pronunciation lexicon provides constraints on the combination of the sound units being modeled at the lowest linguistic level to form a lexical word. The language model provides the constraints on how words can be connected to form a sentence. The lexicon and the language model are independent of the acoustic observations. The decoder is responsible for deciding the most probable and legitimate word sequence.

1.2 Pronunciation Modeling in ASR

Speech exchanges information between a speaker and a listener. It begins with a message in the brain, which activates muscular movements to produce speech

sounds. Speech is a complex combination of information from many different levels, including discourse, semantics, syntax, phonology, phonetics and acoustics. This makes the actual realization of a particular speech sound contain a great deal of variability. Pronunciation variations seriously deteriorate the performance of an ASR system if they are not handled properly. Factors causing pronunciation variation can be divided into two categories: inter-speaker and intra-speaker. The major difficulty for unrestricted, speaker-independent continuous speech recognition is due to the diversified speaker characteristics such as dialectal accents, speaking styles, gender and psychological conditions. Even the speech produced by the same speaker may contain substantial variation, which may be caused by co-articulation, speaking rate, physical and emotional condition [5][6]. All these problems make continuous speech recognition a very difficult task.

Given the high variability of human speech, the mapping between the acoustic features and the underlying linguistic representations is not one-to-one. Different linguistic symbols can give rise to similar speech sounds while each symbol may have multiple pronunciations. Pronunciation modeling of ASR is aimed at providing a mechanism by which speech recognition systems can be adapted to pronunciation variability.

Pronunciation variations can be roughly classified into two types: phone change and sound change [6][7]. A phone change happens when a canonical (*baseform*) phoneme is realized as another (*surfaceform*) phoneme. The baseform pronunciation is assumed to be the "standard" pronunciation that the speaker is supposed to use. Surfaceform pronunciations are the actual pronunciations that different speakers may use. Phone change can be considered as the baseform phoneme being substituted by another (surfaceform) phoneme. A sound change happens at a lower level, e.g. phonetic or sub-phonetic level. Acoustically, the sound unit is neither the baseform nor any surfaceform phoneme.

Phone change can be handled by replacing the baseform transcription by the actual pronunciation observed in acoustic signal, which is the surfaceform transcription. This is accomplished by augmenting the standard baseform lexicon with additional pronunciation variants for each word entry [8]-[11] or expanding the

search space to include those variations during sentence decoding [12]. M.K. Liu *et al* suggested to build an accent-specific Chinese syllable pronunciation variation dictionary by using context-independent and context-dependent syllable confusion matrices [8]. C. Huang *et al* proposed a method of accent modeling through pronunciation dictionary adaptation (PDA) [9]. These are the most straightforward methods to expand the lexicon by using realistic variation information directly observed from speech data. However, due to imperfect recognition results, the observed variation may not be accurate. To solve this problem, Byrne *et al* started from a hand-labeled corpus to build an augmented pronunciation lexicon using an iterative approach [10]. This augmented lexicon is supposed to contain more accurate surfaceform information to cope with pronunciation variation.

To handle a sound change, pronunciation modeling must be applied at a lower level, for example, at state or Gaussian mixture level in a Hidden Markov Model based ASR system. The acoustic model is usually trained with only the knowledge about baseform pronunciations and no alternative pronunciations are considered at all. It is assumed that the speakers always follow the standard pronunciations and realize them exactly all the time. This convenient but obviously inadequate assumption renders the acoustic model thus trained to be unable to represent the variations of speech sounds. It would be useful to refine the acoustic model by taking into account the realistic pronunciations [6][7][13][14]. Y. Liu and M. Saraclar used the surfaceform model in the existing set of acoustic models to refine the baseform model. Y. Liu proposed using partial change phone model (PCPM) as well as auxiliary decision tree to model partial changes [6]. M. Saraclar *et al* suggested to refine the acoustic model by sharing the Gaussian mixture pdf's [14]. In this method, all the mixture components in the surfaceform models are used to enrich the baseform models. This may lead to a problem that some inappropriate surfaceform mixture components are also used. V. Venkataramani *et al* refined the acoustic model by MLLR method [13]. It requires extra training data with pronunciation variations for the adaptation process. M. Saraclar *et al* also trained a new set of acoustic model based on both baseform and surfaceform pairs [7].

1.3 Objectives of the Thesis

The main objective of this thesis is to investigate different types of pronunciation variations in Cantonese speech and incorporate pronunciation model (PM) in Cantonese ASR in order to improve the performance of recognition.

Two types of variations, phone change and sound change, are considered. As their characteristics are different, different methods of modeling are used. We replace the baseform phoneme by the surfaceform phoneme to handle phone change. This can be done by incorporating the PM into the lexicon to form an augmented lexicon with variation information. However, adding all variations into the lexicon will cause confusion results in degradation of recognition. Therefore, different ranking and pruning methods are investigated in order to include only those useful variations.

In order to cope with sound changes, we investigate different algorithms to refine the acoustic model. The refinement of acoustic model includes sharing of Gaussian mixture components and mixture adaptation.

Apart from augmenting the knowledge source, PM can be incorporated in the search process. Pronunciation modeling in decoding process not only can deal with phone change by expanding the search space, but also can handle sound change by modifying the calculation of the state output probability to include variation information.

1.4 Thesis Outline

In the next two chapters, the background knowledge for Cantonese continuous speech recognition will be provided. Cantonese phonology and phonetics will be introduced in Chapter 2. The concept of phone change and sound change will also be explained. The fundamentals of large-vocabulary continuous speech recognition (LVCSR) for Cantonese will be discussed in Chapter 3.

The construction of pronunciation model is introduced in Chapter 4. The PMs we used are context-independent IF confusion matrix and context-dependent decision tree pronunciation model. These models will be used for pronunciation modeling at different levels, as discussed in Chapter 5, 6 and 7.

In Chapter 5, pronunciation modeling at lexical level to handle phone change will be presented. Different methods of constructing a pronunciation variation dictionary (PVD) will be evaluated with recognition experiments.

In Chapter 6, we focus on incorporating PM in acoustic model to handle sound change. Different techniques are investigated to refine the acoustic model to include the variation information by sharing or adaptation of Gaussian mixture components.

In Chapter 7, we present the method of pronunciation modeling at decoding level. The search space is expanded dynamically during sentence decoding for handling phone change and the calculation of the state output probability is modified to deal with sound change by including variation information.

Chapter 8 will conclude this thesis with some suggestions for future research.

Reference

[1] L.R. Bahl *et al*, "A Maximum Likelihood Approach to Continuous Speech Recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol.5, pp.179-190, 1983.

[2] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in *Proceedings of the IEEE*, Vol.77, no.2, pp.257 –286, 1989.

[3] L.E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes", *Inequalities*, Vol.1, pp.1-8, 1972.

[4] W.N. Choi, "An Efficient Decoding Method for Continuous Speech Recognition Based on a Tree-Structured Lexicon", *M.Phil. Thesis*, The Chinese University of Hong Kong, 2001.

[5] M.Y. Tsai *et al*, "Pronunciation Variation Analysis with respect to Various Linguistic Levels and Contextual Conditions for Mandarin Chinese", in *Proceedings of Eurospeech-01*, Vol.2, pp.1445-1448, Alborg, 2001.

[6] Y. Liu, "Pronunciation Modeling for Spontaneous Mandarin Speech Recognition", *Ph.D. Thesis*, The Hong Kong University of Science and Technology, 2002.

[7] M. Saraclar *et al*, "Pronunciation Ambiguity vs Pronunciation Variability in Speech Recognition", in *Proceedings of ICASSP-00*, Vol.3, pp.1679-1682, Istanbul, 2000.

[8] M.K. Liu *et al*, "Mandarin Accent Adaptation Based on Context-Independent/Context-Dependent Pronunciation Modeling", in *Proceedings of ICASSP-00*, Vol.2, pp.1025-1028, Istanbul, 2000.

[9] C. Huang *et al*, "Accent Modeling Based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech Recognition", in *Proceedings of ICSLP-00*, Vol.3, pp.818-821, Beijing, 2000.

[10] W. Byrne, *et al*. "Pronunciation Modelling Using a Hand-labelled Corpus for Conversational Speech Recognition", in *Proceedings of ICASSP-98*, Vol.1, pp.12-15, Seattle, 1998.

[11] W. Byrne, *et al*. "Automatic Generation of Pronunciation Lexicon for Mandarin Spontaneous Speech", in *Proceedings of ICASSP-01*, Vol.1, pp.569-572, Salt Lake City, 2001.

[12] P. Kam *et al*, "Modeling Pronunciation Variation for Cantonese Speech Recognition", in *Proceedings of PMLA-02*, pp.12-17, Denver, 2002.

[13] V. Venkataramani *et al*, "MLLR Adaptation Techniques for Pronunciation Modeling", *ASRU-01*, CDROM, Trento, 2001.

[14] M. Saraclar *et al*, "Pronunciation Modeling by Sharing Gaussian Densities across Phonetic Models", in *Proceedings of Eurospeech-99*, Vol.1, pp.515-518, Hungary, 1999.

Chapter 2

The Cantonese Dialect

Speech signals are composed of analog sound patterns that serve as the basis for a discrete and symbolic representation of the spoken language. These discrete symbols are typically phonemes, syllables, and words. Phoneme is the basic contrastive sound in the phonological system of a particular language or dialect [1]-[3]. A word is made up of a sequence of phonemes, which essentially define the pronunciation of the words. Table 2.1 shows the phoneme sequence for the Chinese phrase 香港中文大學 (CUHK) spoken in Cantonese.

Chinese phrase	香	港	中	文	大	學
Phoneme sequence (IPA[1])	h œŋ	k ɔŋ	ts ʊŋ	m ɐn	t ai	h ɔk
Initial/Final sequence (LSHK[2])	/h/ /oeng/	/g/ /ong/	/z/ /ung/	/m/ /an/	/d/ /aai/	/h/ /ok/

Table 2.1: Phoneme sequence of Chinese phrase 香港中文大學 spoken in Cantonese.

To study the phenomenon of pronunciation variation in Cantonese speech, a good understanding of this language is indispensable. This chapter will describe the phonology and phonetics of Cantonese. Phonology is the science of language that deals with the distribution and patterning of speech sounds and the rules that govern the formation of valid sounds from the sound units. Phonetics is the study of speech sounds and their production, classification and transcription.

---

[1] International Phonetic Alphabet
[2] The phonetic symbols for Cantonese proposed by the Linguistic Society of Hong Kong

2.1 Cantonese – A Typical Chinese Dialect

Cantonese is one of the most dominant Chinese dialects used in Southern China, Hong Kong and among many overseas Chinese communities. It is spoken by more than 60 millions people all over the world [1]. The basic unit of written Cantonese is Chinese character [2]. Chinese characters are ideographic, meaning that characters contain no information about pronunciation. They are generally homophonic [2]. Each pronunciation will map to many Chinese characters. On the other hand, Chinese is homographic [2], meaning that the same character can have several pronunciations with different meanings.

Each Chinese character has its own meaning(s) and can play a linguistically independent role [2]. A Chinese word may consist of one or more characters. Each word has a specific meaning and can be used individually. In written Chinese, words are connected together one after another in a sentence without explicit boundaries. Segmentation of a sentence into words by different readers may be different. There are more than ninety thousand words in Chinese and the number of commonly used characters is about ten thousand. Most (more than 70%) Chinese words are bisyllabic. Monosyllabic words form a substantial set of frequently used words.

The pronunciation of Cantonese is usually represented in the form of syllables [3][4]. Syllable is the basic unit in spoken Cantonese. Each Cantonese syllable represents many Chinese characters while each Chinese character can have several pronunciations represented by different syllables. Some examples of homophones and homographs in Cantonese are listed in Table 2.2.

Homophones	zung1	中, 宗, 忠, 盅, 終, 縱, 鐘,…
	man4	文, 民, 玟, 雯, 紋, 聞,…
Homographs	行	hong2, hong4, hang4, haang4
	生	sang1, saang1

Table 2.2: Example of Cantonese homophones and homographs.

2.1.1 Cantonese Phonology

As a spoken language, Cantonese is quite different from Western languages. Like Mandarin, Cantonese is monosyllabic. Each Chinese character is pronounced as a single syllable sound. Cantonese is also a tonal language. The tone of a syllable carries lexical meaning, i.e. the variation in the pitch pattern of a syllable changes it to another character. Cantonese is said to have nine citation tones according to the pitch contours as shown in Figure 2.1. The first six tones are called *non-entering tone* while the remaining ones are called *entering tone*. The entering tones, carried by syllables that end with *-p*, *-t*, *-k*, are generally regarded as short counterparts of the non-entering tones 1, 3 and 6 respectively. As a result, a six-tone system is commonly used for Cantonese.

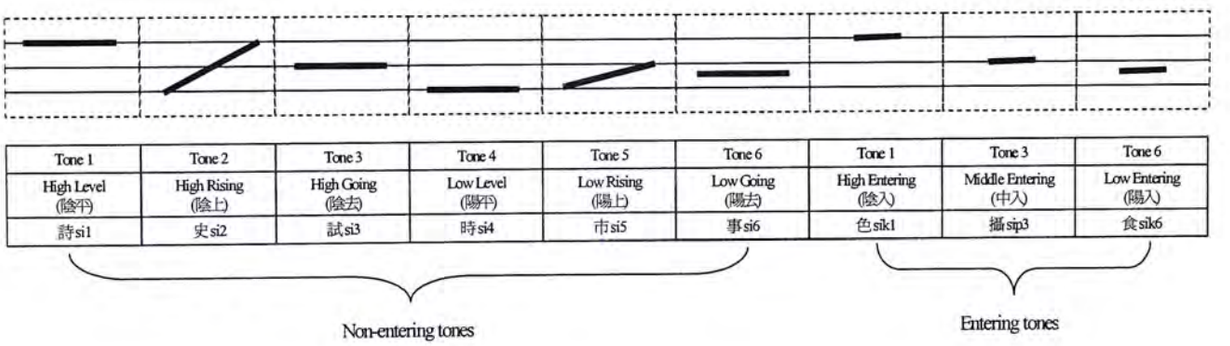


Figure 2.1: Pitch profiles of nine citation tones in Cantonese.

A Cantonese syllable has a structured form of a beginning *Initial* (I) followed by a *Final* (F) [3][4]. A Final can be further decomposed into a vowel nucleus and a consonant coda. Table 2.3 shows the structure of a Chinese word. The Chinese word 我們 (we) consists of two syllables. The first syllable, "ngo", is formed by the Initial /ng/ and the Final /o/. The second syllable is formed by the Initial /m/ and the Final /un/.

Chinese word	Chinese character	Base syllable	IF units
我們	我	ngo	/ng/ /o/
	們	mun	/m/ /un/

Table 2.3: Structure of a Chinese word.

Altogether there are 19 Initials and 53 Finals in Cantonese. Initials and Finals are combined under certain phonological constraints. There are more than 600 legitimate Initial-Final (IF) combinations listed in Appendix I [5]. Each of these IF combinations is also referred to as a base syllable. For phonemic transcription at IF level, the LSHK scheme is adopted in our work. IPA symbols for the Cantonese Initials and Finals are also given in Appendix II.

If different tones are considered, there are about 1,800 Cantonese tonal syllables. The phonological structure of Cantonese syllables is shown as in Figure 2.2. The numbers in the brackets () are the total number of the respective units. The units shown in square brackets [] are optional and may not appear in a syllable. We define the *null* Initial to represent a deleted Initial in a syllable with only the Final part. In Cantonese, all Initial onsets are consonants while all nuclei are vowels. Cantonese codas fall into three main classes: stop, nasal and vowel. Except for the short vowel *-a-*, the vowel nucleus can be a Final by itself.

<table>
<tr><td colspan="3">Tonal Syllable (1,800)</td></tr>
<tr><td colspan="2">Base Syllable (665)</td><td rowspan="3">Tone (6)</td></tr>
<tr><td>Initial (19)</td><td colspan="2">Final (53)</td></tr>
<tr><td>[Onset] (19)</td><td>Nucleus (8)</td><td>[Coda] (8)</td></tr>
</table>

Figure 2.2: Phonological structure of Cantonese syllables.

In speech recognition, the base syllable and the tone can be recognized individually because they primarily concern different acoustic aspects of speech signals. In this thesis, we will only focus on recognition of the base syllable.

2.1.2 Cantonese Phonetics

In linguistic theory, sound units can be divided into two types: phonemes and phones [3][4]. Phoneme is the smallest speech unit in the formation of a particular language or dialect. It forms the smallest set of unambiguous symbols that altogether will be sufficient for representing the language [3][4]. Phoneme is language and dialect dependent. Replacing any of the phoneme in a syllable would result in another

syllable, for example, replacing Initial /z/ in "zung" by /c/ results in another syllable "cung".

Phones, on the other hand, are the smallest sound-building units that are physically differentiable [3][4]. Different phones are formed physically by changing the place and manner of articulation during speech production. They are the fundamental sound categories that describe the range of acoustic features. Phones are generally classified into two categories: consonants and vowels.

Consonants are typically featured by noise-like properties in acoustic speech signals [3]. Different consonants are related to where and how the air flowing in the vocal tract is interrupted. In Cantonese, consonants always play the role of Initials, e.g. /b/, /d/, /g/, /p/, /t/, /k/, and codas, e.g. *-p*, *-t*, *-k*, etc.

Vowels are featured by the periodic and voiced properties of speech signals. They are generated by periodic oscillation of the vocal cord producing periodic air flowing through the vocal tract. Different vowels are produced by changing the size of the vocal tract. In Cantonese, vowels always play the role of syllable nucleus, for example, *-aa-*, *-a-*, *-i-*, *-yu-*, *-u-*, *-e-*, *-oe-*, *-eo-* and *-o-*.

2.2 Pronunciation Variation in Cantonese

Acoustic speech can be represented in terms of either phonemes or phones. Phonemic transcription, also known as baseform or canonical transcription, represents a spoken utterance by a sequence of phonemes [6]. Baseform transcription is the standard pronunciation that a speaker is supposed to use. It does not contain variation information.

On the other hand, phonetic transcription, also referred as surfaceform transcription, is the transcription in accordance with actual phone realizations. The representation of speech in terms of baseform or surfaceform may be very different. Table 2.4 shows the baseform and some possible surfaceform transcriptions for the word 我們.

Chinese word	Baseform	Possible surfaceforms
我們	ngo mun	ngo mun
		o mun
		ngo wun
		o wun

Table 2.4: Baseform and some possible surfaceform transcriptions of the word 我們

2.2.1 Phone Change and Sound Change

Pronunciation variations can be roughly classified into two types: phone change and sound change [3][6][7]. Phone change is the realization of a baseform phoneme by another surfaceform phoneme where the surfaceform can be identified. It is a complete change of one phoneme to another, for example, /n/ changes to /l/, symbolized as /n/→/l/.

Sound change is the pronunciation variation between two phonemes. Even human transcribers can hardly agree on the identity of the surfaceform. It is a partial change [6] of baseform phoneme with its surfaceform, for example, /n/ varies with /l/, symbolized as /n/~/l/. When the pronunciation of a phone is ambiguous between the realizations of two phonemes, it is not appropriate to label the phone by either one of these two phonemes. Thus, sound change cannot be modeled by simply substituting the canonical phoneme with another phoneme.

Over the past 15 years, sociolinguists have focused their attention on phonetic variations in Cantonese by correlating phonetic variables with social characteristics of speakers such as sex, age and educational level. These sociolinguistic studies have revealed that systematic patterns underlie the phonetic variations [3]. Table 2.5 shows the observations from several sociolinguistic studies of phonetic variations in Cantonese.

Initial consonants	/n/→/l/	Phone change of nasal to lateral.
	/n/~/l/	Sound change between nasal and lateral
	/ng/→/null/	Phone change of velar nasal to null Initial.
	/ng/~/null/	Sound change between velar nasal and null Initial.
	/gw/→/g/	Phone change of labialized velar to delabialized velar before back round vowel /o/. (Delabialization of labialized velar.)
	/gw/~/g/	Sound change between labialized and delabialized velars.
Nasal syllabics	/ng/→/m/	Phone change of velar nasal to bilabial syllabic before/after labial consonants. (Labial assimilation of velar nasal.)
	/ng/~/m/	Sound change between velar and bilabial nasal syllabic.
Final consonants	-ng→-n	Phone change of velar nasal Final to dental nasal Final.
	-ng ~ -n	Sound change between velar nasal Final and dental nasal Final.
	-k→-t or -p	Phone change of velar stop Final to dental or glottal stop Final.
	-k ~ -t or *-p*	Sound change between velar stop Final and dental or glottal stop Final.

Table 2.5: Observations of phonetic variations in Cantonese from sociolinguistic studies

One of the reasons that explains these observations is the distinct physiological characteristics among different speakers [3]. For example, /n/→/l/, /ng/→/null/, /gw/→/g/ is correlated with the sex and age of a speaker [8]. Bourgerie found that female uses more /l/ for /n/ and /null/ for /ng/ than male. The older age group has a much lower frequency of /l/, /null/ and /g/ than younger speakers.

The use of /l/ for /n/, /null/ for /ng/ and /g/ for /gw/ are also inversely correlated with the formality of the speech situation [8]; as the level of formality declines, the frequency of /l/, /null/ and /g/ increases.

Bauer stated that these variations may also be related to the developments in neighboring dialects of the Pearl River Delta [3]: in Panyu and Shunde, /null/ regularly corresponds to both standard Cantonese Initial /ng/ and /null/, but in the Dongguan-Guancheng dialect, /ng/ corresponds to both standard Cantonese Initial /ng/ and /null/. The plain velar Initial /g/ corresponds regularly to the standard Cantonese labialized velar /gw/ in Conghua, Zhongshan-Shiqi, Xinhui, Taishan. Bauer even attributed one's pronunciation changes to influence from one's mother

who was from the Dongguan district where Dongguan area had reported phonetic changes of -ng→-n, *-k→-t* or *-p*.

Phone change and sound change can also occur through the mutual influence of adjacent phonemes [3]. *Assimilation* is a kind of phonetic change in which one phoneme becomes similar to a neighboring phoneme by acquiring a phonetic feature of it. When the preceding syllable ended in a nasal consonant, Ho found that a nasal assimilation effect in which subjects used more /n/ than /l/ [9], for example, "soeng nei" 想你 (think of you). Labial dissimilation has probably been the cause of the change /gw/→/g/ when they occur before the vowel nucleus /o/ [3], for example, changing "gwok" 國 (country) to "gok" 角 (corner). The sequence of the two lip-rounded segments /w/ and /o/ has become redundant or unnecessary with the second one driving out the first. The change /ng/→/m/ is due to the fact that when velar nasal /ng/ occurred in the environment of a bilabial sound segment, say /p/, its place of articulation assimilated to bilabial under the influence of the neighboring bilabial sound /p/ or *-p*. For example, "sap ng" 十五 (fifteen) becomes "sap m" through the perseverance of the bilabial closure of Final *-p* into the articulation of the following nasal syllabic. This phenomenon is termed as *perseveratory assimilation* [3].

Other pronunciation variations may occur due to dialectal accent of non-native speakers. These speakers may have difficulties to master some of the Cantonese pronunciations. They sometimes use the pronunciation of their own native language to pronounce a Cantonese word, for example, "ngo" 我 (me) will be pronounced as "wo" by a native Mandarin speaker.

2.2.2 Notation for Different Sound Units

In this section, the notations for different sound units we used in this thesis are defined. The baseform and the surfaceform sequences at Initial-Final level are denoted as *B* and *S,* respectively. Both of them are expressed using the LSHK scheme. A baseform IF and a surfaceform IF are denoted by *b* and *s* respectively. The notation / / represents an IF, for example /aa/. Vowel and coda can be represented by *-vowel-* and *-coda*, for example, *-aa-*, *-a-*, *-i-* and *-p*, *-t*, *-k* respectively.

{ } indicates an IF sequence, for example {/ng/ /o/ /m/ /un/}. " " refers to a syllable, for example "ngo". Surfaceform phoneme will be used only to represent the realization of baseform phoneme having phone change, such as, /b/ → /p/, /ng/ → /null/, /m/ → /w/, /gw/ → /g/, /n/→ /l/, /aai/ → /ai/, etc. It will not be used to represent sound change.

2.3 Summary

In this chapter, we gave a general introduction to the Cantonese dialect. The phonology and the phonetics of Cantonese were described. The definition of baseform transcription and surfaceform transcription is given. We also discussed about the two major categories of pronunciation variations, namely phone change and sound change, in Cantonese. Phone change can be represented by substitution with a different surfaceform phoneme while the ambiguous pronunciations caused by sound change would make surfaceform transcription uncertain.

Reference

[1] B.F. Grimes et al, Ethnologue, Languages of the World, SIL International, 2000.

[2] Yuan Ren Chao, *A Grammar of Spoken Chinese*, University of California Press, 1965.

[3] R.S. Bauer *et al*, *Trends in Linguistics, Studies and Monographs 102, Modern Cantonese Phonology*, Mouton de Gruyter, Berlin, New York, 1997.

[4] W.K. Lo, "Cantonese Phonology and Phonetics: an Engineering Introduction", *Internal Document*, Speech Processing Laboratory, Department of Electronic Engineering, the Chinese University of Hong Kong, 1999.

[5] Y.W. Wong, "Large Vocabulary Continuous Speech Recognition for Cantonese," *M.Phil. Thesis*, The Chinese University of Hong Kong 2000.

[6] Y. Liu, "Pronunciation Modeling for Spontaneous Mandarin Speech Recognition", *Ph.D. Thesis*, The Hong Kong University of Science and Technology, 2002.

[7] M. Saraclar *et al*, "Pronunciation Ambiguity vs Pronunciation Variability in Speech Recognition", in *Proceedings of ICASSP-00*, Vol.3, pp.1679-1682, Istanbul, 2000.

[8] D.S. Bourgerie, "A Quantitative Study of Sociolinguistic Variation in Cantonese", *Ph.D. Thesis*, The Ohio State University, 1990.

[9] M.T. Ho, "(n-) and (l-) in Hong Kong Cantonese: A Sociolinguistic Case Study", *M.A. Thesis*, University of Essex, 1994.

Chapter 3

Large-Vocabulary Continuous Speech Recognition for Cantonese

In this chapter, the fundamental principles of large-vocabulary continuous speech recognition (LVCSR) will be reviewed. Subsequently the details of a Cantonese LVCSR system will be described. This system will be used as the experimental baseline for the investigation of various pronunciation modeling approaches.

The ultimate goal of a speech recognition system is to convert the input speech utterance into its written form. The technology is multi-disciplinary, requiring the use of advanced techniques in signal processing, pattern recognition and linguistic processing. One of the most successful and widely used approaches of speech recognition is the statistical approach. This approach, without relying on the knowledge of an expert, offers a way to systematically organize the knowledge about the speech communication process. Given an input utterance, speech recognition is formulated as a probabilistic process to determine the most likely word sequence. It involves three knowledge sources, namely acoustic model, pronunciation lexicon and language model, as shown in Figure 3.1.

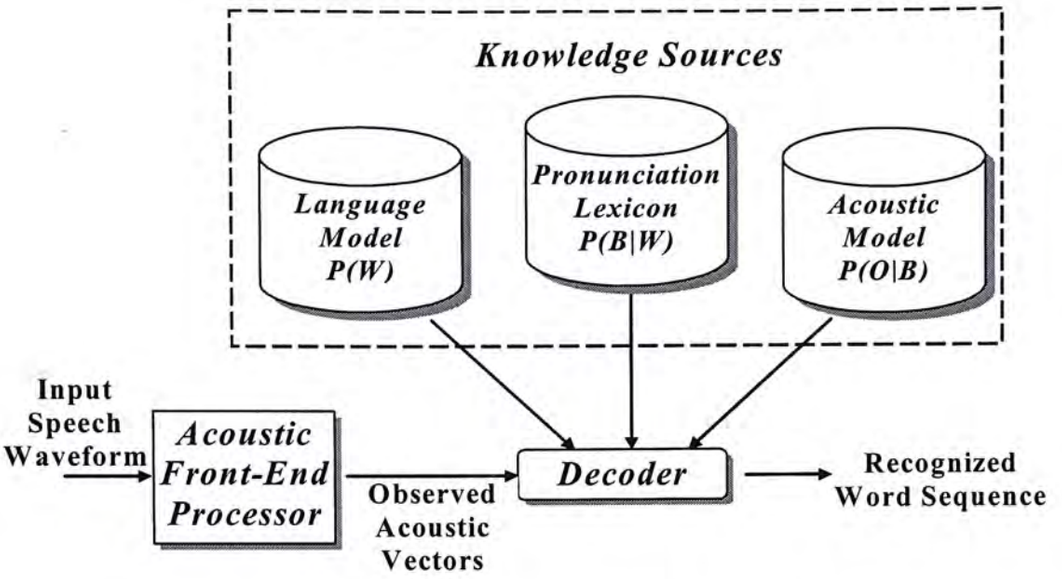


Figure 3.1: Block diagram of a typical speech recognition system.

3.1 Feature Representation of the Speech Signal

Digitized speech signal first go through the acoustic front-end, which extracts a sequence of feature vectors from the time-domain signals. The feature vectors provide a compact spectral and temporal representation of the speech signal. Feature extraction is performed on a frame basis. Each short-time frame is analyzed to generate a feature vector. The most commonly used features in speech recognition include Mel Frequency Cepstral Coefficient (MFCC) [1], Linear Predictive Coding (LPC) [2], Perceptual Linear Prediction (PLP) [3], etc. They all use a small number of parameters to represent the properties of speech signal. In this research, MFCC will be used.

3.2 Probabilistic Framework of ASR

Given an acoustic signal, a sequence of feature vectors O is obtained by feature extraction. The goal of ASR is to find the most probable word sequence W that maximizes the probability $P(W|O)$, i.e.

$$W^* = \arg\max_W P(W \mid O) \qquad (3.1)$$

$P(W|O)$ is also commonly known as the *a posterior* probability. According to the Bayesian decision rule, equation (3.1) can be rewritten as

$$W^* = \arg\max_W P(O \mid W)P(W) \qquad (3.2)$$

where $P(W)$ denotes the *a priori* probability of the word sequence W and $P(O|W)$ represents the probability of the acoustic features O being observed when W is spoken. $P(W)$ is given by the language model while $P(O|W)$ is computed based on the acoustic model and the pronunciation lexicon. Usually W is represented as a sequence of sub-word units, denoted by B. If the acoustic model is built based on these sub-word units, equation (3.2) becomes

$$W^* = \arg\max_W P(W)P(O \mid B)P(B \mid W) \qquad (3.3)$$

where $P(O|W)$ has been decomposed into $P(O|B)$ and $P(B|W)$. $P(O|B)$ is the probability of O being observed when B is given. It is computed from the sub-word acoustic model. $P(B|W)$, obtained from the lexicon, gives the probability that W is pronounced as the sub-word sequence B. For conventional lexicon in which a single realization is assumed for each word, $P(B|W)$ always equals to 1.0.

3.3 Hidden Markov Model for Acoustic Modeling

Acoustic model computes the probability of the acoustic features O being observed when the word sequence W is given. Word-level acoustic model is seldom used for large vocabulary applications, as the number of models required would be too large to be practical. Instead, sub-word level acoustic model is often used, for examples, phoneme or IF models (for Chinese). Sub-word acoustic model gives the probability $P(O|B)$ for O given the sub-word sequence B.

A powerful statistical method for representing the speech signal is Hidden Markov Model (HMM) [4]. An HMM is a finite-state machine. Each state is defined

with a probability density function (pdf). Each transition between the states is also governed by a probability. Figure 3.2 shows the structure of an HMM. The number of states depends on the complexity of the unit being modeled. In practice, we known the observed vector sequence O only, while the underlying state sequence is hidden. Therefore, it is termed as "Hidden" Markov Model.

The pdf at each HMM state is typically a mixture of multivariate Gaussian distribution functions. At a particular time instant, the probability of the feature vector o_t being generated from state j, also termed as state output probability, is computed as

$$p_j(o_t) = \sum_{m=1}^{M} w_{jm} N(o_t; \mu_{jm}, \Sigma_{jm})$$
$$N(o_t; \mu_{jm}, \Sigma_{jm}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_{jm}|}} e^{-\frac{1}{2}(o_t-\mu_{jm})'\Sigma_{jm}^{-1}(o_t-\mu_{jm})} \quad (3.4)$$

where M is the number of Gaussian mixture components in the j-th state, and w_{jm} is the weight for the m-th mixture component. $N(o_t; \mu_{jm}, \Sigma_{jm})$ denotes the multivariate Gaussian distribution with the mean vector μ_{jm} and covariance matrix Σ_{jm}.

In Figure 3.2, the HMM that models the Cantonese Initial /b/ is shown as an example. It has three states denoted by $I\_b(1)$, $I\_b(2)$ and $I\_b(3)$. Each state is associated with M Gaussian pdf's to model the output distribution. Given the acoustic observation sequence, $o_1, o_2, \ldots, o_7$, the state output probability is given by $p_1(o_1), p_1(o_2), \ldots, p_3(o_7)$. The probability of the transition from state i to state j is denoted by a_{ij}.

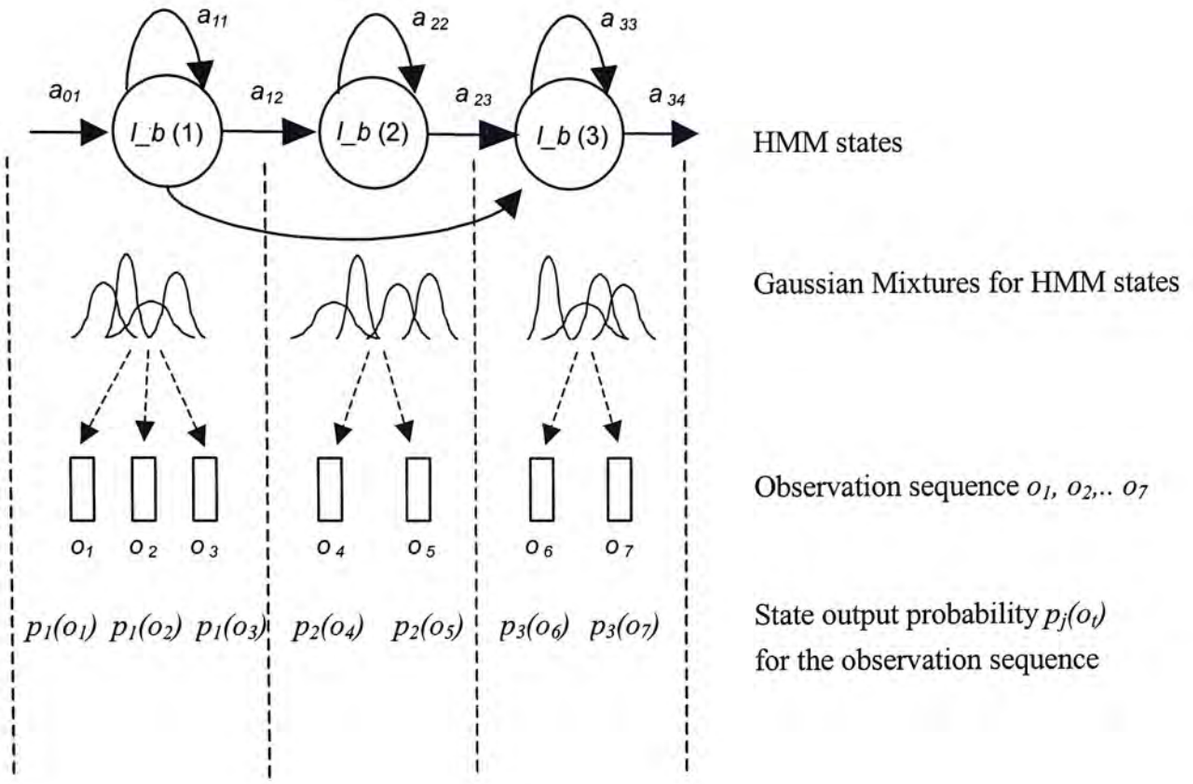


Figure 3.2: An example of HMM.

An HMM is fully specified when the state transition probabilities a_{ij} and the means μ_{jm}, co-variance matrix Σ_{jm} and mixture weights w_{jm} are given. These parameters are determined in the training process using the Baum-Welch re-estimation algorithm (forward-backward algorithm) [4][5].

Figure 3.3 shows different levels of acoustic representation. Given a sequence of acoustic feature vectors as the training data, the individual vectors are first aligned to the states. The probability functions are estimated from the statistics of all the training vectors assigned to a particular state. Therefore, each state is associated with a number of Gaussain mixture component pdf's.

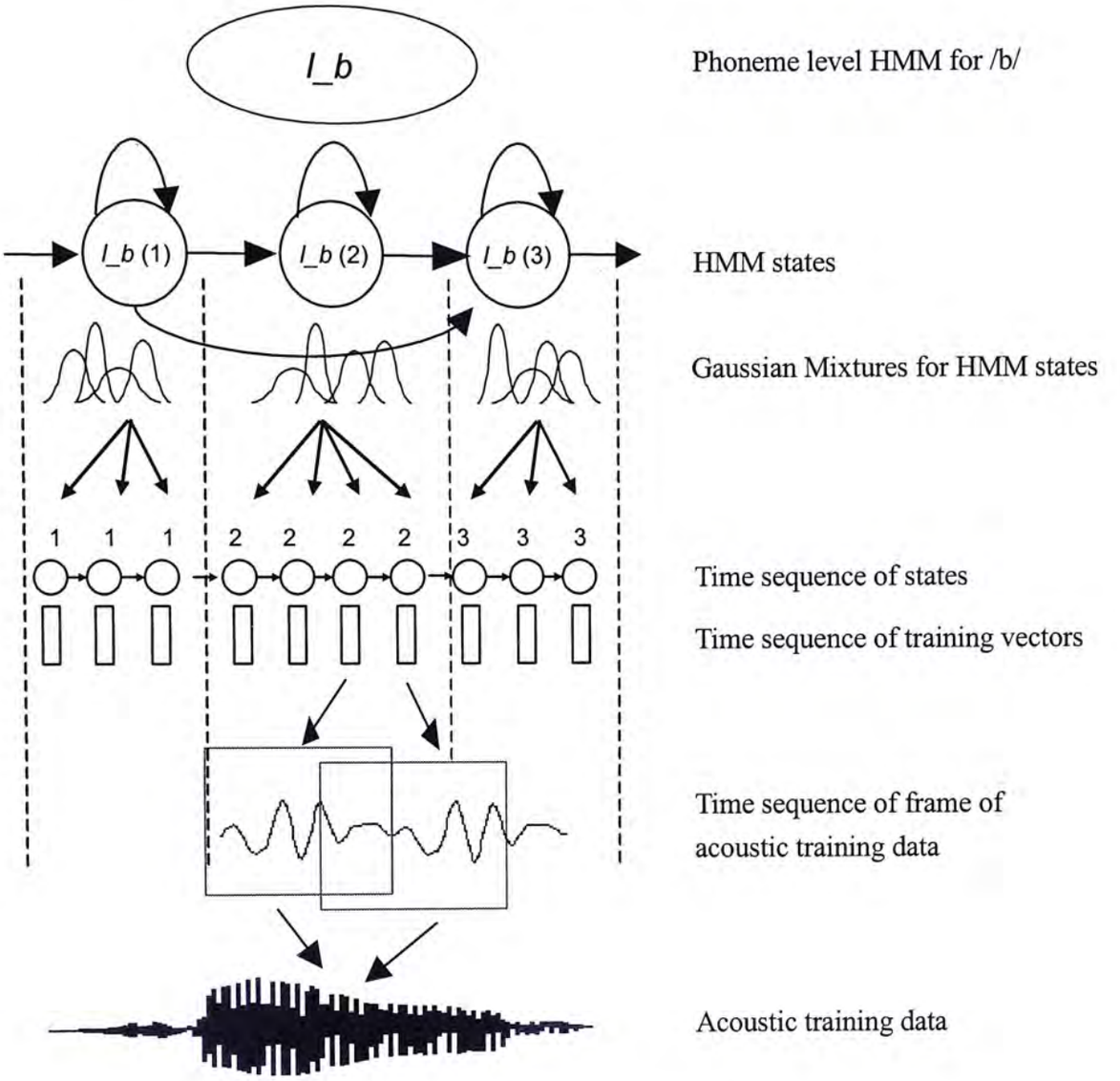


Figure 3.3: Different levels of acoustic representation.

Due to co-articulation, the acoustic realization of a phoneme is usually affected by its neighbors. Context-dependent acoustic model is often used. In this research, right context biphone models are employed. For example, the HMM *I_b* that models the Initial /b/ is expanded to a set of right context-dependent HMMs, which are denoted by *I_b+F_aa*, *I_b+F_an*, etc. The prefixes *I_* and *F_* denote Initial and Final models respectively. "+" denotes the connection of the base phone and the right context in a biphone model. In such a way, the total number of models would increase dramatically. To deal with this problem, decision-tree based state clustering approach [6] are used to allow sharing of model parameters among similar models.

3.4 Pronunciation Lexicon

Pronunciation lexicon essentially provides constraints on the combination of sub-word acoustic model to form a word model which describes the pronunciation of the word in terms of these sub-word units. It contains a baseform transcription for each word in the form of a sub-word sequence.

3.5 Statistical Language Model

A language model provides the constraints on how words can be concatenated together to form a sentence. Let $W = w_1, w_2, ..., w_n$ be a sequence of n words. The *a priori* probability of W is given by

$$P(W) = P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, ..., w_{i=1}) \qquad (3.5)$$

where $P(w_i|w_1,w_2,\ldots,w_{i-1})$ is the probability that the word w_i is preceded by the sequence $w_1,w_2,\ldots,w_{i-1}$. In reality, it is impossible to consider the entire word history. Instead, only a few preceding words are considered. This leads to the so-called n-gram language model, in which n-1 preceding words are considered. For example, bi-gram (n = 2) specifies the probability of w_{i-1} followed by w_i, i.e. $P(w_i|w_{i-1})$. This probability is usually obtained with a statistical approach from a large amount of training data. Let $c(w_{i-1},w_i)$ be the frequency count of the sequence (w_{i-1},w_i), i.e. w_{i-1} followed by w_i, and $c(w_{i-1})$ be the total count of w_{i-1} in the training data. The bi-gram probability can be computed by

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})} \qquad (3.6)$$

3.6 Decoding

The decoding process, also known as search process, finds an optimal sequence of words from a search space, which is formed by the pronunciation lexicon, the acoustic model and the language model. The search space is a compact structure that covers all legitimate word sequences. The algorithms for search are generally categorized as one-pass versus multi-pass search. In a one-pass search, all knowledge sources are used at a time to decode an utterance, whilst in a multi-pass search, knowledge sources are applied at different stages during decoding.

The forward Viterbi search is commonly employed [7]-[9]. Viterbi algorithm is a time-synchronous search, which employs dynamic programming technique to process all possible paths at the same time and to keep only the one with maximum score. The score is the cumulative probability density of the observations given by the HMMs and the score given by the LM.

3.7 The Baseline Cantonese LVCSR System

In this research, the effectiveness of different pronunciation modeling strategies will be compared based on a Cantonese continuous speech recognition system. In addition to the LVCSR task, the pronunciation models are also evaluated in a domain-specific task (stock domain). Details of the baseline system and the recognition tasks are given in the following sections.

3.7.1 System Architecture

A Cantonese LVCSR system consists of the components depicted in Figure 3.4. In our work, MFCC is used as the acoustic features. Each speech frame is represented by a 39 dimensional feature vector with 12 MFCCs and the speech energy of the frame, as well as their first and second order derivatives. The analysis window is 25 ms with 10 ms frame shift.

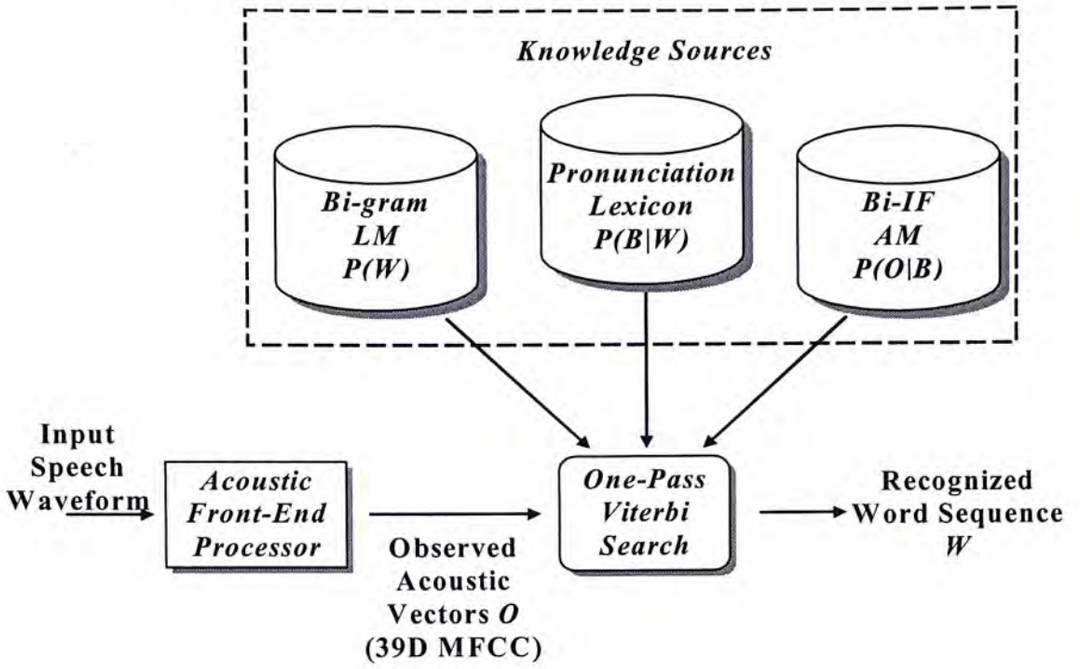


Figure 3.4: Block diagram of Cantonese LVCSR.

For Cantonese LVCSR, right context-dependent Initials and Finals are usually used as the basic units for acoustic modeling using HMM. In this research, the acoustic model being used is a set of cross-word bi-IF HMMs trained with 20 hours of continuous speech from the CUSENT corpus, which will be described in detail in the following section. The number of states for modeling Cantonese Initial and Final are three and five respectively. The number of Gaussian mixture components for each state is 16.

The pronunciation lexicon contains Chinese word entries with the corresponding baseform IF sequence transcription. In the LVSCR task, the lexicon consists of about 6,500 entries in which about 60% are poly-character words and the others are single-character words [6]. The baseform pronunciation for each entry is obtained from CUDICT [10]. In stock domain task, the lexicon contains 1,147 words.

The language model used is a word bi-gram using the 6,500-word lexicon mentioned above. The training corpus for N-gram modeling was complied from five Hong Kong newspapers consisting of about 98 million characters. For the stock domain task, the language model used is a word bi-gram train from 2095 stock queries.

We use a one-pass decoder for continuous Cantonese speech recognition [11]. The search space is a tree-structured lexicon constructed based on the baseform lexicon. The search algorithm is forward Viterbi search. A word lattice is resulted from this Viterbi search. When the utterance end is reached, the most probable word sequence is obtained by back-tracing the best path.

3.7.2 Speech Databases

The training speech data we used is the CUSENT corpus developed by the Chinese University of Hong Kong [12]. CUSENT is a read speech corpus of continuous Cantonese sentences, which is designed to be rich in phonetic context. A semi-automatic process was adopted in the creation of the sentence corpus. Chinese sentences are selected from four local newspapers of Hong Kong. The selection ensures that the coverage of intra-syllable (onset-nucleus) and inter-syllable (coda-onset) contexts is adequate and balanced. The corpus includes 5,100 training sentences and 600 testing sentences. The sentences were uttered by speakers of both genders. Table 3.1 gives a summary of CUSENT.

	Training Set	**Testing Set**
No. of Speakers	68	12
No. of Utterances	20 K	1.2 K
No. of Syllables	292.8 K	11.7 K
Total Length (hours)	20	1.1
Average Sentence Length (No. of Syllables)	10.5	9.7

Table 3.1: Statistics of the CUSENT corpus.

The training set of CUSENT is used to train both the acoustic model and the pronunciation model. The pronunciation modeling algorithms are tested in the LVCSR task using CUTEST, which is the test set of CUSENT. CUSENT was not designed specifically for pronunciation modeling. Being a read-speech corpus, it may not contain much variation information. Nevertheless, with its rich phonetic

coverage, commonly occurred variations in read speech are expected to be included. The proposed methods are also evaluated in a domain-specific application of Cantonese ASR that deals with 1300 utterances of spoken queries on stock information recorded from 13 speakers, named as STOCKTEST.

3.8 Summary

In this chapter, we reviewed a statistical approach for LVCSR. The details of a Cantonese LVCSR system were described.

An ASR system consists of three knowledge sources, namely, acoustic model, pronunciation lexicon and language model. Acoustic model gives the probability of the acoustic features being observed when the sub-word sequence is given. Pronunciation lexicon provides constraints on the combination of sub-word units forming a word. Language model gives the probability of a word sequence.

A decoder applies the Viterbi algorithm to find an optimal sequence of words from the search space formed by the knowledge sources.

Reference

[1] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.29, no.2, pp.254-272, 1981.

[2] B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", *J. Acoust. Soc. Amer.*, Vol.55, no.6, pp.1304-1312, 1974.

[3] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", *J. Acoust. Soc. Amer.*, Vol.87, no.4, pp.1738-1752, 1990.

[4] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in *Proceedings of the IEEE*, Vol.77, no.2, pp.257 –286,1989.

[5] L.E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes", *Inequalities*, Vol.1, pp.1-8, 1972.

[6] Y.W. Wong, "Large Vocabulary Continuous Speech Recognition for Cantonese", *M.Phil. Thesis*, The Chinese University of Hong Kong, 2000.

[7] A.J. Viterbi, "Error Bounds for Convolutional Codes and Asymptotically Optimal Decoding Algorithm", *IEEE Trans. on Information Theory*, Vol.13, pp.260-269, 1967.

[8] K.F. Lee *et al*, "Large Vocabulary Speaker-Independent Continuous Speech Recognition using HMM," in *Proceedings of the ICASSP-88*, Vol.1, pp.123-126, 1988.

[9] H. Ney, "Dynamic Programming Parsing for Context-Free Grammars in Continuous Speech Recognition", *IEEE Trans. on Signal Processing*, Vol.39, no.2, pp.336-341, 1991.

[10] CUPDICT: Cantonese Pronunciation Dictionary (Electronic Version), http://dsp.ee.cuhk.edu.hk/speech/page/corpus/Documents/culex.pdf, Dept. of Electronic Engineering, The Chinese University of Hong Kong, 2003.

[11] W.N. Choi, "An Efficient Decoding Method for Continuous Speech Recognition Based on a Tree-Structured Lexicon", *M.Phil. Thesis,* The Chinese University of Hong Kong, 2001.

[12] W.K. Lo *et al*, "Development of Cantonese Spoken Language Corpora For Speech Applications", in *Proceedings of ISCSLP-98*, pp.102-107, Singapore, 1998.

Chapter 4

Pronunciation Model

Pronunciation modeling in automatic speech recognition (ASR) is aimed at providing a mechanism by which the recognition systems can be adapted to pronunciation variation. This is done with a descriptive or predictive pronunciation model (PM) from which surfaceform pronunciations can be derived from the baseform pronunciation and its phonetic context. The basis for establishing a PM is the information about pronunciation variation. Generally speaking, the approaches can be divided into two categories: knowledge-based and data-driven [1].

In the knowledge-based approach, pronunciation variation is derived from linguistic knowledge. Based on linguistic studies or enumerated information dictionaries, certain rules for pronunciation variation are formulated. These rules typically concern deletions, insertions and substitutions of phonemes. The effectiveness of knowledge-based approaches depends on whether there exist appropriate linguistic references that fit the intended use. As a matter of fact, linguistic studies have a completely different perspective and focus from engineering applications. The pronunciation variation information thus obtained is usually inadequate to describe what is happening in real speech data.

In the data-driven approach, the information about pronunciation variation is obtained from speech data. By contrasting the actual realization of these data to their baseform pronunciations, pronunciation variants are observed. Reliable labeling of acoustic realization can be attained by human inspection of spectrographic display. However, this is very time-consuming and costly, especially because the amount of required data is usually large. In practice, the acoustic realizations are obtained automatically using speech recognition techniques. In comparison with manually labeled data, automatically recognized transcriptions are more appropriate for

pronunciation modeling because they are based on the same acoustic evidence as the ASR system [2]. Riley *et al* proposed to use a hand-labeled corpus as a bootstrap to establish a set of rules, which provide constraints for subsequent automatic transcription [2].

This research adopts the data-driven approach. PM is developed based on a large corpus, namely CUSENT, which contains 20 hours of continuous speech collected at the Chinese University of Hong Kong (CUHK) [3]. Two different PMs are used: IF confusion matrix (CM) and decision tree pronunciation model (DTPM). CM is a context-independent PM that predicts surfaceform from only the baseform. DTPM is a context-dependent PM which makes prediction of surfaceform depending on the phonetic context of the baseform.

4.1 Pronunciation Modeling at Different Levels

Pronunciation modeling can be done at different levels, for example, word level, phone level and state level [4].

Word-level pronunciation model **(WLPM)** specifies the probability that a word is pronounced as a particular surfaceform. It can be built by observing the realization of each word directly from the PM training data. However, many word entries in the lexicon may not be covered in the training data. In most cases, WLPM is built from phone-level pronunciation model (PLPM), where alternative pronunciations of a word can be obtained by replacing a phoneme in the word by a surfaceform phoneme.

Phone-level pronunciation model **(PLPM)** gives the probability that a baseform phoneme sequence is pronounced as a surfaceform phoneme sequence. This is equivalent to find a set of the possible variants for a particular phoneme (IF) unit. It provides a probabilistic description of the mapping between baseform phonemes (IF) and surfaceform phonemes (IF).

PLPM can be obtained by aligning the baseform transcription with the surfaceform transcription of a training corpus and computing the frequency of occurrences for each surfaceform. As a result, a probabilistic confusion matrix is obtained [5]. Alternatively, decision-tree based approach can be used to build a context-dependent PLPM, which is able to make prediction of surfaceform phonemes (IF) given the baseform phoneme (IF) and its context [4][6]. These two types of PLPMs will be discussed in the later section.

For HMM based acoustic model, it is also possible to develop the PM at a sub-phonetic level, i.e. HMM state level. ***State-level pronunciation model*** **(SLPM)** gives the probability that a baseform state sequence is pronounced as the surfaceform state sequence.

SLPM can be obtained by aligning baseform state sequence with surfaceform state sequence and computing the frequency of occurrences for each surfaceform state. SLPM can be used to modify the acoustic model in a way that the states of each sub-word HMM can either be adapted by the surfaceform states, or include the parameters of the surfaceform states. The SLPM has a finer resolution than PLPM. However, the number of parameters of SLPM would be significantly greater than that of the PLPM, obviously because of the additional information being included. This calls for more storage space and computation time. Also, the state sequence obtained in the recognition process is constrained by the acoustic model which is built at phone level. Thus we believe that PLPM and SLPM contain similar information. PLPM will be used throughout this research.

As stated in Chapter 3, speech recognition can be formulated as a probabilistic search process as follows

$$W^* = \arg\max_{W} P(W)P(O \mid B)P(B \mid W) \qquad (4.1)$$

where $P(B|W)$ is the probability that W is realized as the sub-word sequence B. If B is defined by the baseform pronunciation lexicon, $P(B|W)$ always equals to 1.0 because B is the only legitimate realization of W. By incorporating PLPM, the

probability $P(S_k|B)$ for the k-th pronunciation variant sequence S_k of B is introduced, and equation (4.1) is re-written as

$$W^* = \arg\max_W P(W)P(O \mid S_k)P(S_k \mid B)P(B \mid W) \qquad (4.2)$$

where $P(O|S_k)$ is the acoustic likelihood of S_k.

4.2 Phone-level pronunciation model and its Application

Two types of PLPM are discussed in this section: IF confusion matrix (CM) and decision tree pronunciation model (DTPM). CM is built directly by observing the realizations of all baseform phonemes in the training data. It makes a prediction of surfaceform from the baseform without considering its phonetic context. The training of DTPM applies an optimization process to cluster a set of phonemes according to their contextual information. Decision tree based prediction makes use of the phonetic context of the baseform. It is considered to be more precise than context-independent prediction as pronunciation variation is obviously affected by co-articulation.

4.2.1 IF Confusion Matrix (CM)

An IF confusion matrix (CM) characterizes the mapping between baseform IF and surfaceform IF, and for each surfaceform realization, specifies its probability. CM is obtained by the following procedures as illustrated in Figure 4.1:

1. The baseform transcription for the training corpus, for example CUSENT, is obtained from the baseform dictionary, which consists of standard Cantonese pronunciation of the words.

2. Surfaceform transcription is obtained from the output of phone (IF) recognition. The phone recognition is constrained such that the recognized output must be a sequence of I-F pairs. This constraint greatly enhances the recognition accuracy over unconstrained phone recognition. By aligning the recognized surfaceform with the baseform IF sequence, a phone recognition accuracy of 90.33% is observed.

3. For each utterance, the surfaceform transcription is aligned with its baseform transcription using dynamic programming. The confused pairs of baseform and surfaceform IF units are identified.

4. For a particular baseform IF unit b and surfaceform IF unit s_k, the total number of times that b is confused with s_k is counted and denoted by $C(b \rightarrow s_k)$. Then the variation probability (VP), $P(s_k|b)$, is estimated as

$$P(s_k \mid b) = \frac{C(b \rightarrow s_k)}{\sum_k C(b \rightarrow s_k)} \qquad (4.3)$$

5. A threshold is set to prune those less frequent surfaceform pronunciations in order to assure the augmented lexicon does not contain irrelevant pronunciations. Such odd events are probably due to recognition errors. Indeed, unconstrained phone recognition is known to be fairly erroneous. On the other hand, including all variations would increase the homophone rate and cause severe confusion in recognition. Adding these pronunciations will deteriorate the recognition performance. The threshold can be set in terms of either absolute count or variation probability.

Typically, for each baseform IF unit, a number of possible surfaceform units are found. Table 4.1 shows part of the CM in the form of a table for the Initial units /m/ and /ng/, and the Final units /o/ and /un/. Both /m/ and /ng/ have one alternative variation, whilst /o/ and /un/ do not. The full matrix obtained from CUSENT is shown in Appendix III.

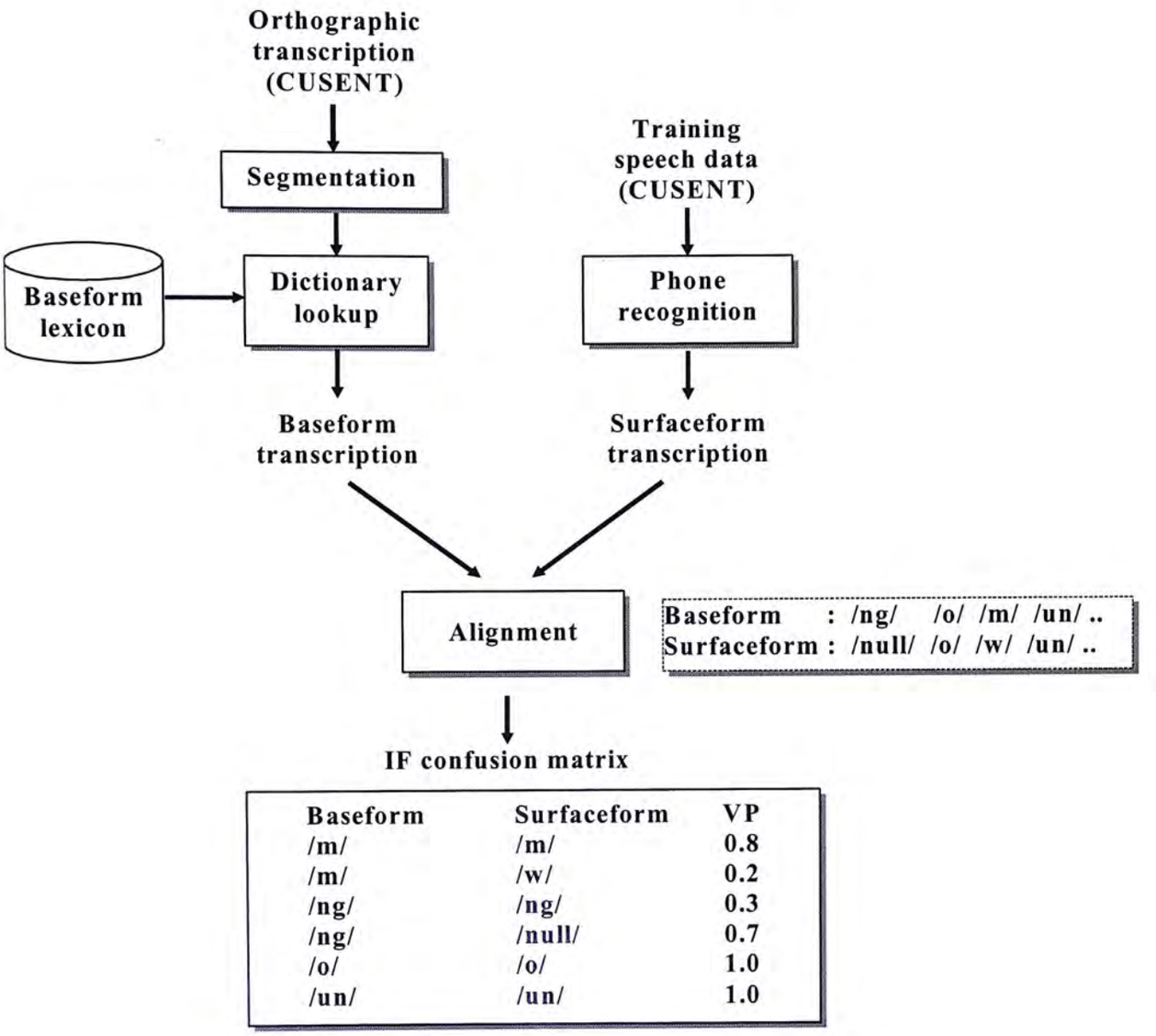


Figure 4.1: Construction of CM.

Baseform *b*	Surfaceform *s*	Variation Probability (VP)
/m/	/m/	0.8
/m/	/w/	0.2
/ng/	/ng/	0.3
/ng/	/null/	0.7
/o/	/o/	1.0
/un/	/un/	1.0

Table 4.1: CM in table form for Initial /m/ and /ng/, and Final /o/ and /un/ with the corresponding variation probabilities.

Both the baseform and the surfaceform representations use the same set of phonemic units, which are modeled by the same set of acoustic models. Replacing a baseform IF by a surfaceform IF is equivalent to using another (surfaceform) IF model to produce the acoustic score. If a speaker pronounces the phoneme with variation, this surfaceform acoustic score is supposed to be higher than the baseform acoustic score. In this way, we provide another path having a higher acoustic score in the search process.

4.2.2 Decision Tree Pronunciation Model (DTPM)

Decision tree pronunciation model (DTPM) is essentially a context-dependent PM used to predict the surfaceform IFs given the baseform IF. As shown in Figure 4.2, a decision tree contains many nodes that are organized in a hierarchical way. Each node in the decision tree is featured by a binary question (yes/no answer) about the phonetic features regarding the context of the baseform. The leaves of the tree illustrate the best predictions (surfaceform IFs) based on the training data.

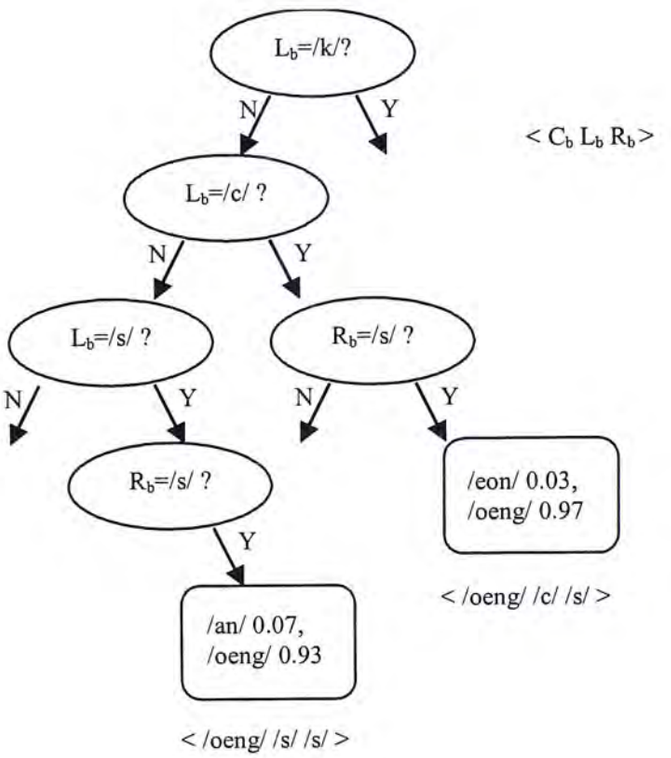


Figure 4.2: Decision tree based prediction of pronunciation variation for the Final /oeng/.

The training process of a DTPM is shown as in Figure 4.3. The baseform IF transcription together with the surfaceform transcription obtained from the phone recognition form a set of training vectors with phonetic context. The training applies an optimization process to cluster a set of phoneme with contextual information [7]. A set of questions about the phonetic context are designed. When the lexical tree grows, all possible questions are tried at each node to split the data. The question that generates the best partitions is selected. The best question is the one that minimizes the total conditional entropy of the surfaceform realizations of a phoneme given its phonetic context. This process is applied recursively on each branch, until the stopping criterion is met. The stopping criterion requires a minimal number of samples on the parent node and child node.

One decision tree is built for each Initial and Final. Therefore, a total of 73 decision trees are needed for Cantonese.

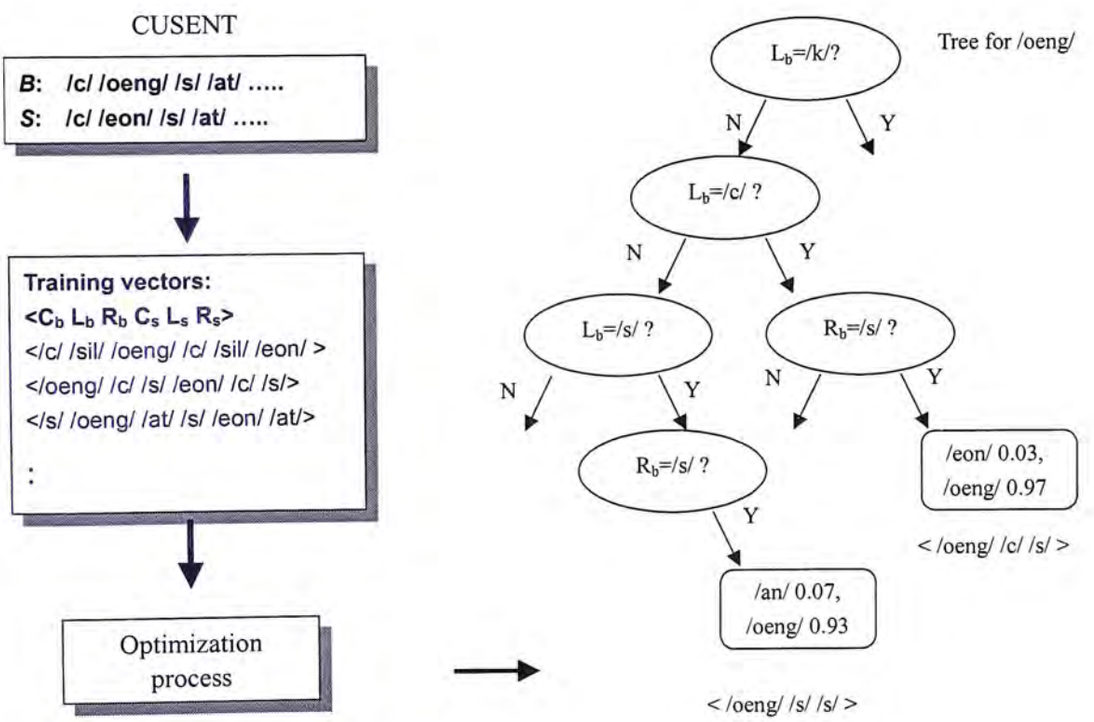


Figure 4.3: Construction of decision tree.

In our research, different sets of decision trees are designed for different purposes. ***Context-dependent decision tree (CDDT)*** concerns the left and right phonetic context. ***Left context-dependent decision tree (LCDDT)*** concerns only the left phonetic context of the baseform and surfaceform.

As the amount of training data is limited, the number of training samples retained at a leaf node may be very small. The prediction in each leaf may correspond to only some rarely occurred training samples. Therefore, we divide the phone set into classes according to their phonetic features and design a set of questions concerning these features for building the tree, termed as ***phonetic class decision tree (PCDT)***. We believe that phones belonging to the same class will have similar effect on their neighborings. The question set is listed in Appendix IV. ***Left phonetic class decision tree (LPCDT)*** considers only the phonetic features of the preceding phoneme. Appendix V shows a complete CDDT and PCDT for /aang/ obtained from CUSENT.

Let the baseform and surfaceform units under consideration be denoted as C_b and C_s. Let L_b and L_s be the left baseform context and the left surfaceform context respectively, and R_b and R_s be the right baseform context and the right surfaceform context respectively. DTPM can be used to predict the surfaceform IF (C_s) given the baseform IF (C_b) and the phonetic context (L_b, R_b). For prediction, the baseform IF together with its left and right context form a vector <C_b L_b R_b>. An example is shown in Figure 4.4. The input vector representing the baseform unit /oeng/ is </oeng/ /c/ /s/>. This vector is then processed by the respective decision tree to obtain the predicted pronunciation variants and the corresponding VPs.

CDDT and PCDT can be used to refine the CM [6][8]. This will be discussed in the following section. On the other hand, LCDDT and LPCDT can be incorporated in the search process to make online surfaceform prediction for baseform phoneme [8] since the right context for an IF model in the search space is not known in the forward Viterbi search.

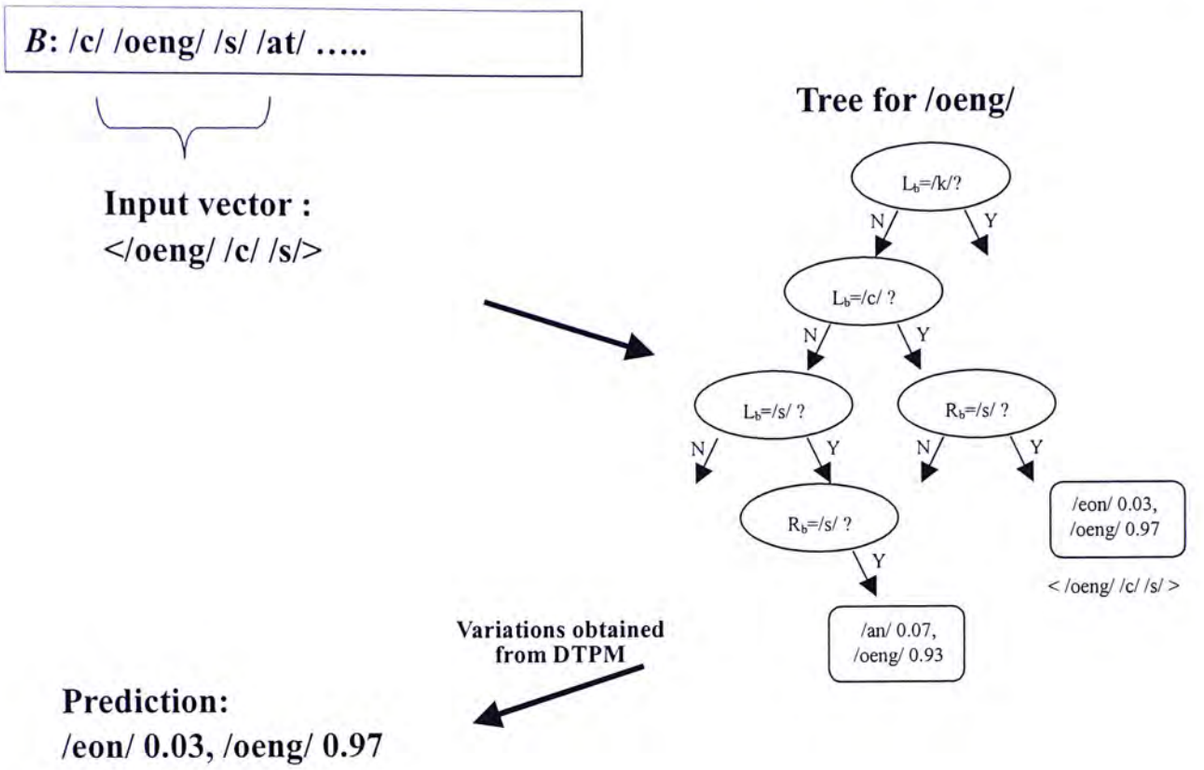


Figure 4.4: Prediction of variations using DTPM.

4.2.3 Refinement of Confusion Matrix

As stated in Section 4.2.1, CM can be obtained from the alignment between baseform transcription and phone recognition output. The phone recognition output may contain errors. In this section we describe a method to obtain more accurate surfaceforms by using CDDT or PCDT [6][8]. The refined surfaceforms are used to modify the CM. The method can be summarized by the following procedures as illustrated in Figure 4.5.

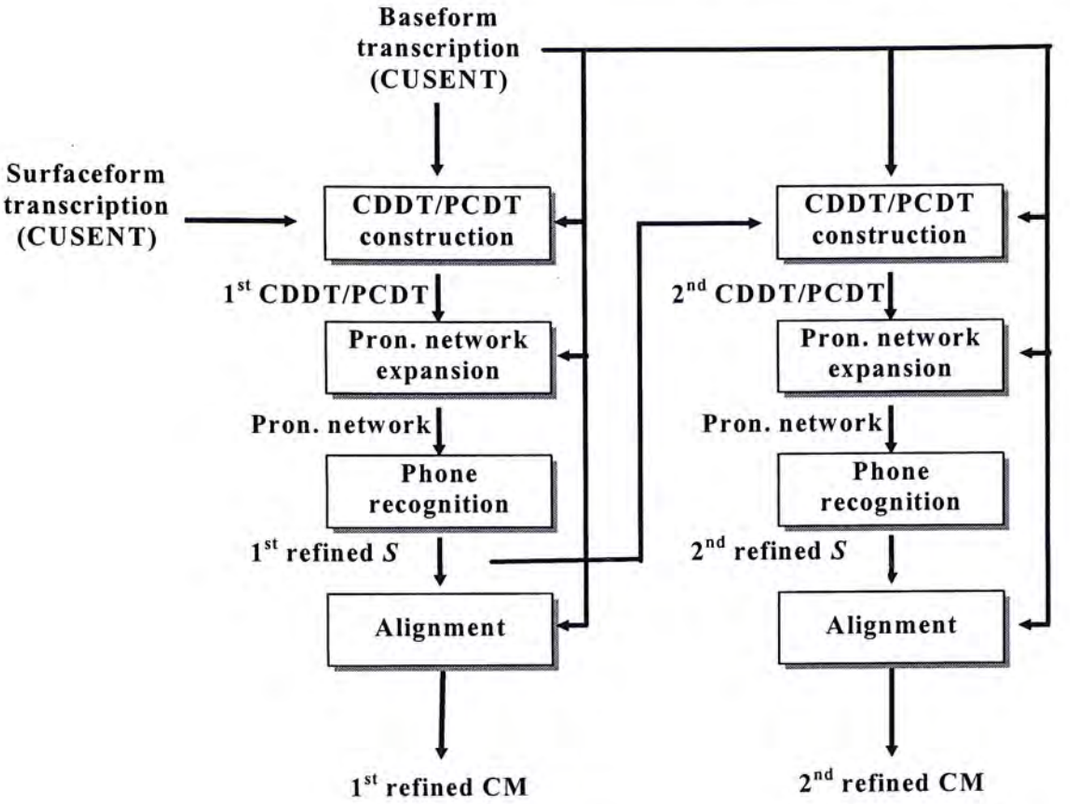


Figure 4.5: CM refinement by DTPM.

1. Baseform transcription is obtained from the baseform dictionary. Surfaceform transcription is obtained by automatic phone recognition.

2. The baseform and surfaceform transcriptions are used as training data to construct the first set of DTPM, denoted as 1st CDDT/PCDT.

3. DTPM is applied to the baseform transcription to obtain a pronunciation network with pronunciation alternatives.

4. Constrained phone recognition is performed on the same data using the pronunciation network. The result is a new surfaceform transcription, named as 1st refined *S*. The refined surfaceform attains a phone accuracy of 92.15%, as compared with 90.33% in the unrefined case.

5. Refined surfaceform and the original baseform transcription are used to establish a refined CM, named as 1st refined CM.

6. The alignment between the baseform and the refined surfaceform transcription over the training data are used to train another set of DTPM, named as 2nd CDDT/PCDT. Repeating step 2 to step 5, 2nd refined CM can be obtained in the same way except the surfaceform pronunciation are chosen from the alternatives generated by a new set of DTPM.

4.3 Summary

In this chapter, we introduce the role of PM in ASR, i.e. providing a mechanism by which ASR can be adapted to pronunciation variability. The construction and uses of different types of PMs are discussed.

We have adopted a data-driven approach to obtain pronunciation variation from a set of Cantonese speech data. Two types of phone-level pronunciation model (PLPM) are constructed: IF confusion matrix (CM) and decision tree pronunciation model (DTPM).

CM is a context-independent PLPM which gives the probability that a baseform phoneme is realized as a surfaceform phoneme. CM can be refined by using DTPM. DTPM is essentially a context-dependent PLPM used to predict the surfaceform phonemes (IF) given the baseform phoneme (IF) and its phonetic context. Context-dependent decision tree (CDDT) concerns the left and right phonetic context. Left context-dependent decision tree (LCDDT) concerns only the left phonetic context. If the phone set is divided into classes according to their phonetic features and questions are designed based on the phonetic features, the trees are termed as phonetic class decision tree (PCDT). Left phonetic class decision tree (LPCDT) considers only the phonetic features of the preceding phoneme.

References

[1] H. Strick *et al*, "Modeling Pronunciation Variation for ASR: A Survey of the Literature", *Speech Communication*, Vol.29, pp.225-246, 1999.

[2] M. Riley *et al*, "Stochastic Pronunciation Modeling from Hand-labeled Phonetic Corpora", *Speech Communication*, Vol.29, pp. 209-224, 1999.

[3] W.K. Lo *et al*, "Development of Cantonese Spoken Language Corpora For Speech Applications", in *Proceedings of ISCSLP-98*, pp.102-107, Singapore, 1998.

[4] Y. Liu, "Pronunciation Modeling for Spontaneous Mandarin Speech Recognition", *Ph.D. Thesis*, The Hong Kong University of Science and Technology, 2002.

[5] M.K. Liu *et al*, "Mandarin Accent Adaptation Based on Context-Independent/Context-Dependent Pronunciation Modeling", in *Proceedings of ICASSP-00*, Vol.2, pp.1025-1028, Istanbul, 2000.

[6] W. Byrne *et al*. "Pronunciation Modeling Using a Hand-labeled Corpus for Conversational Speech Recognition", in *Proceedings of ICASSP*-98, Vol.1, pp.12-15, Seattle, 1998.

[7] http://festvox.org/docs/speech_tools-1.2.0/x3475.htm

[8] P. Kam *et al*, "Modeling Pronunciation Variation for Cantonese Speech Recognition", in *Proceedings of PMLA-02*, pp.12-17, Denver, 2002.

Chapter 5

Pronunciation Modeling at Lexical Level

This chapter will be focused on the incorporation of pronunciation model (PM) into the pronunciation lexicon to deal with phone change. This is accomplished by augmenting the standard baseform lexicon with additional pronunciation variants to construct a pronunciation variation dictionary (PVD).

The pronunciation lexicon in our Cantonese LVCSR system defines how a word is formed by Initial (I) and Final (F) units. This is essentially to provide constraints on the combination of IF units. Conventionally, the lexicon includes only the baseform transcription for each word. If a phone change occurs in the pronunciation of a word, the baseform transcription will no longer reflect the actual pronunciation [1]. If such alternative pronunciations are not included in the lexicon, the correct word can never by recognized because the respective surfaceform IF sequence is not allowed in the search space [1]-[6]. Instead, another word that is acoustically similar to that surfaceform pronunciation will probably be retrieved.

For example, the baseform transcription for the Chinese word 我們 is {/ng/ /o/ /m/ /un/}. If /ng/ can be completely realized as /null/ and /m/ can be realized as /w/, then the surfaceform transcription {/null/ /o/ /m/ /un/}, {/ng/ /o/ /w/ /un/} and {/null/ /o/ /w/ /un/} should be added into the lexicon. In this way, four paths are allowed to represent 我們 in the search space.

The augmented lexicon PVD is a word-level pronunciation model (WLPM) which gives the probability $P(S_{W,k}|W)$ that a word sequence W is being pronounced

as the k-th surfaceform pronunciation, $S_{W,k}$ [6]. If multiple pronunciations of words are included in the lexicon, $P(B|W)$ in the conventional search equation given in equation (4.1) is replaced by $P(S_{W,k}|W)$. Thus we have

$$W^* = \arg\max_W P(W)P(O \mid S_{W,k})P(S_{W,k} \mid W) \qquad (5.1)$$

where $P(O|S_{W,k})$ is obtained from the acoustic model.

In our research, phone-level pronunciation model (PLPM) is used to build the PVD to handle pronunciation variations. Alternative pronunciations of a word can be obtained by replacing a phoneme in the word by a surfaceform phoneme.

5.1 Construction of PVD

PVD is an augmented lexicon that includes alternative pronunciations of words. To build a PVD, we have to decide what the variants are and which of them should be included in the lexicon. We adopt the data-driven approach as described in Chapter 4 originated by M.K. Liu *et al* [2] to derive the pronunciation variants from a PLPM. Specifically, we use IF confusion matrix (CM) and refined IF confusion matrix (refined CM), which give the possible variants with the corresponding VPs of a particular baseform IF unit as discussed in Section 4.2.1 and 4.2.3.

The entire process of PVD construction is shown as in Figure 5.1. "PVD" means the pronunciation variation dictionary built directly using a CM. "1st PVD" and "2nd PVD" are the pronunciation variation dictionaries built by the 1st and 2nd refined CM.

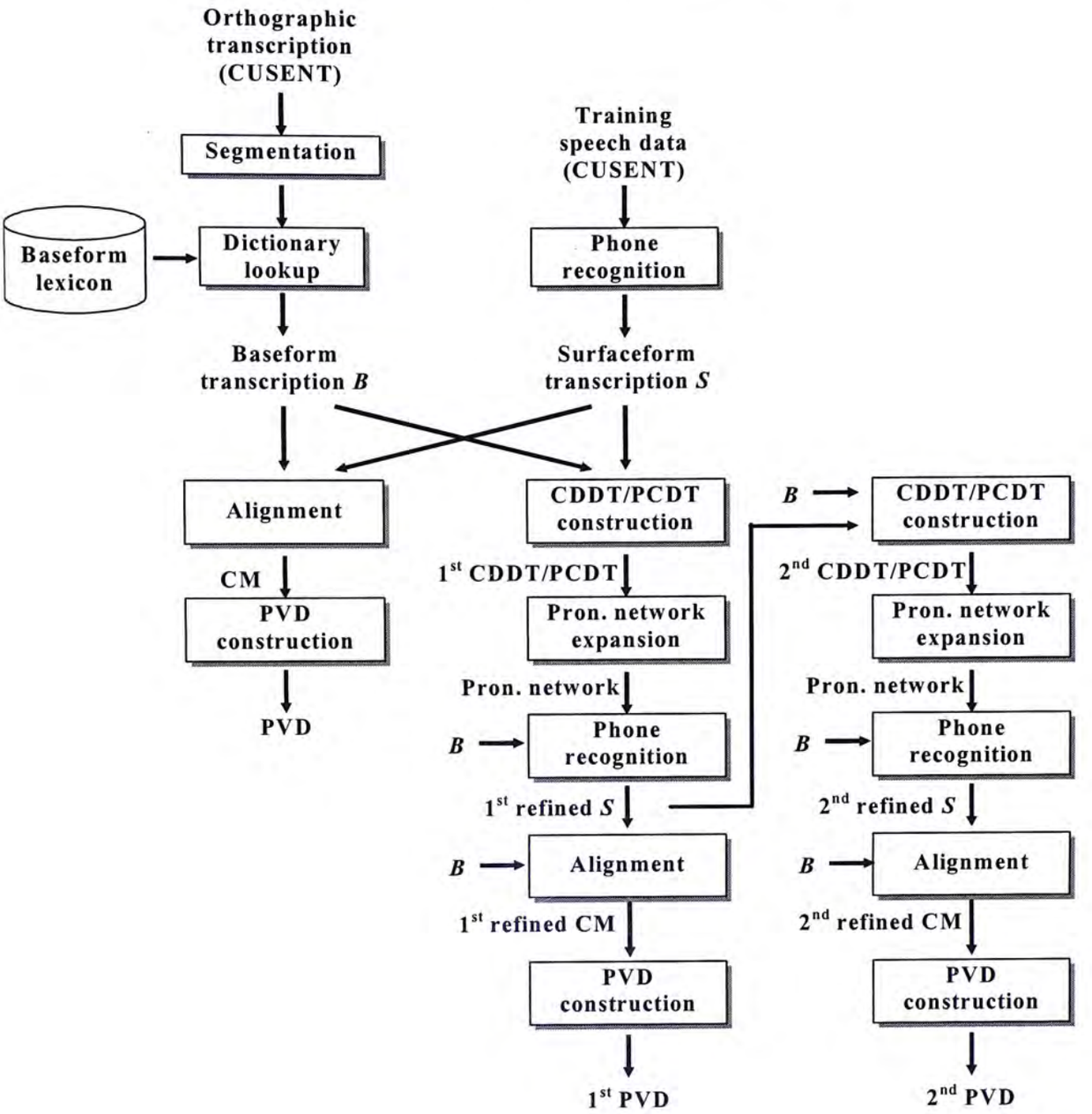


Figure 5.1: PVD construction by CM and refined CM.

PVD contains the surfaceform transcriptions $s_{w_i,k}$ for the i-th word w_i with the corresponding word variation probabilities $P(s_{w_i,k}|w_i)$. This probability is obtained by multiplying the variation probabilities (VPs) of all individual surfaceform IFs composing the word given by equation (5.2).

$$\begin{aligned} P(s_{w_i,k} \mid w_i) &= P(s_{w_i,k} = s_{k1}s_{k2}..s_{kN} \mid b_{w_i} = b_1 b_2..b_N)P(b_{w_i} \mid w_i) \\ &= \prod_{n=1}^{N} P(s_{kn} \mid b_n)P(b_{w_i} \mid w_i) \end{aligned} \qquad (5.2)$$

where $b_1\ b_2 \ldots b_N$ and $s_{k1}\ s_{k2} \ldots s_{kN}$ are the baseform and k-th surfaceform IF sequences for the word w_i respectively. $P(b_{w_i}|w_i)$, the probability of the baseform IF sequence given w_i, is always equal to one. $P(s_{kn}|b_n)$ is the VP for the n-th surfaceform IF in the transcription.

Table 5.1 shows part of the PVD with the word w_i be 我們. The baseform IF sequence is {/ng/ /o/ /m/ /un/}. From Table 4.1 in Chapter 4, we have $P(/null/|/ng/)$ = 0.7, $P(/o/|/o/)$ = 1 $P(/m/|/m/)$ = 0.8, and $P(/un/|/un/)$ = 1. Then, by equation (5.2), $P(\{/\text{null}/\ /\text{o}/\ /\text{m}/\ /\text{un}/\}|\text{我們})$ = 0.7 x 1 x 0.8 x 1 = 0.56.

w_i	b_{w_i}	$s_{w_i,k}$	$P(s_{w_i,k}\|w_i)$
我們	/ng/ /o/ /m/ /un/	/ng/ /o/ /m/ /un/	0.24
		/null/ /o/ /m/ /un/	0.56
		/ng/ /o/ /w/ /un/	0.06
		/null/ /o/ /w/ /un/	0.14

Table 5.1: The word 我們 with its surfaceforms and word variation probabilities.

The probability $P(S_{W,k}|W)$ is composed of the word variation probabilities $P(s_{w_i,k}|w_i)$ of each word in the word sequence by equation (5.3). This probability can be used in the decoding process to find a particular pronunciation variant that maximizes the probability $P(W|O)$.

$$P(S_{W,k}\mid W) = \prod_i P(s_{w_i,k}\mid w_i) \qquad (5.3)$$

5.2 PVD Pruning by Word Unigram

Word frequency is an important factor to be considered in building the PVD. In Chinese, a longer word generally has a smaller chance to occur. Many researchers have proved that words with a small unigram tend to have fewer variations in their pronunciations [7][8]. However, in our approach of building the PVD, a larger number of variants tend to be introduced for long words. This is because these

variants are obtained by phoneme substitution and long words consist of more phonemes.

In order to solve this problem, we use the word unigram to control the number of alternative pronunciations to be added into the baseform lexicon. The higher the unigram, the more the alternatives should be added. The procedures are as follows:

1. PVD is built using the method above by CM. However, the threshold is set looser in order to include more variations at the beginning.

2. The word alternatives in the PVD are ranked by $P(s_{w_i,k}|w_i)$.

3. Word unigrams for all words in the PVD are found from the LM.

4. The word unigram is used to scale the number of variations N_{w_i} for each word. The N_{w_i} variations with the highest word variation probabilities $P(s_{w_i,k}|w_i)$ will be included in the PVD. The maximum number of variations for each word was limited to 10. The word unigram ranges from 0 to 0.04. There are 97.6% of the words with unigram between 0 and 0.025. The word unigram together with a scaling factor of 4000 limits the range of N_{w_i} from 0 to 10.

5.3 Recognition Experiments

5.3.1 Experiment 1 — Pronunciation Modeling in LVCSR

In Experiment 1, the effectiveness of using different PVDs is evaluated in Cantonese LVCSR. The factors affecting the performance of PVD will be investigated.

Experimental Conditions:

As described in Section 3.7, the testing data is from CUTEST, which contains 1200 sentences (about 1.1 hours) recorded from 6 male speakers and 6 female speakers. 39-dimensional MFCC feature vectors are used. The acoustic model is a set of cross-word bi-IFs. The search engine is a one-pass decoder based on tree-structure lexicon [9]. PM is trained with the 20 hours CUSENT corpus.

Experimental Results:

(1) The use of PVDs with different VP thresholds

Table 5.2 shows the recognition results with PVDs that adopt different VP thresholds (VP Th) ranging from 0.02 to 0.2. "WER" stands for word error rate. "No. of IF variants" is the total number of IF variants including the originally 73 IFs. "PVD size" is the total number of entries in the PVD. It is found that the use of PVD achieves a better performance of recognition. The extent of improvement varies with VP Th. If the threshold is too stringent, i.e. VP Th is very high, many frequently pronounced variations may not be included in the PVD. If VP Th is small, a large number of variants would be included. As a result, the search space is enlarged and more ambiguities are introduced to the searching process. A threshold of 0.05 appears to give the most significant improvement on the accuracy. In this case, the average number of pronunciation variants (including the baseform) per IF unit is 1.30 (95/73). The average number of variants per word is 1.33 (8568/6451).

	Baseline	VP Th 0.02	**VP Th 0.05**	VP Th 0.10	VP Th 0.15	VP Th 0.20
WER (%)	25.34	23.91	**23.49**	23.7	23.64	23.58
Relative WER Reduction (%)		5.64	**7.30**	6.47	6.71	6.95
No. of IF variants	73	129	**95**	82	79	78
PVD size	6451	20840	**8568**	7356	7210	7171

Table 5.2: WER(%) of LVCSR task using PVDs with different VP Th.

(2) The use of different PVDs with VP Th = 0.05

Table 5.3 shows the recognition results of using different PVDs with the same VP Th = 0.05. It is noted that all PVDs have similar performance. The purpose of iterative refining is to deal with possible phone recognition errors in order to obtain a refined surfaceform. From the results, refining the PVD with CDDT or PCDT seems not to be helpful. This may be due to the fact that the amount of data used to train the context-independent CM is quite large. The mis-recognized surfaceforms tend to be very diverse as opposed to the case with fewer training data. The VPs for these mis-recognized surfaceforms are much smaller than that of the baseform and the surfaceforms that are true pronunciation variations. Therefore, after pruning with the VP Th, the CM would no longer include those mis-recognized surfaceforms. In other words, most of the unreliable information caused by phone recognition error has already been removed.

On the other hand, the acoustic model being used is a set of context-dependent HMMs. This may explain why adding contextual information via decision tree does not help much.

It is found that PCDT does not perform better than CDDT. The choice of the additional paths in the expanded pronunciation network during phone recognition also depends on the acoustic model. Although CDDT may produce more unreliable paths than PCDT, as long as the true path is there, it will be retrieved by the decoder with the aids of AM. Therefore, dividing the phone set into classes according to their phonetic features for building the tree may not do any further help.

	Baseline	PVD	1st CDDT PVD	2nd CDDT PVD	1st PCDT PVD	2nd PCDT PVD
WER (%)	25.34	23.49	23.54	23.55	23.53	23.56
Relative WER Reduction (%)		7.30	7.10	7.06	7.14	7.02
No. of IF variants	73	95	93	93	93	93
PVD size	6451	8568	8358	8358	8358	8358

Table 5.3: WER(%) of LVCSR task using different PVDs with VP Th = 0.05.

Result Analysis for PVD with VP Th = 0.05:

To have a better understanding about how the PVD really affect the recognition performance, the recognition results of using the PVD with VP Th = 0.05 is compared with that of the baseline system. Table 5.4 shows the performance table of all the variants that were added to the lexicon. It reveals the relationship between the recognition performance and several important factors including the frequency of occurrences of baseform and surfaceform units, variation probability, lexical tree expansion factor and character level confusion. The followings are the definitions of the symbols used in this table.

b - Baseform IF unit
s - Surfaceform IF unit
I - Number of characters improved in character recognition
D - Number of characters degraded in character recognition
T - No change in character recognition performance
O_B - Occurrence count of *b* in testing data
O_S - Occurrence count of *s* in testing data
VP - Variation Probability
N_S - No. of nodes for *s* in the original baseform lexical tree
N_{expS} - No. of nodes for *s* in the expanded lexical tree
EF - Expansion factor = N_{expS}/N_S
C_S - No. of characters represented by *s* in original baseform lexicon
C_{expS} - No. of characters represented by *s* in expanded lexicon
X_S - No. of confusing characters represented by *s* in expanded lexicon

b	*s*	*I*	*D*	*T*	O_B	O_S	*VP*	*s*	N_S	N_{expS}	*EF*	C_S	C_{expS}	X_S
aak	aa	5	1	11	156	369	0.108	aa	118	219	1.86	83	134	104
aat	aa	8	1	17	176	369	0.099							
aak	aat	6	0	7	156	176	0.084	aat	50	101	2.02	21	51	33
aang	aan	1	1	4	20	412	0.205	aan	94	112	1.19	88	106	61
aang	an	1	2	2	20	383	0.063	an	166	252	1.52	120	167	136
ang	**an**	**17**	**1**	**52**	**191**	**383**	**0.271**							
aang	ang	0	1	1	20	191	0.076	ang	68	86	1.26	29	47	31
aap	ap	1	1	3	45	220	0.069	ap	50	75	1.50	20	43	27
ak	at	6	0	16	90	333	0.173	at	84	127	1.51	42	62	41
ek	e	0	0	6	33	82	0.115	e	56	68	1.21	32	43	29
eng	ing	0	0	2	46	378	0.099	ing	196	209	1.07	141	156	118
im	in	1	1	3	81	508	0.060	in	172	213	1.24	92	128	106
it	i	1	0	5	99	416	0.056	i	253	295	1.17	114	147	130
ng	**m**	**10**	**2**	**65**	**90**	**6**	**0.815**	m	1	11	11	1	9	9
ok	o	2	0	14	237	396	0.080	o	119	203	1.71	75	119	107
on	ong	5	1	12	97	419	0.075	ong	126	151	1.20	104	127	39
gw	**g**	**21**	**4**	**35**	**226**	**915**	**0.301**	g	189	245	1.30	208	245	97
kw	k	0	0	2	16	339	0.147	k	84	87	1.04	70	86	23
n	**l**	**70**	**13**	**210**	**302**	**735**	**0.768**	l	160	218	1.36	174	220	184
ng	**null**	**34**	**8**	**146**	**257**	**154**	**0.595**	null	15	62	4.13	21	82	38
null	ng	1	0	4	154	257	0.072	ng	47	62	1.32	61	82	38
Total		190	37	617					6601	7351				

Table 5.4: Performance table with VP Th = 0.05

The total number of surfaceform IFs being recognized is 844. It means that there are 844 times that the surfaceform models give higher scores than the baseform models. Out of the 844 cases, 190 cause improvement in character recognition, 37 cause degradation and 617 do not have any effect. The improvement is due to the fact that pronunciation variation can be represented by a more realistic surfaceform model. For example, if a person mis-pronounces the Chinese character 百 (hundred) as {/b/ /aat/} instead of {/b/ /aak/}, the surfaceform model /aat/ gives a higher acoustic score. The original baseform dictionary only contains the mapping between 百 and {/b/ /aak/}, 八 (eight) and {/b/ /aat/}. However, the PVD also contains the realization of 百 as {/b/ /aat/} Therefore, the decoder will be able to retrieve the correct word 百. The degradation is due to confusion. In the PVD, each word may have more pronunciations and consequently each pronunciation is now representing more characters. For example, {/b/ /aat/} represents both characters 百 and 八.

Factors Affecting the Performance of PM

From Table 5.4, we can observe that most of the added variations lead to performance improvement, in particular, /ang/→/an/, /ng/→/m/, /gw/→/g/, /n/→/l/ and /ng/→/null/. This is related to a number of factors: (1) occurrences of baseform and surfaceform IF, (2) variation probability, (3) lexical tree expansion factor, and (4) character level confusion

(1) Frequency of occurrences of baseform and surfaceform

The more frequent the occurrence of a baseform IF is, the more probably its surfaceform variants would be touched. For example, O_B for Initial /ng/ and Final /ng/ are much larger than O_S for Initial /null/ and Final /m/. Therefore, the number of improvement is large. On the other hand, higher occurrence of a surfaceform IF in testing data O_S increases the chance of confusion. For example, O_S for /g/ and /l/ are much larger than O_B for /gw/ and /n/, introducing a lot of confusion.

The occurrence of the baseform IF in testing data O_B is related to the word unigram of the word containing that IF. The larger the word unigram, the larger the

occurrence of the IF, the more the variation should be added. Therefore, word unigram can be used for PVD pruning.

(2) Variation probability (*VP*)

Being obtained from a large amount of acoustic data, variation probability reflects the likeliness of a realistic pronunciation. Adding the variations with large VPs makes it possible to handle frequently occurred variations. Therefore, this leads to improvement of recognition performance. Also, VP contributes to the path score in the search process. For example, even O_B is much smaller than O_S for /ang/→/an/ and /gw/→/g/ and /n/→/l/, they show a large number of improvement due to large VPs.

(3) Lexical tree expansion factor

Lexical tree expansion factor (*EF*) is the ratio of the count of an IF in expanded surfaceform lexical tree (N_{expS}) to the count in original baseform lexical tree (N_S). In general, large expansion factor will increase the confusability during decoding and cause degradation in recognition.

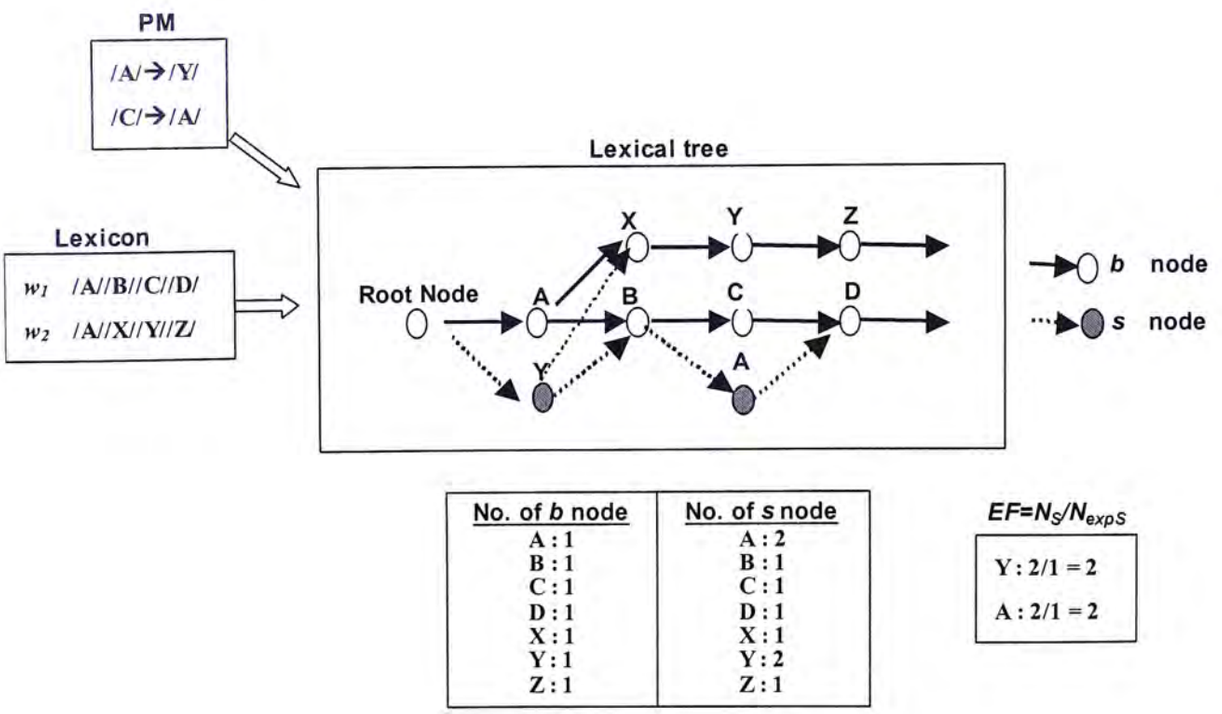


Figure 5.2: Calculation of lexical tree expansion factor.

Figure 5.2 shows a simple example that explains the expansion factor. Suppose the lexicon contains only two words, w_1 and w_2, with IF transcription /A/ /B/ /C/ /D/ and /A/ /X/ /Y/ /Z/. With the PM predicting /A/ to be /Y/ and /C/ to be /A/, the numbers of surfaceform nodes for both /Y/ and /A/ are increased by one. Originally, /Y/ only corresponds to one character. After the expansion, it represents 2 characters. If the speaker really utters the baseform /Y/, only one possible character can be chosen in the original search space. However, in the expanded lexical tree, the choice becomes more, the confusability is increased. For example, the expansion factor for /ng/ → /null/ is large (EF = 4.13), therefore, it causes a large number of degradation (D = 8).

(4) Character level confusion

A Chinese character consists of an Initial (I) and a Final (F). The introduction of a surfaceform I will give another representation for this character with the succeeding F. This character will confuse with the existing characters formed by the surfaceform I and F in the baseform lexicon. Similarly, the introduction of a surfaceform F will cause confusion to the existing character formed by the preceding I and the surfaceform F. Character level confusion considers the overlapping of the preceding or succeeding IF for the baseform and surfaceform. If the overlapping is large, the baseform IF and the surfaceform unit share the same set of preceding or succeeding units. Then, the introduction of s given b will cause a lot of confusions.

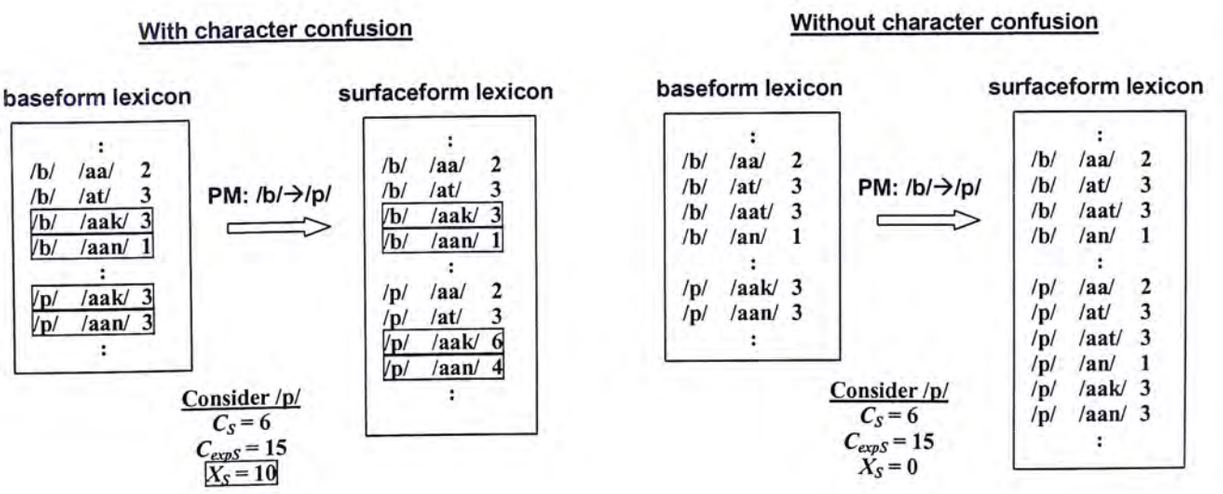


Figure 5.3: Calculation of character level confusion.

For example, as shown in Figure 5.3, /p/ is a surfaceform of /b/. The tables show the occurrences of the syllables with Initial /b/ and /p/ in the baseform or surfaceform lexicon. The number of each entry is the number of characters represented by that IF combination in the baseform or surfaceform lexicon. In the case with character confusion, the baseform lexicon contains 2 characters with an Initial /b/ followed by a Final /aa/, 3 characters with an Initial /b/ followed by Final /at/, 3 characters followed by /aak/ and 1 characters followed by /aan/. And, there are 3 characters with /p/ followed by /aak/, 3 characters followed by /aan/. The number of characters represented by /p/ in the original baseform lexicon C_S is 6. Allowing /b/ to be realized as /p/ will make the number of characters represented by /p/ in the expanded lexicon C_{expS} becomes 15. 10 out of 15 are being shared among /b/ and /p/. This will cause confusion.

If there is no character confusion, /b/ and /p/ do not have overlapping succeeding Final. No character is being shared among /b/ and /p/ after lexicon expansion.

We can see that larger character level confusion will result in more performance degradation. For example, /n/ → /l/ shows large number of degradation.

Result Analysis for Different PVDs:

Table 5.5 is the performance table when different PVDs are used. It can be seen that the behavior of the variants is similar among all PVDs.

b	*s*	PVD			1st CDDT PVD			2st CDDT PVD			1st PCDT PVD			2st PCDT PVD		
		I	*D*	*T*	*I*	*D*	*T*	*I*	*D*	*T*	*I*	*D*	*T*	*I*	*D*	*T*
aak	aa	5	1	11	5	2	11	5	2	11	5	2	11	5	2	11
aat	aa	8	1	17	8	1	15	8	1	15	8	1	15	8	1	15
aak	aat	6	0	7	4	0	6	4	0	6	4	0	6	4	0	6
aang	aan	1	1	4	1	1	4	1	1	4	1	1	4	1	1	4
aang	an	1	2	2	0	2	2	0	2	2	0	2	2	0	2	2
ang	**an**	**17**	**1**	**52**	**17**	**1**	**50**	**17**	**1**	**51**	**17**	**1**	**50**	**17**	**1**	**50**
aang	ang	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1
aap	ap	1	1	3	1	1	3	1	1	3	1	1	3	1	1	3
ak	at	6	0	16	5	0	15	5	0	14	5	0	14	5	0	14
ek	e	0	0	6	0	0	6	0	0	6	0	0	6	0	0	6
eng	ing	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2
im	in	1	1	3	-	-	-	-	-	-	-	-	-	-	-	-
it	i	1	0	5	-	-	-	-	-	-	-	-	-	-	-	-
ng	**m**	**10**	**2**	**65**	**10**	**2**	**65**	**10**	**2**	**65**	**10**	**2**	**65**	**10**	**2**	**65**
ok	o	2	0	14	2	0	11	2	0	11	2	0	12	2	0	12
on	ong	5	1	12	5	2	11	5	2	11	5	2	11	5	2	11
gw	**g**	**21**	**4**	**35**	**21**	**3**	**35**	**21**	**4**	**35**	**21**	**3**	**35**	**21**	**3**	**35**
kw	k	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2
n	**l**	**70**	**13**	**210**	**70**	**13**	**208**	**71**	**13**	**207**	**71**	**13**	**208**	**70**	**13**	**208**
ng	**null**	**34**	**8**	**146**	**35**	**9**	**146**	**36**	**9**	**146**	**35**	**9**	**146**	**35**	**9**	**148**
null	ng	1	0	4	1	0	4	1	0	4	1	0	4	1	0	4
Total		190	37	617	185	38	597	187	39	596	186	38	597	185	38	599

Table 5.5: Performance table of using different PVDs with VP Th = 0.05.

By analyzing the recognition results in detail, we observe that there are three Initials that are always confused. The *labialized (lip-rounded) velar* /gw/ is confused with *delabialized velar* /g/. *Nasal* /n/ is confused with the *lateral (tongue rolled)* /l/. *Nasal* /ng/ is always deleted. It is found that pronunciation variations for the Finals occur mainly in codas. *Nasal* codas, for example, *-ng*, *-n* and *-m* are always confused. Unvoiced *stops*, for example, *-k*, *-t* and *-p* are also easily confused. These observations agree with that from the linguistic study described in Section 2.2.1.

5.3.2 Experiment 2 — Pronunciation Modeling in Domain Specific task

In this experiment, the methods described above are evaluated in a continuous Cantonese speech recognition application in the stock domain. This experiment aims at investigating domain dependence of pronunciation variations in Cantonese. We also try to examine the effectiveness of a PM trained with a relatively small amount of data.

Experiment Conditions:

The testing data, STOCKTEST, contains 1300 sentences (about 65 minutes) recorded from 13 speakers. The acoustic model and the search engine are the same as the previous experiment. PM is trained by the CUTEST corpus.

Experimental Results:

(1) The use of PVDs with different VP Th

Table 5.6 shows the recognition results with PVDs that adopt different VP Th from 0.05 to 0.25. It can be seen that the use of PVD achieves a better performance of recognition and the extent of improvement varies with VP Th.

Different from the previous experiment, the result shows that a threshold of 0.05 does not give the most significant improvement on the accuracy. For the general domain task in Experiment 1, it is observed that most of the variants added will lead to more improvement than degradation although the threshold is small (VP Th = 0.05). However, the case is different in this experiment. This is because the PM training data for stock task are 1200 utterances from CUTEST. The amount is not large enough to train a reliable PM. Therefore, when the threshold is set to 0.05, many unreliable variations are added due to imperfect recognition. However, in the general domain, the training data contain 20341 utterances from CUSENT. As there are larger amount of data for training the PM, the PM is more reliable. Even the

threshold is set small, the variations added are still reflecting the actual realization of the baseform.

From Table 5.6, it is found that the optimal threshold of VP is 0.2. The average number of variations per IF unit for this threshold is 1.11 (81/73). The average number of variations per word is 1.21 (1511/1247). By using this threshold, a relative WER reduction of 7.30% can be achieved.

	Baseline	VP Th 0.05	VP Th 0.10	VP Th 0.15	**VP Th 0.20**	VP Th 0.25
WER (%)	12.06	12.01	11.46	11.25	**11.18**	11.26
Relative WER Reduction (%)		0.41	4.98	6.72	**7.30**	6.63
No. of IF variants	73	110	93	87	**81**	80
PVD size	1247	6213	1887	1671	**1511**	1499

Table 5.6: WER(%) of stock domain task using PVDs with different VP Th.

(2) The use of different PVDs with VP Th = 0.2

Table 5.7 shows the recognition results of using different PVDs with same VP Th = 0.2. "2nd CDDT/ PCDT PVD" performs a little better than "1st CDDT PVD" and "PVD". This may be due to the fact that the number of data used to train the CM is small. The mis-recognized surfaceform seem to be biased to certain IFs. The VPs for these surfaceforms are comparable to those which are true pronunciation variations. If small VP Th is used for pruning, these mis-recognized surfaceforms cannot be pruned away. On the other hand, if large VP Th is used for pruning, useful surfaceforms will also be removed. Therefore, the CM obtained is not very reliable. Refining the CM by "1st CDDT" and "2nd CDDT" or by using "PCDT" will reduce the number of mis-recognized surfaceforms, making the CM more reliable.

	Baseline	PVD	1st CDDT PVD	2nd CDDT PVD	1st PCDT PVD	2nd PCDT PVD
WER (%)	12.06	11.18	11.21	11.17	11.17	11.17
Relative WER Reduction (%)		7.30	7.04	7.38	7.38	7.38
No. of IF variants	73	81	81	81	81	81
PVD size	1247	1511	1511	1511	1511	1511

Table 5.7: WER(%) of stock domain task using different PVDs with VP Th = 0.2.

Result Analysis for PVD using VP Th = 0.05:

Table 5.8 shows the performance table of part of the variants that were added to the lexicon with VP Th = 0.05. The total number of surfaceform IFs being recognized is 1427. Out of the 1427 cases, 109 cause improvement in character recognition and 71 cause degradation. The net improvement accounts for the little increase in recognition accuracy.

b	*s*	*I*	*D*	*T*	O_B	O_S	*VP*	*s*	N_S	N_{expS}	*EF*	C_S	C_{expS}	X_S
aa	aak	1	4	14	639	84	0.064	aak	13	133	10.23	6	41	30
aat	aak	0	6	0	13	84	0.235							
aap	aak	0	0	3	18	84	0.375							
ak	aak	0	0	1	43	84	0.122							
at	ak	1	2	14	895	43	0.062	ak	10	64	6.4	4	11	0
o	ou	1	6	28	1859	480	0.059	ou	130	263	2.023	24	44	36
n	l	2	8	45	58	220	0.854	l	51	95	1.86	27	43	13
ng	l	1	2	24	921	220	0.061							
eon	an	0	3	7	72	509	0.080	an	28	50	1.79	14	31	19
w	m	3	4	27	1013	926	0.064	m	85	122	1.44	26	46	20
b	*s*	*I*	*D*	*T*	O_B	O_S	*VP*	*s*	N_S	N_{expS}	*EF*	C_S	C_{expS}	X_S
ng	null	47	9	695	921	427	0.720	null	37	69	1.86	5	14	5
ek	e	5	0	16	39	55	0.219	e	12	22	1.83	7	8	2
ng	m	28	5	58	98	102	0.819	m	12	25	2.08	1	3	3
gw	g	8	5	52	163	1931	0.257	g	148	165	1.11	58	68	28
Total		109	71	1247					2820	3748				

Table 5.8: Performance table with VP Th = 0.05.

From Table 5.8, it is found that some variations can never give improvement and have to be filtered out. The upper part shows the variations added causing more degradation. The lower part shows the variations added causing more improvement.

In this domain-specific task, the lexical tree expansion factor and character level confusion are more significant than those in the general domain task. This is because in a domain specific task, the dictionary size is small. After expansion, even a small increase in the number of entries will show large effect in the whole lexicon and lexical tree. Therefore, expansion factor and character level confusion are more sensitive in the domain specific task.

For example, /aa/, /aat/, /aap/ and /ak/ can all be realized as /aak/. Originally, /aak/ occurs in the baseform lexical tree only 13 times in 6 different characters. After adding the surfaceforms in the lexicon, /aak/ occurs 133 times in the expanded

lexical tree representing 41 characters. Both the expansion factor and character level confusion are very large. If one makes the correct pronunciation of /aak/, there will be 41 instead of 6 possible choices of characters. The wrong word would be retrieved more often.

In general, the higher the expansion factor or larger the character level confusion, the larger the confusability, the poorer the recognition performance. The larger the VP and O_B, the more the chance for the PM to be applied results in more improvement.

Result Analysis for PVD using VP Th = 0.2:

Table 5.9 shows the performance table of all the variants that were added to the lexicon with VP Th set to be 0.2. The expansion factor and character level confusion are reduced results in less degradation. It is observed that most of the variants added will lead to more improvement than degradation.

b	*s*	*I*	*D*	*T*	O_B	O_S	*VP*	*s*	N_S	N_{expS}	*EF*	C_S	C_{expS}	X_S
aap	aak	0	0	3	18	84	0.375	aak	13	17	1.31	6	8	0
ak	at	0	0	16	43	895	0.516	at	54	64	1.19	7	11	0
ang	an	3	0	70	112	509	0.538	an	28	37	1.32	14	21	6
ek	e	5	0	16	39	55	0.219	e	12	22	1.83	7	8	2
ng	m	30	2	65	98	102	0.883	m	12	25	2.08	1	3	3
on	ong	0	0	0	0	397	0.349	ong	41	43	1.05	17	18	3
gw	g	8	3	54	163	1931	0.257	g	148	165	1.11	58	68	28
n	l	2	6	52	58	220	0.854	l	51	63	1.24	27	34	7
ng	null	51	11	702	921	427	0.766	null	37	69	1.86	5	14	5
Total		99	22	979	1100				2820	2929				

Table 5.9: Performance table with VP Th = 0.2.

Result Analysis for different PVDs using VP Th = 0.2:

Table 5.10 is the performance table when different PVDs are used. Same variations are added in all PVDs. The performance for the variants is similar among all PVDs.

b	*s*	PVD			1st CDDT PVD			2st CDDT PVD			1st PCDT PVD			2st PCDT PVD		
		I	*D*	*T*	*I*	*D*	*T*	*I*	*D*	*T*	*I*	*D*	*T*	*I*	*D*	*T*
aap	aak	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3
ak	at	0	0	16	0	0	17	0	0	17	0	0	17	0	0	17
ang	an	3	0	70	3	0	69	3	0	69	3	0	70	3	0	70
ek	e	5	0	16	5	0	16	5	0	16	5	0	16	5	0	16
ng	m	30	2	65	30	2	65	30	2	65	30	2	65	30	2	65
gw	g	8	3	54	8	3	54	8	3	54	8	3	54	8	3	54
n	l	2	6	52	2	6	52	2	6	52	2	6	52	2	6	52
ng	null	51	11	702	51	11	709	51	11	708	51	11	708	51	11	704
Total		99	22	978	99	22	985	99	22	984	99	22	985	99	22	981

Table 5.10: Performance table of using different PVDs with VP Th = 0.2.

5.3.3 Experiment 3 — PVD Pruning by Word Unigram

Experiment Conditions:

In Experiment 3, the method of PVD pruning by word unigram is evaluated in the same stock domain-specific task. The acoustic model, PM and the search engine are the same as the previous experiment. The word unigram obtained by LM is used for PVD pruning.

Experiment Results:

From Table 5.11, it is observed that the performance of pruned PVD achieves the best performance. This shows that word unigram is an important factor for controlling the number of variations to be added in the PVD.

	Baseline	PVD (VP Th 0.05)	Pruned PVD by word unigram
WER (%)	12.06	12.01	11.08
Relative WER Reduction (%)		0.41	8.13
No. of IF variants	73	110	110
PVD size	1247	6213	1863

Table 5.11. WER(%) of stock domain task using PVD pruned by word unigram.

5.4 Summary

In this chapter, we discussed the incorporation of pronunciation model (PM) into the pronunciation lexicon to handle phone change. IF confusion matrix (CM) is used for augmenting the baseform lexicon with additional pronunciation variants to build a Pronunciation Variation Dictionary (PVD). From the result, it is found that that the use of PVDs achieves a better performance of recognition. We also investigated the use of CDDT and PCDT to obtain the refined PVDs. It is found that decision-tree based refinement of PVD does not lead to additional performance improvement when large amount of training data is used. However, refined PVD performs a little better when small amount of training data is used. Word unigram can be used for PVD pruning. It is observed that the performance of pruned PVD achieves the best performance.

The effectiveness of different sets of PVDs is evaluated in a Cantonese LVCSR task and a stock domain task. It is found that a small VP Th can be used for sufficient PM training data and a larger VP Th must be used if the amount of training data is not enough. The performance of the variants added can be explained by several factors including the (1) occurrences of baseform and surfaceform IF, (2) variation probability, (3) lexical tree expansion factor, and (4) character level confusion.

Reference

[1] Y. Liu, "Pronunciation Modeling for Spontaneous Mandarin Speech Recognition", *Ph.D. Thesis*, The Hong Kong University of Science and Technology, 2002.

[2] M.K. Liu *et al*, "Mandarin Accent Adaptation Based on Context-Independent/Context-Dependent Pronunciation Modeling", in *Proceedings of ICASSP-00*, Vol.2, pp.1025-1028, Istanbul, 2000.

[3] W. Byrne *et al*. "Pronunciation Modeling Using a Hand-labeled Corpus for Conversational Speech Recognition", in *Proceedings of ICASSP*-98, Vol.1, pp.12-15, Seattle, 1998.

[4] M. Eichner *et al*, "Data – Driven Generation of Pronunciation Dictionaries in the German Verbmobil Project – Discussion of Experimental Results", in *Proceedings of ICASSP-00*, Vol.3, pp.1687-1690, Istanbul, 2000.

[5] T. Sloboda *et al*, "Dictionary Learning for Spontaneous Speech Recognition", in *Proceedings of ICSLP-96*, Vol.4, pp.2328-2331, Philadelphia, 1996.

[6] P. Kam *et al*, "Modeling Pronunciation Variation for Cantonese Speech Recognition", in *Proceedings of PMLA-02*, pp.12-17, Denver, 2002.

[7] M.Y. Tsai *et al*, "Improved Pronunciation Modeling by Properly Integrating Better Approaches for Baseform Generation, Ranking and Pruning", in *Proceedings of PMLA-02*, pp.77-82, Denver, 2002.

[8] H. Schramm *et al*, "Efficient integration of multiple pronunciations in a large vocabulary decoder", in *Proceedings of ICASSP-00*, Vol.3, pp.1659-1662, Istanbul, 2000.

[9] W.N. Choi, "An Efficient Decoding Method for Continuous Speech Recognition Based on a Tree-Structured Lexicon", *M.Phil. Thesis*, The Chinese University of Hong Kong, 2001.

Chapter 6

Pronunciation Modeling at Acoustic Model Level

In this chapter, we will discuss the incorporation of pronunciation model (PM) into the acoustic model to deal with sound change. This is done by refining the acoustic model to include variation information.

We have discussed the methods to handle phone change which happens when a canonical phoneme is realized as a different phoneme. Such a change can be modeled by converting the baseform phoneme to a surfaceform phoneme. On the other hand, sound change happens at a lower level, i.e. phonetic or sub-phonetic level [1][2]. When sound change occurs, the pronunciation is ambiguous between the baseform phoneme and its surfaceform phoneme. It cannot be modeled by simply replacing the canonical phoneme with another phoneme. To deal with a sound change, pronunciation modeling must be applied at a sub-model level, for example, at the states of HMMs or the Gaussian mixture components of HMM states.

In general, acoustic models are trained with the assumption that the training data follow the baseform pronunciations exactly. This convenient but apparently wrong assumption renders the acoustic model thus trained to be inadequate to represent the variations of speech sounds. It would be useful to refine the acoustic model by taking into account the realistic pronunciations.

Three algorithms of acoustic model refinement are investigated in this chapter: (1) sharing Gaussian mixture components of HMM states in both baseform and surfaceform models; (2) adapting the mixture components of the baseform

models towards those of the surfaceform models; (3) selectively reconstructing new acoustic model through sharing or adapting.

6.1 Hierarchy of HMM

HMM is a finite state model that characterizes the acoustic signal in a statistical way. The number of states is determined by the complexity of a phone. The observation probability at each HMM state is usually modeled by a continuous probability density function (pdf). In the simplest case, it is a Gaussian distribution.

If the acoustic signals contain a great deal of variability, the distribution of features may not be well represented by a single Gaussian pdf. Instead, the acoustic vectors for training a particular state are divided into M categories, each being represented by a Gaussian pdf. Therefore, each HMM state is associated with a mixture of M multivariate Gaussian pdf's that model the acoustic variation. If more mixture components are used, the statistical distribution of the acoustic signals would be better represented.

Figure 6.1 shows the HMMs for the Cantonese Initials /b/, /d/ and /p/. Each model has 3 states. Each state in an HMM is associated with M Gaussian mixture component pdf's, denoted as $m(1),m(2)...m(M)$.

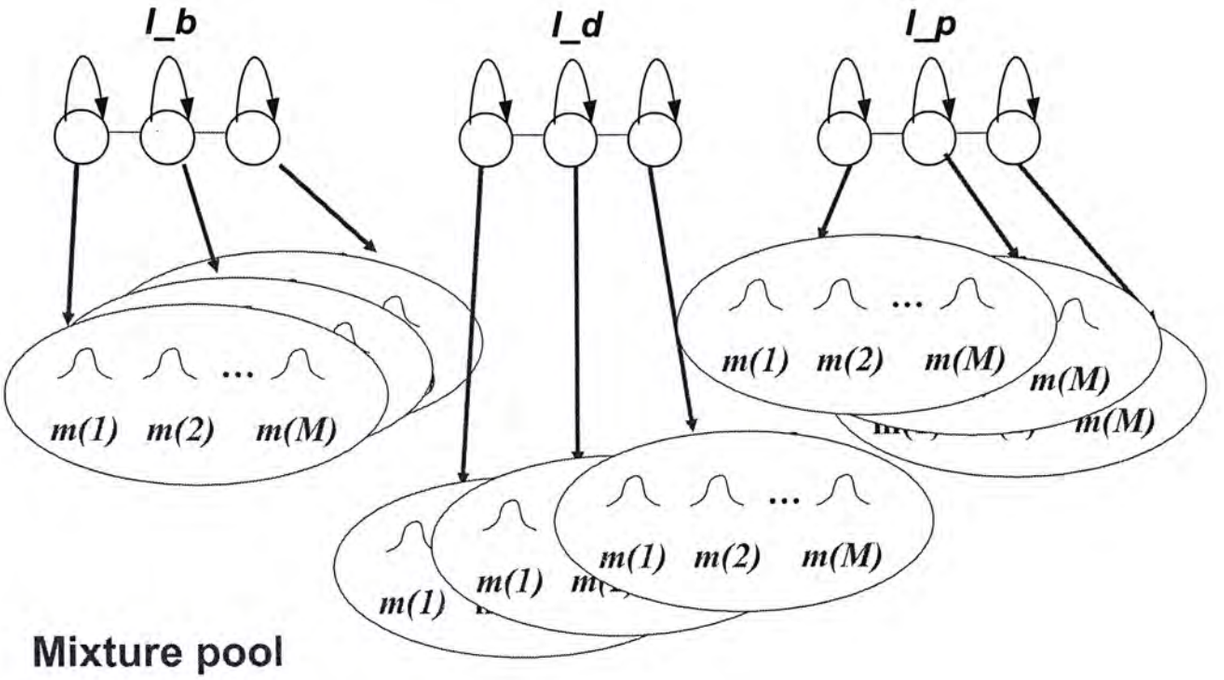


Figure 6.1: HMMs for Cantonese Initials /b/, /d/ and /p/.

6.2 Sharing of Mixture Components

In this approach, the mixture components in the surfaceform models are used to enrich the baseform models such that they have a better coverage of acoustic variability [3].

Suppose that the predicted surfaceform of the baseform phoneme /b/ is /p/. This prediction can be made by the pronunciation model (PM) as described in Chapter 4. CM with VP Th = 0.05 is used here as the PM. As shown in Figure 6.2, each of the phonemes /b/ and /p/ is modeled by a three-state HMM. We can use the surfaceform model *I_p* to refine the baseform model *I_b*. This is done by tying the mixture component pdf's of the surfaceform model to the baseform model. The parameters of the model *I_p* are included in the model *I_b* to form a complementary acoustic model to represent the realistic acoustics.

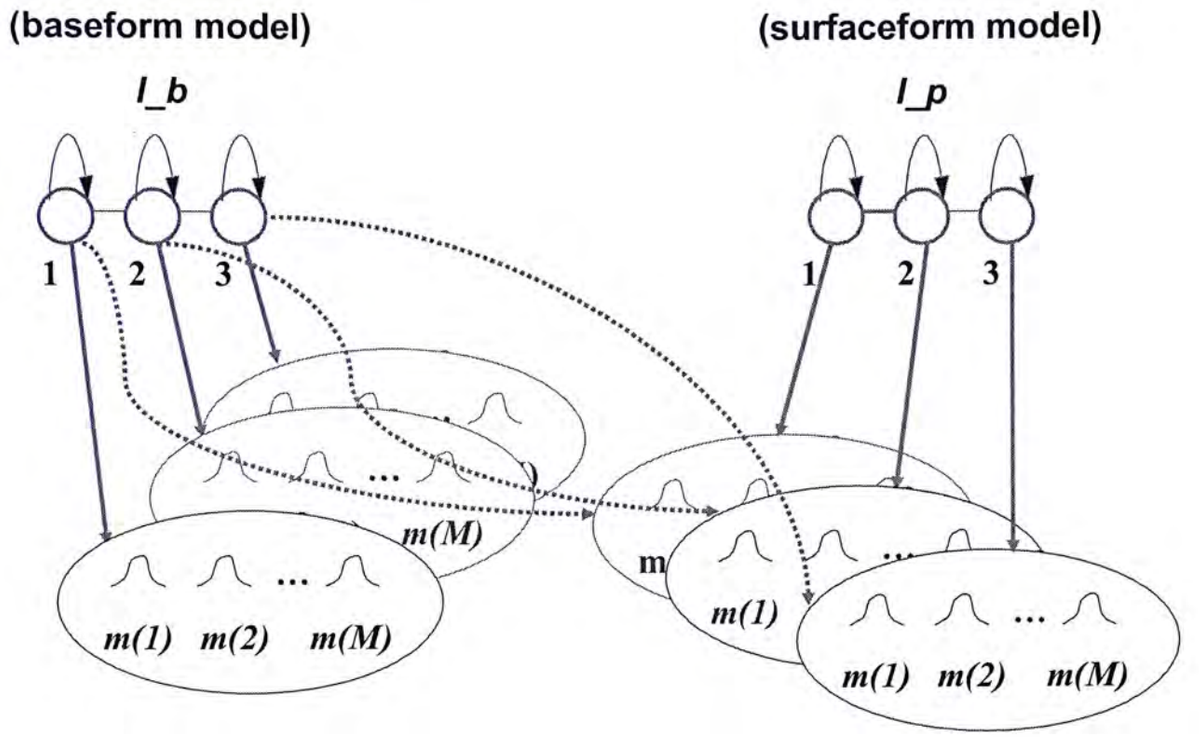


Figure 6.2: Mixture component sharing of surfaceform model *I_p* with baseform model *I_b*.

We first align the states of the baseform and surfaceform models. It is assumed that both models have exactly the same number of states. Then the *n*-th state in the baseform model is mapped to the *n*-th state of the surfaceform model.

For each state of the baseform model, the observation probability density function is modified by including the contribution of the surfaceform output observation pdf's. Let the observation pdf of the original baseform state *j* be

$$p_j(o_t) = \sum_{m=1}^{M} w_{jm} N(o_t; \mu_{jm}, \Sigma_{jm}) \qquad (6.1)$$

where M is the number of Gaussian mixture components, and w_{jm} is the weight for m-th mixture component of state j. The modified pdf is given as

$$p_j'(o_t) = P(s_k = b \mid b) \cdot p_j(o_t) + \sum_{\substack{k=1 \\ s_k \neq b}}^{K} P(s_k \mid b) \cdot q_{kj}(o_t) \qquad (6.2)$$

where K is the total number of surfaceform pronunciations for the baseform pronunciation b, $q_{kj}(o_t)$ is the output pdf of the j-th state of the k–th surfaceform s_k, $P(s_k=b|b)$ is the VP for a baseform to be realized as itself, and $P(s_k|b)$ is the VP for a baseform to be realized as s_k.

The number of mixtures in the modified baseform model depends on the number of surfaceform pronunciations. More surfaceform pronunciations will bring in more mixture components to the modified baseform model. As the number of mixture components of each state is changed after the sharing process, re-estimation of model parameters is needed.

6.3 Adaptation of Mixture Components

Although the previous approach yields an acoustically complementary model, it also increases the model size by including more mixture components. As a result, it costs more storage for the additional model parameters and requires more computation in decoding. On the other hand, if the surfaceform model is actually very similar to the baseform model, including those similar mixture components in the modified baseform model can be unnecessarily superfluous.

In this section, we describe another method that refines the baseform model by adapting its existing Gaussian mixture components instead of introducing extra components. The states of the baseform and surfaceform models are first aligned to form a one-to-one mapping. The baseform mixture component pdf's are adapted towards the nearest surfaceform components. Within each pair of states, we need to pair up the baseform and surfaceform mixture component pdf's with the smallest distance as shown in Figure 6.3.

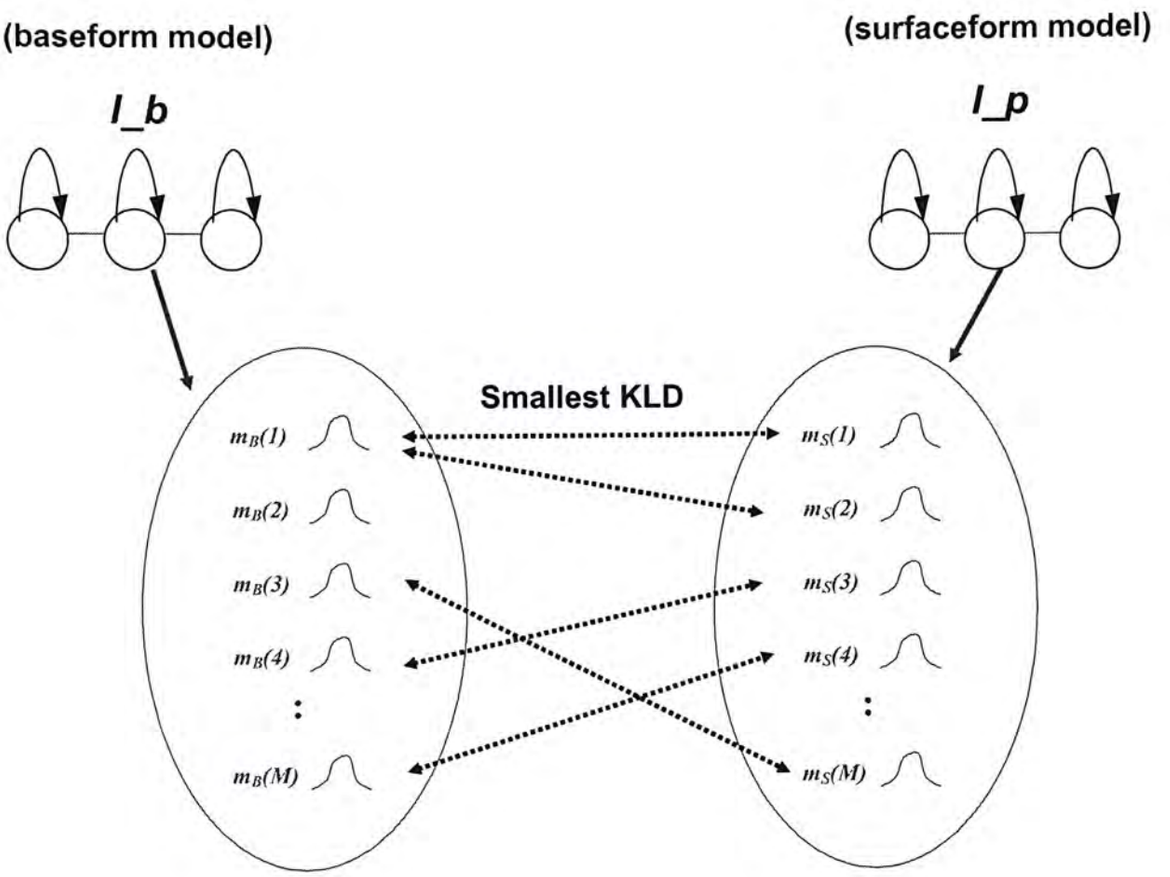


Figure 6.3: Mapping between baseform and surfaceform mixture component pdf's with smallest KLD.

The distance between two mixture component pdf's can be calculated by the Kullback-Leibler divergence (KLD) [4], which is an information-theoretic measure for finding the similarity between two given pdf's. Specifically, when these pdf's, denoted as f and g, are multivariate Gaussian, the symmetric KLD has a closed form as

$$d(f,g) = \frac{1}{2} trace\{(\Sigma_f^{-1} + \Sigma_g^{-1})(\mu_f - \mu_g)(\mu_f - \mu_g)^T + \Sigma_f \Sigma_g^{-1} + \Sigma_g \Sigma_f^{-1} - 2I\} \quad (6.3)$$

where μ and Σ are the mean vectors and the co-variance matrices of the two pdf's respectively.

Let $m_B(i)$, $m_S(i)$, for i = 1 to M, be the M baseform and surfaceform mixture components respectively in that pair of states. We compute the KLDs between all possible pairs of mixture components, $(m_B(i), m_S(j))$. Each of the surfaceform

mixture components is paired up with the nearest baseform mixture component in KLD. That is, for each $m_S(j)$, we find

$$\hat{i} = \arg\min_{m_B(i)} d(m_B(i), m_S(j)) \qquad (6.4)$$

As a result, a particular baseform mixture component $m_B(i)$ may be associated with k surfaceform components. For example, in Figure 6.3, there are two surfaceform pdf's, $m_S(1)$ *and* $m_S(2)$ being mapped to the baseform pdf $m_B(1)$. To modify $m_B(i)$, we first compute the centroid c_S of the k surfaceform mixture components, weighted by the corresponding mixture weights. For example, the centroid pdf between $m_S(1)$ *and* $m_S(2)$ is shown as in Figure 6.4.

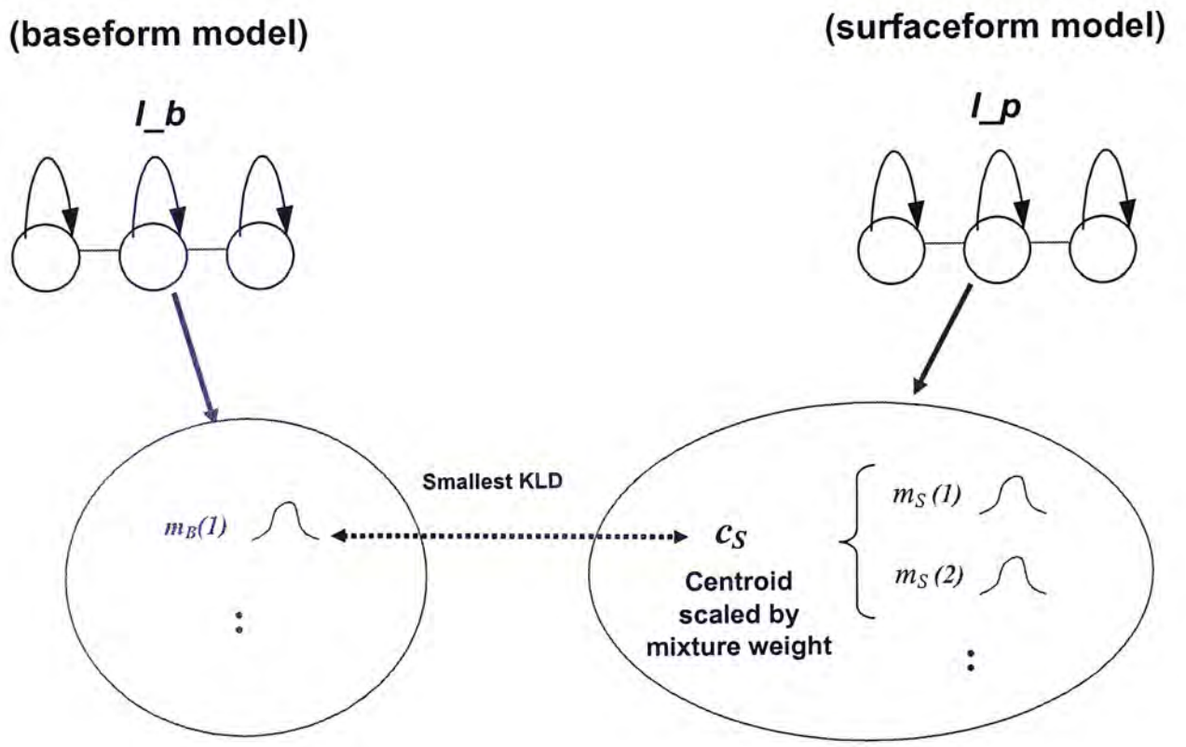


Figure 6.4: Centroid c_S of 2 surfaceform mixture components, $m_S(1)$ *and* $m_S(2)$.

If the baseform has K surfaceform pronunciations, there will be K centroids generated. Let them be denoted as $c_{S_1} \ldots c_{S_K}$. All these K individual centroids and the baseform component are weighted by the corresponding VPs. A combined centroid is then obtained from these centroids and the baseform component. This VP weighted combined centroid becomes the modified baseform mixture $m_B(i)'$. For example, Figure 6.5 shows that baseform /b/ has two surfaceforms, /p/ and /d/. Then,

$m_B(1)$ is associated with two centroids, c_{S_1} and c_{S_2}. $m_B(1)$, c_{S_1} and c_{S_2} are weighted by the corresponding VPs, 0.8, 0.15 and 0.05. A centroid of these 3 mixture components is computed. This VP weighted centroid is the modified baseform mixture $m_B(1)'$.

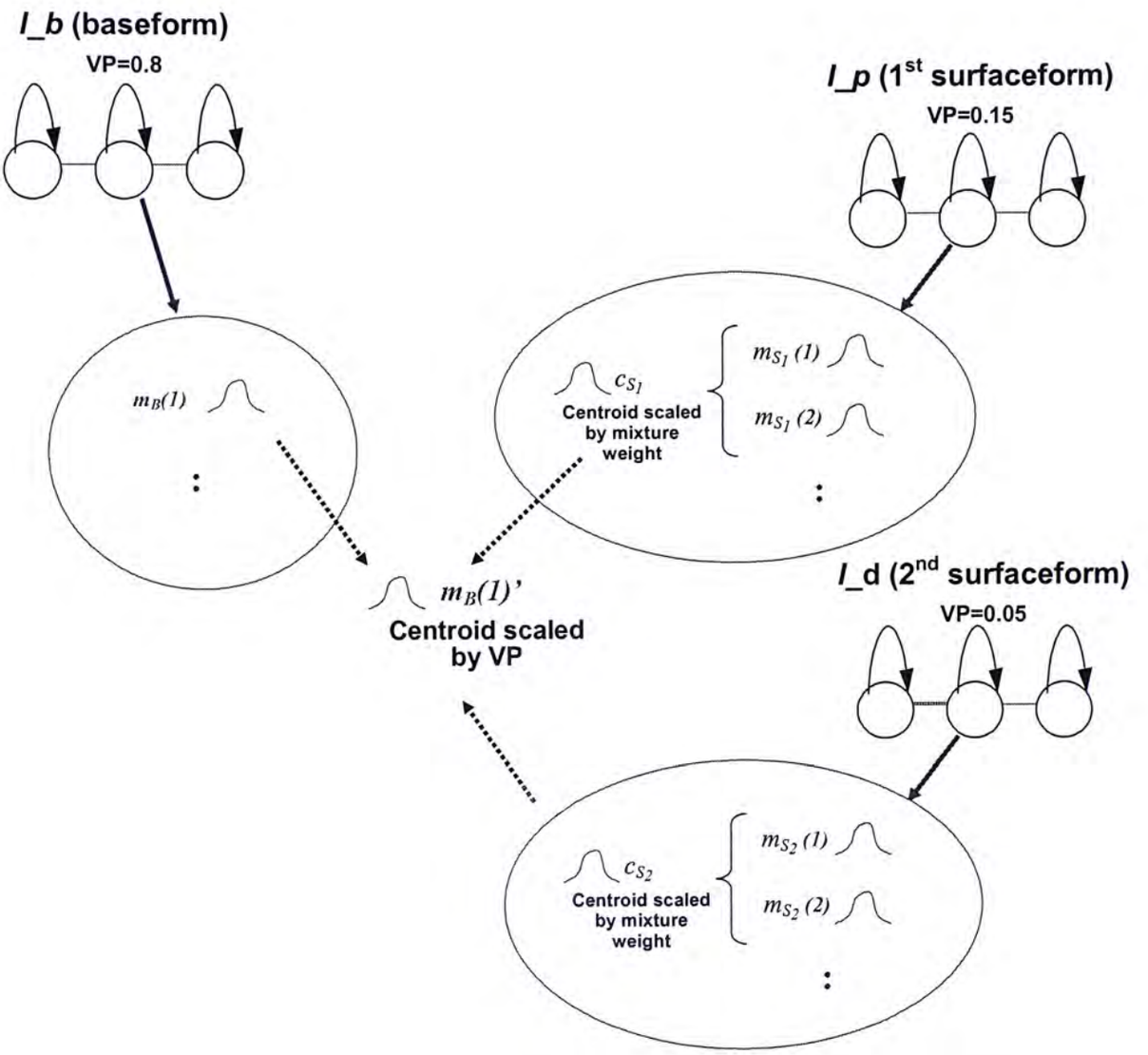


Figure 6.5: VP weighted centroid of 2 surfaceform centroids and the baseform component.

The weighted centroid, f_c, of k mixture components can be found by minimizing the weighted divergence as

$$\{\mu_c', \Sigma_c'\} = \arg\min_{\mu_c, \Sigma_c} \sum_{n=1}^{k} a_n d(f_c, f_n) \qquad (6.5)$$

where a_n is the weighting coefficient of the n-th pdf, f_n. In calculating the individual centroid c_S for the k surfaceform mixtures, the mixture weight is used for a_n. In

calculating the VP weighted optimal centroid, $m_B(i)'$, the VP is used for a_n. Similar to the derivation in [4], if diagonal co-variances are used, the i-th component of the centroid is

$$\mu_c'(i) = \frac{\Sigma_{n=1}^{k} a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i)) \mu_n(i)}{\Sigma_{n=1}^{k} a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i))}$$
$$\Sigma_c'(i) = \sqrt{\frac{\Sigma_{n=1}^{k} a_n [\Sigma_n(i) + (\mu_c(i) - \mu_n(i))^2]}{\Sigma_{n=1}^{k} a_n \Sigma_n^{-1}(i)}} \qquad (6.6)$$

6.4 Combination of Mixture Component Sharing and Adaptation

Both of the approaches described in Section 6.2 and 6.3 attempt to adjust the baseform models using the mixture components in the surfaceform models. In the case of adaptation, the baseform pdf's are shifted towards the corresponding surfaceform pdf's. If the surfaceform pdf is far away from the baseform one, the extent of modification would be substantial and consequently the modified pdf may fail to model the original baseform. In the case of mixture sharing, as we mentioned earlier, including more mixture components into the baseform models may be superfluous. Thus, we propose to combine these two approaches. The idea is to perform adaptation using the surfaceform components that are close to the baseform, and at the same time, to use those relatively distant components for sharing.

We first try to analyze the KLD between the mixture component pdf's in the baseform HMMs and the nearest surfaceform pdf's in the surfaceform HMMs. Let s_k be a predicted surfaceform realization of the baseform phoneme b. In our research, the acoustic model is a set of right-context bi-IF HMMs. Given the context-independent IF unit b, we can find a number of context-dependent HMMs that correspond to b or s_k. For each state pair, the KLDs for all possible mixture pairs, i.e. $(m_B(i), m_S(j))$ are computed. Each of the surfaceform mixture components is paired up with the nearest baseform mixture component in KLD. After that, M KLDs can be obtained if each state contains M mixture components. If N HMMs are found for

the unit *b*, we can compute *N* x *M* KLDs for each state. The distributions of these KLDs are plotted for this variation pair. Figure 6.6 shows the KLD distribution for some of the variation pairs /aak/→/aa/, /aak/→/aat/, /aang/→/aan/ and /aang/→/an/.

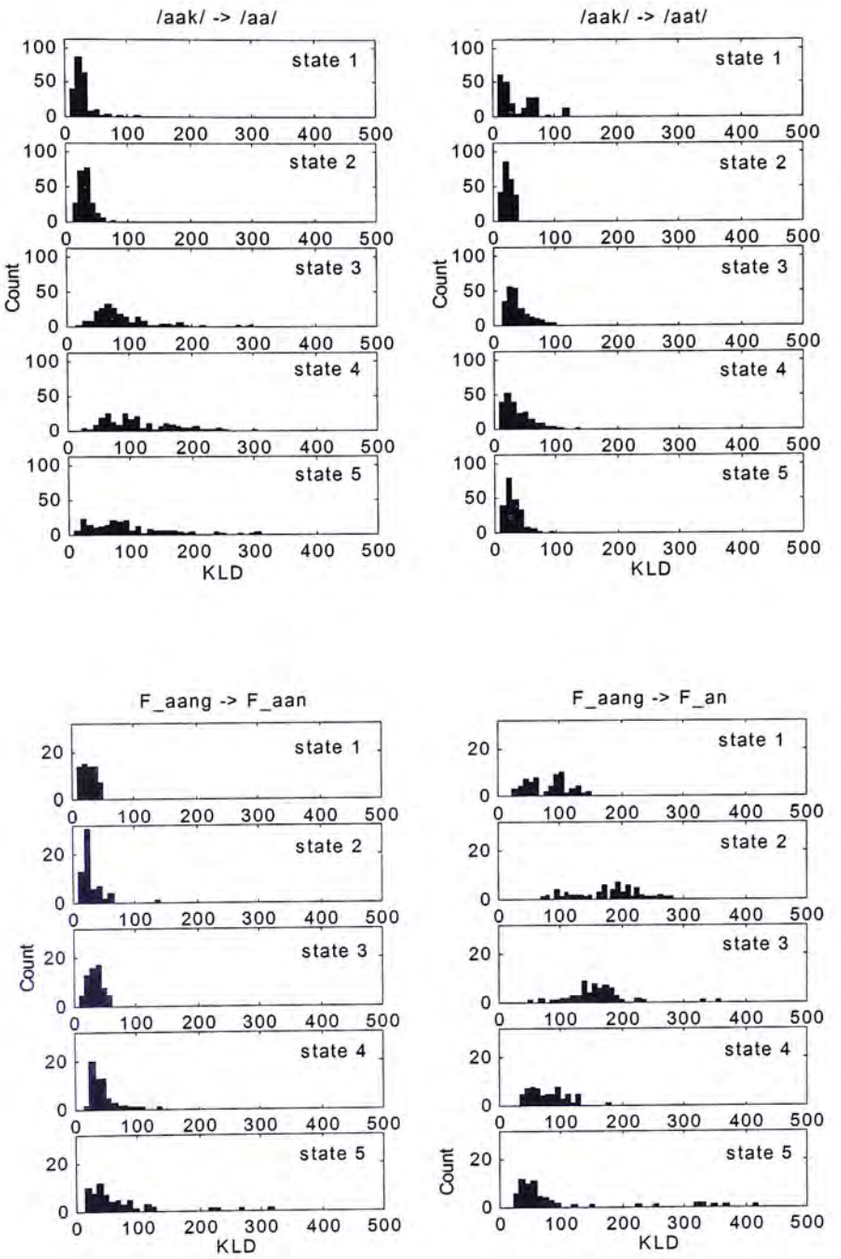


Figure 6.6: KLD distributions for variation pair /aak/→/aa/, /aak/→/aat/, /aang/→/aan/ and /aang/→/an/.

From the KLD distributions for different variations, the following observations can be made:

(1) The distribution of KLD varies among different variation pairs.

(2) Two main types of distributions can be identified. One is showing consistently small KLD (<50), while the others showing a wider range of KLD values.

(3) Small value of KLD is obtained when a vowel nucleus remains unchanged or a consonant Initial/coda is substituted by another phoneme in the same phone class. For example, for the cases of /aat/→/aa/, /aak/→/aat/, the baseform and surfaceform units have the same vowel nucleus. It is noted that the first two states of their HMMs show very small KLD values. Similar observation is made in the last 3 HMM states for /aak/→/aat/, /ak/→/at/, /aang/→/aan/, /im/→/in/, which involves substituted codas. It is also found in the first few states in mis-pronounced Initials, for example, /n/→/l/, /ng/→/null/ and /null/→/ng/.

(4) Widely-distributed KLD is observed for those confused pairs of phonemes with the vowel nucleus completely changed or the coda deleted. For example, the baseform and surfaceform units have different vowel nuclei for the cases of /aang/→/ang/, /aap/→/ap/, /eng/→/ing/. It is noted that the first 3 states of their HMMs show large KLD values. Similar phenomenon is observed in the last 3 HMM states for /aak/→/aa/, /aat/→/aa/, /ek/→/e/, /it/→/i/, /ok/→/o/ in which stop codas are deleted. Such kind of distribution is also found in the last 2 states in velar lip-rounded Initials mixed with velar Initials, for example, /gw/→/g/, /kw/→/k/.

Based on the above discussion, we plot the KLD distributions according to the type of variation (see Figure 6.7). These types of variation are: (1) Finals with deleted stop coda; (2) Finals with stop coda interchanged; (3) Finals with nasal coda interchanged; (4) Finals with vowel identity changed; (5) mis-pronounced Initial and (6) velar lip-rounded Initial changes to velar Initial.

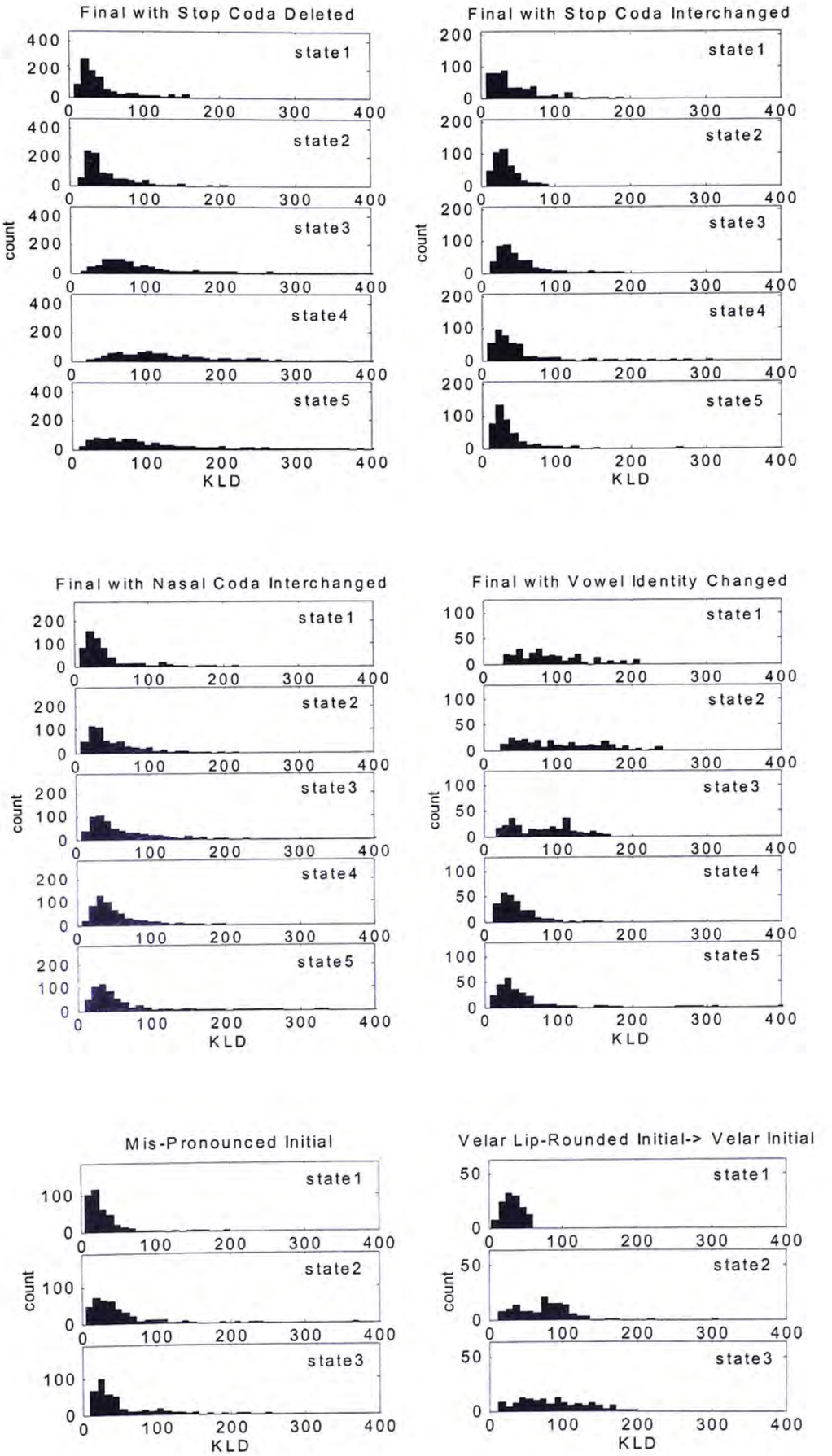


Figure 6.7: KLD distributions for different types of pronunciation variations.

Small KLD means that the mixture components of the baseform and surfaceforms are similar. In this case, the baseform components are adapted towards the surfaceform. In the case of a widely-distributed KLD, the surfaceform components should not be used to adapt the baseform components, but be kept along with the modified baseform model in order to characterize irregular pronunciations explicitly. In this way, a combined approach of baseform model refinement is formulated as shown in Table 6.1. "S1", "S2",..., "S5" stand for state 1, state 2,..., state 5 of an HMM respectively. "A" and "S" stand for adaptation and sharing of mixture components respectively. At each state, either adaptation or sharing is performed based upon their KLD distributions.

	S1	S2	S3	S4	S5
Final with Deleted Stop Coda	A	A	S	S	S
Final with Stop Coda Interchanged	A	A	A	A	A
Final with Nasal Coda Interchanged	A	A	A	A	A
Final with Vowel Identity Changed	S	S	S	A	A
Mis-Pronounced Initial	A	A	A	--	--
Velar Lip-Rounded Initial change to Velar Initial	A	S	S	--	--

Table 6.1: Mixture combination in different states using adaptation or sharing for different variation types.

6.5 Recognition Experiments

Experimental Conditions:

CUTEST is used as the testing data. The acoustic model is refined by the methods mentioned above. IF confusion matrix with VP Th = 0.05 is used.

Experimental Results:

The experimental results for different refining methods are shown as in Table 6.2. Three baseline systems are prepared with different number of mixture components in the acoustic model. "Sharing" refers to HMM mixture component sharing discussed in Section 6.2. "Adaptation" refers to HMM mixture component adaptation discussed in Section 6.3. "Combined" refers to HMM refinement using both "sharing" and "adaptation" discussed in Section 6.4. The number of mixture components in the acoustic model for "Baseline 1", "Baseline 2" and "Baseline 3" are intentionally made the same as that of "Adaptation", "Sharing" and "Combined" respectively, so that fair comparison would be possible. In general, a better recognition performance can be attained by refining the parameters in the baseform acoustic model.

	Baseline 1 (32144)	Baseline 2 (37505)	Baseline 3 (34042)	Sharing (37505)	Adaptation (32144)	Combined (34042)
No retrain	25.34	24.34	24.93	24.38	24.70	24.87
Retrained	N/A	N/A	N/A	23.96	N/A	24.57

Table 6.2: WER(%) of LVCSR task using three different HMM refining methods. Figures inside () are the numbers of mixture components in different model sets.

Experimentally we found that the optimal threshold of variation probability (VP) to prune less frequent surfaceforms is 0.05. The KLD thresholds for HMM sharing and adaptation are set at 300 and 50 respectively. By using these thresholds, the relative WER reductions are 3.79% and 2.53% for "sharing" and "adaptation" respectively compared with "Baseline 1". The error reduction can be further improved to 5.45% when the model parameters are re-estimated in the case of "sharing". As no extra mixture component is included in the models for "adaptation", no re-estimation is done to prevent any loss of surfaceform information.

From the results, it is found that including surfaceform mixture components in the acoustic model gives better recognition performance than adapting the

baseform mixture components. This may be due to the fact that extra mixture components are used for representing the acoustic model. The number of mixture components used in the model set for "sharing" is 16.7% more than that for "adaptation". Another reason is that not all surfaceform mixture components are appropriate for adaptation. Some of them may represent irregular or idiosyncratic pronunciations. Therefore, we may want to choose selectively those useful mixture components for refining the acoustic model rather than merge them all with the baseform mixtures.

Combining "sharing" and "adaptation" reduces WER by 1.85%. The relative error reduction can be further improved to 3.04% when the models are re-estimated. This approach keeps a small number of mixture components in the model set while solving the problems in "adaptation". With only 6.4% more mixture components than the "adaptation" approach, it obtains good performance improvement.

In general, better recognition accuracy can be attained with a model set having larger number of mixtures. Both "sharing" and "combined" increases the number of mixtures of the model set. In order to have a fair comparison, "baseline 2" and "baseline 3" are trained so that the number of mixtures of the model sets is the same as "sharing" and "combined" respectively.

Comparing with "baseline 2", the relative WER reduction for "sharing" change from 5.45% to 1.56%. Comparing with "baseline 3", the relative WER reduction for "combined" change from 3.04% to 1.44%. This illustrates that the improvement is partially due to pronunciation modeling and partly contributed by the increase in number of mixtures of the model set.

6.6 Result Analysis

In this section, we try to analyze the recognition results in detail and explain the causes that lead to performance improvement. The baseline recognition results are used for a contrastive reference to reveal which variations cause improvement.

6.6.1 Performance of Sharing Mixture Components

Table 6.3 is a performance table for the first method "sharing". M_s is the surfaceform model used to refine the baseform model M_b. "Improve" refers to the count of improved cases with the respective total count of occurrences. Positive value means improvement while negative value means degradation. For example, by using the surfaceform models *F_aa* and *F_aat* to refine the baseform model *F_aak*, a net improvement of 6 is attained in 156 occurrences of /aak/. In most situations, the refined models show improvement in recognition accuracy. 2316 IFs in the recognized output correspond to the modified models. 97 out of these 2316 are improved.

M_s used to refine M_b			Confused M_s		
M_b	M_s	Improve	M_b	M_s	Improve
F_aak	*F_aa/F_aak/ F_aat*	6 / 156	*F_aa*	*F_aa*	-4 / 369
F_aang	*F_aan/F_aang/ F_an/F_ang*	3 / 20	*F_aan* *F_an*	*F_aan* *F_an*	-2 / 412 -2 / 383
F_aat	*F_aa/F_aat*	2 / 176			
F_ak	*F_ak/F_at*	3 / 90	*F_at*	*F_at*	-1 / 333
F_ang	*F_an/F_ang*	19 / 191			
F_im	*F_im/F_in*	2 / 81	*F_in*	*F_in*	-5 / 508
F_ng	*F_m/F_ng*	10 / 90	*F_m*	*F_m*	-2 / 6
F_ok	*F_o/F_ok*	7 / 237	*F_o*	*F_o*	-3 / 396
F_on	*F_on/F_ong*	4 / 97	*F_ong*	*F_ong*	3 / 419
I_gw	*I_g/I_gw*	7 / 226	*I_g*	*I_g*	-7 / 915
I_n	*I_l/I_n*	21 / 302	*I_l*	*I_l*	-5 / 735
I_ng	*I_ng/I_null*	15 / 257			
I_null	*I_ng/I_null*	1 /154			
Total		97 / 2316			

Table 6.3: Performance table for "sharing".

The attained improvement can be explained as in Figure 6.8, which uses the pdf's of the baseform model *I_gw* and the surfaceform model *I_g* as examples. Originally, if a speaker pronounces /gw/ correctly, the acoustic probability of the model *I_gw* is higher than that of the surfaceform model *I_g*. Thus, correct recognition can be obtained. On the other hand, if one pronounces with variations such that the observation O_{gw} is closer to the surfaceform model, the output

probability of the surfaceform model *I_g* will be higher than that of the baseform model *I_gw*.

If the baseform model is modified by PM, the pdf of the refined baseform model *I_gw*' becomes a combination of the baseform and surfaceform mixture components. The output probability is now higher for the refined baseform model. Then the correct phoneme can be recognized.

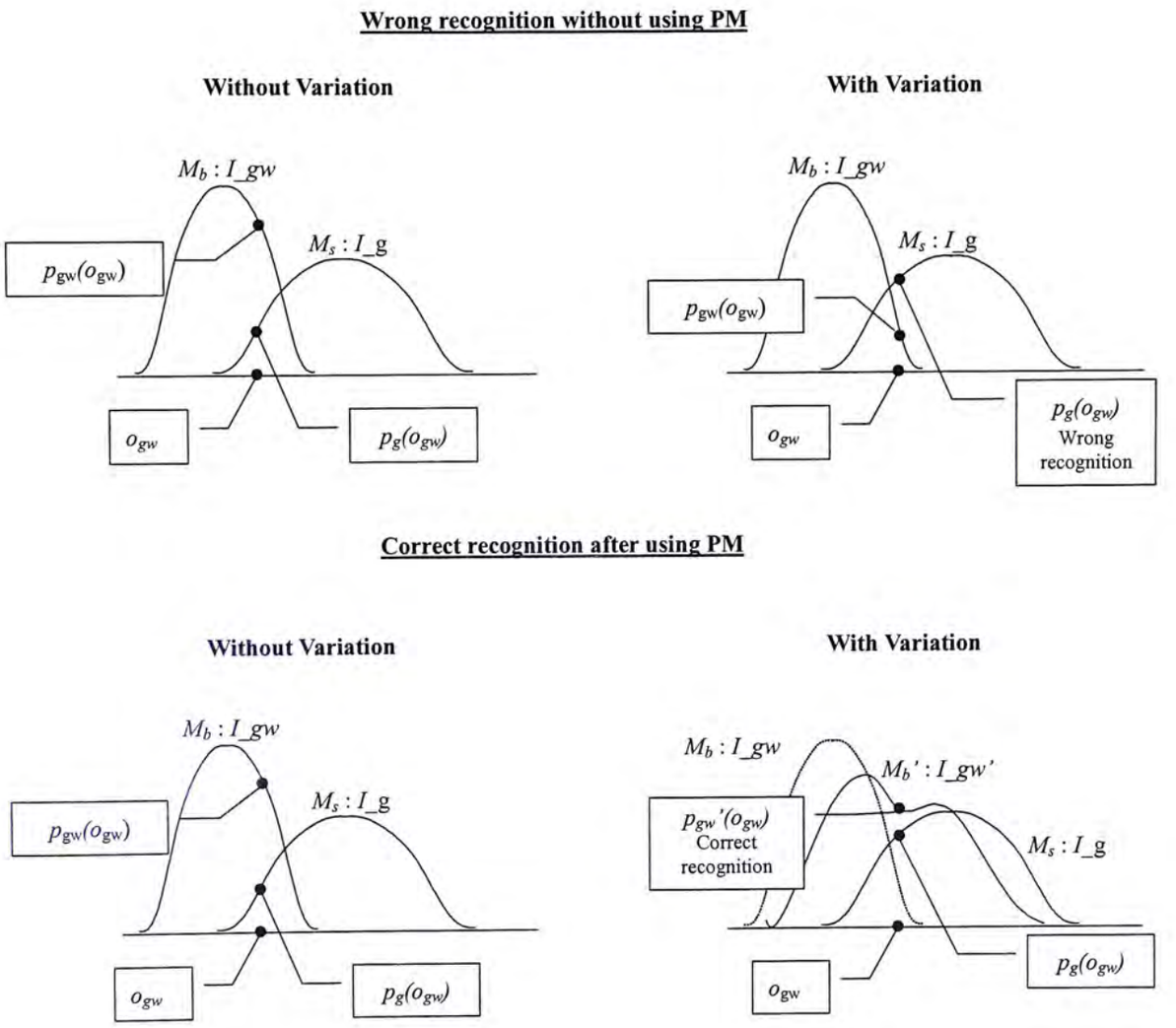


Figure 6.8: Improvement in recognizing the baseform /gw/ after mixture component sharing.

However, the incorporation of PM may also cause confusion when the surfaceform *s* is pronounced. The column "Confused M_s" in Table 6.3 shows that being the surfaceform model of others, degradation in recognition will be introduced. This can be explained as in Figure 6.9. /g/ is the surfaceform of /gw/. If one pronounces /g/ correctly, originally, the output probability is higher for the correct model *I_g*. If PM is used to modified *I_gw*, *I_gw'* will contain surfaceform mixtures of *I_g*. This will cause confusion, as the observation is now closer to the refine model *I_gw'*. Recognition error will be introduced.

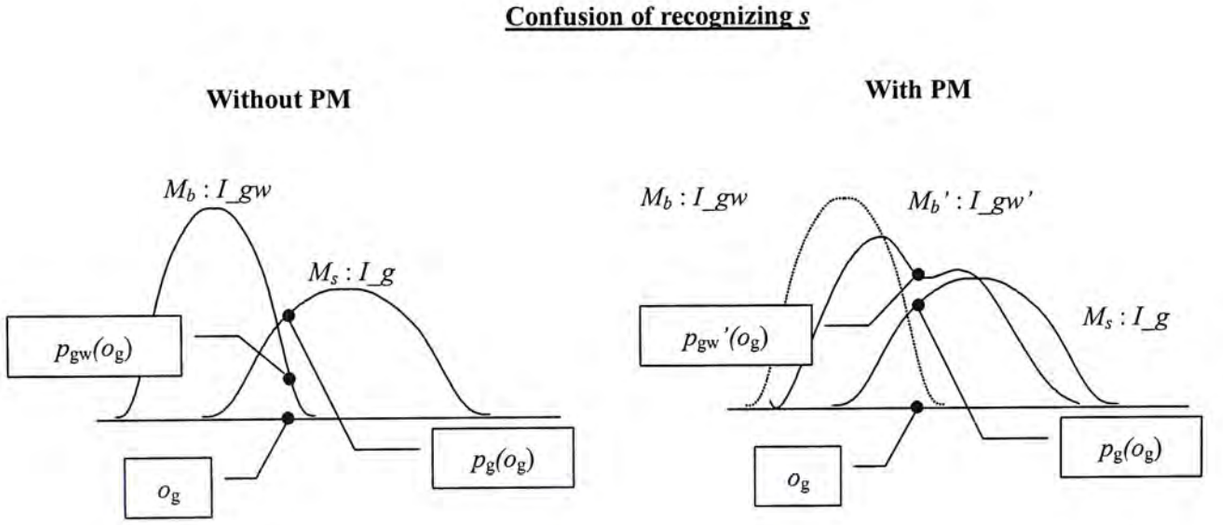


Figure 6.9: Degradation in recognizing the surfaceform /g/ after mixture component sharing.

6.6.2 Performance of Mixture Component Adaptation

Table 6.4 is a performance table for the second method "adaptation". In most cases, the refined models shown improvement in recognition accuracy. This can be explained as in Figure 6.10.

M_s used to refine M_b			Confused M_s		
M_b	M_s	Improve	M_b	M_s	Improve
F_aak	*F_aa/F_aak/ F_aat*	6 / 156	*F_aa*	*F_aa*	2 / 369
F_aang	*F_aan/F_aang/ F_an/F_ang*	3 / 20	*F_aan* *F_an*	*F_aan* *F_an*	-5 / 412 -2 / 383
F_aat	*F_aa/F_aat*	2 / 176			
F_ak	*F_ak/F_at*	3 / 90	*F_at*	*F_at*	2 / 333
F_ang	*F_an/F_ang*	17 / 191			
F_im	*F_im/F_in*	1 / 81	*F_in*	*F_in*	-4 / 508
F_ng	*F_m/F_ng*	10 / 90	*F_m*	*F_m*	0 / 6
F_ok	*F_o/F_ok*	2 / 237	*F_o*	*F_o*	2 / 396
F_on	*F_on/F_ong*	1 / 97	*F_ong*	*F_ong*	-3 / 419
I_gw	*I_g/I_gw*	7 / 226	*I_g*	*I_g*	-1 / 915
I_n	*I_l/I_n*	12 / 302	*I_l*	*I_l*	-16 / 735
I_ng	*I_ng/I_null*	19 / 257			
I_null	*I_ng/I_null*	0 /154			
Total		85 / 2316			

Table 6.4: Performance table for "adaptation".

Similar to the case of "sharing", if variations occur, the output probability of the surfaceform model *I_g* is higher than that of the baseform model *I_gw*. The pdf of the refined baseform model *I_gw*' will shift towards the surfaceform model. The output probability is now higher for *I_gw*'. Then the correct phoneme can be recognized.

On the other hand, if correct pronunciation /g/ is made. The pdf of the refined baseform model *I_gw*' will shift towards the surfaceform model. The observation is now closer to *I_gw*'. Recognition error will be introduced due to confusion.

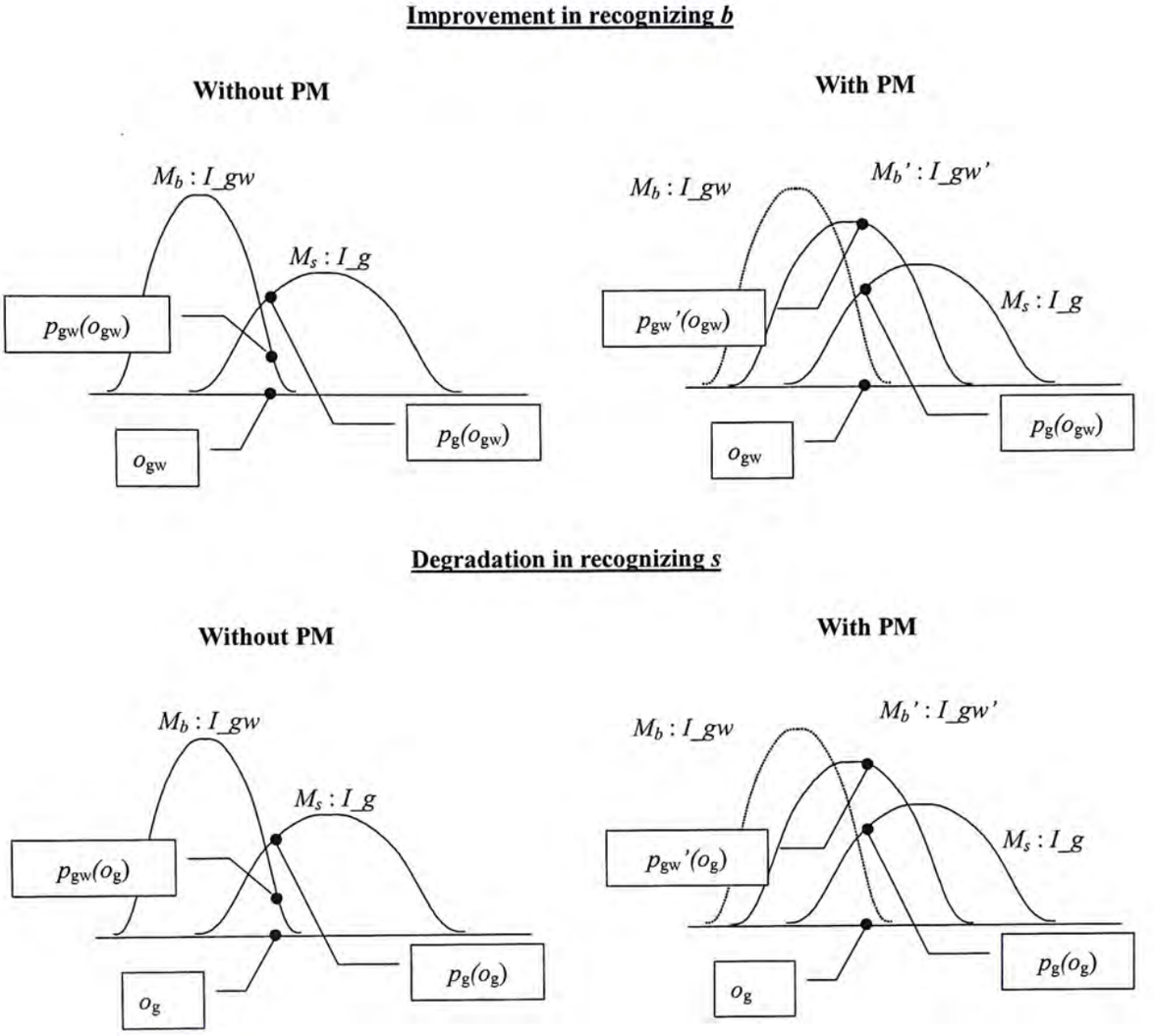


Figure 6.10: Improvement in recognizing *b* and degradation in recognizing *s* after mixture component adaptation.

6.7 Summary

In this chapter, we refine acoustic model for handling sound change in pronunciation by the methods of: sharing the surfaceform Gaussian mixture components with those of the baseform; adapting mixture components of the baseform towards those of the surfaceforms; selectively to share or to adapt the models using the distribution information of the KLD between mixture component pair of baseform and surfaceform.

There are two main types of distributions of KLD between baseform and surfaceform mixture component pairs, one is showing consistently small KLD (<50), while the others showing a wider range of KLD. They actually reflect different types

of pronunciation variations. Small KLD is usually found when a vowel nucleus remains unchanged or a consonant Initial/coda interchanged with another phoneme in the same phone class. Wide-Range KLD are found when a vowel nucleus is changed or a stop coda deleted or velar lip-rounded Initials mixed with velar Initials.

Reference

[1] M. Saraclar *et al*, "Pronunciation Ambiguity VS Pronunciation Variability in Speech Recognition", in *Proceedings of ICASSP-00*, Vol.3, pp.1679-1682, Istanbul, 2000.

[2] Y. Liu, "Pronunciation Modeling for Spontaneous Mandarin Speech Recognition", *Ph.D. Thesis*, The Hong Kong University of Science and Technology, 2002.

[3] M. Saraclar *et al*, "Pronunciation Modeling by Sharing Gaussian Densities across Phonetic Models", in *Proceedings of Eurospeech-99*, Vol.1, pp.515-518, Hungary, 1999.

[4] T.A. Myrvoll *et al*, "Optimal Clustering of Multivariate Normal Distributions Using Divergence and its Application to HMM Adaptation", in *Proceedings of ICASSP-03*, Vol.1, pp.552-555, Hong Kong, 2003.

Chapter 7

Pronunciation Modeling at Decoding Level

Due to co-articulation, the pronunciation of a phoneme is affected by its neighboring context. Pronunciation modeling at lexical level and acoustic model level makes use of IF confusion matrix for context-independent prediction of pronunciation variations. Context-dependent variations caused by co-articulation are not considered. Augmenting the lexicon with pronunciation variants can handle only intra-word variation, i.e. context-dependent variations within the word. In order to deal with context-dependent cross-word variations, we make an attempt to use PM at decoding level.

Both phone change and sound change can be handled at decoding level. In the case of phone change, the search space is expanded dynamically to include variation information during sentence decoding. As for sound change, the computation of acoustic scores during the search process can be modified to take into account the surfaceform information.

7.1 Search Process in Cantonese LVCSR

The decoding process of a speech recognizer aims at finding a sequence of words whose corresponding acoustic and language models best match the input signal. The search process in our baseline Cantonese LVCSR system is a one-pass search [1]. The pronunciation lexicon, the acoustic model and the language model are used to form a search space from which the most likely word sequence is decoded. The

search space is a lexical tree constructed based on the baseform lexicon. The lexical tree specifies all legitimate connections for the baseform bi-IF HMMs. A branch of the lexical tree is shown in Figure 7.1. Each node on the lexical tree represents a base IF which corresponds to all context-dependent HMMs with this base IF as the core. For example, in Figure 7.1, the node with base IF /ang/ carries the bi-IF HMMs of *F_ang+I_z*, *F_ang+I_s* and *F_ang+I_g*.

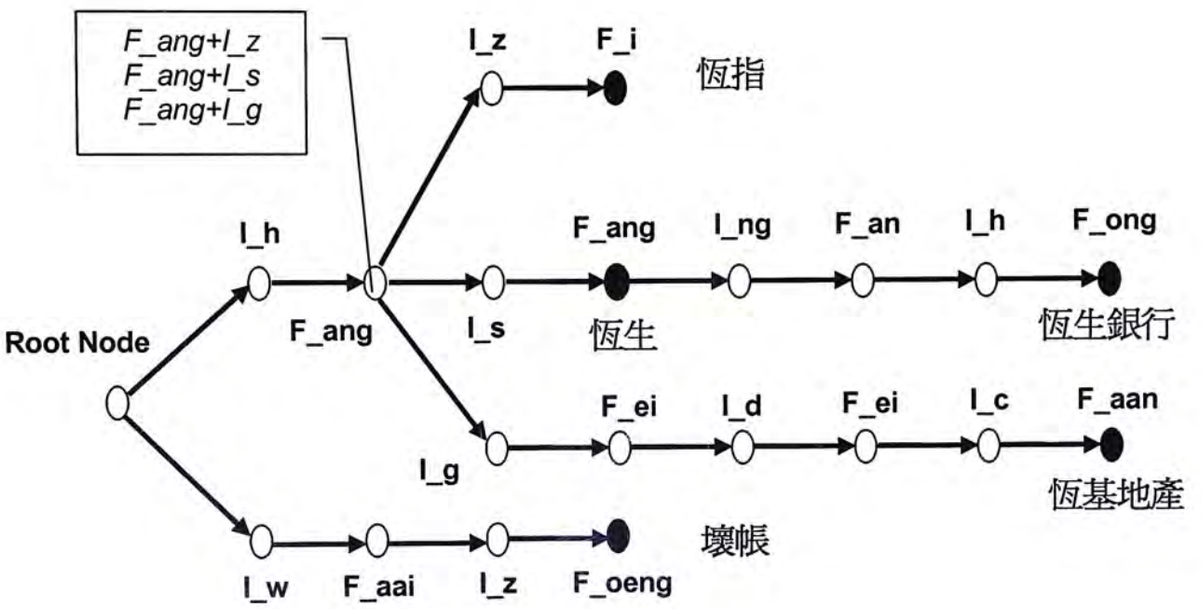


Figure 7.1: A branch of the lexical tree contructed by the baseform lexicon.

Forward Viterbi search is adopted in the LVCSR system. It is a token-based search process. A token is defined with the following items: node identity, path score, path history and the corresponding acoustic model. During the search, each token represents a search path reaching a particular lexical node. The propagation of tokens follows the paths defined in the lexical tree in which only the paths connecting consecutive bi-IF HMMs are activated. Bi-gram language model is used and a word record is created whenever a search path reaches a word-end node. A word lattice is resulted from this Viterbi search. When the utterance end is reached, the most probable word sequence is obtained by back-tracing the best path from the word lattice.

7.2 Model-Level Search Space Expansion

To deal with context-dependent inter-word pronunciation variations, some researchers suggested to add a group of multi-words into the lexicon [2][3]. However, this method can only handle a limited number of these variations. Another way is to incorporate PM at decoding level [4].

To incorporate PM at decoding level, the surfaceform pronunciation dictionary is not needed. The search works all the way with the baseform lexicon. Thus, the search space is the same as the baseline system. On the other hand, the search process is modified in the way that the number of alive tokens is increased to account for the pronunciation variations. Each bi-IF connection is expanded to the predicted surfaceform dynamically during the search. The paths leading to alternative pronunciations are also allowed to propagate in the search process.

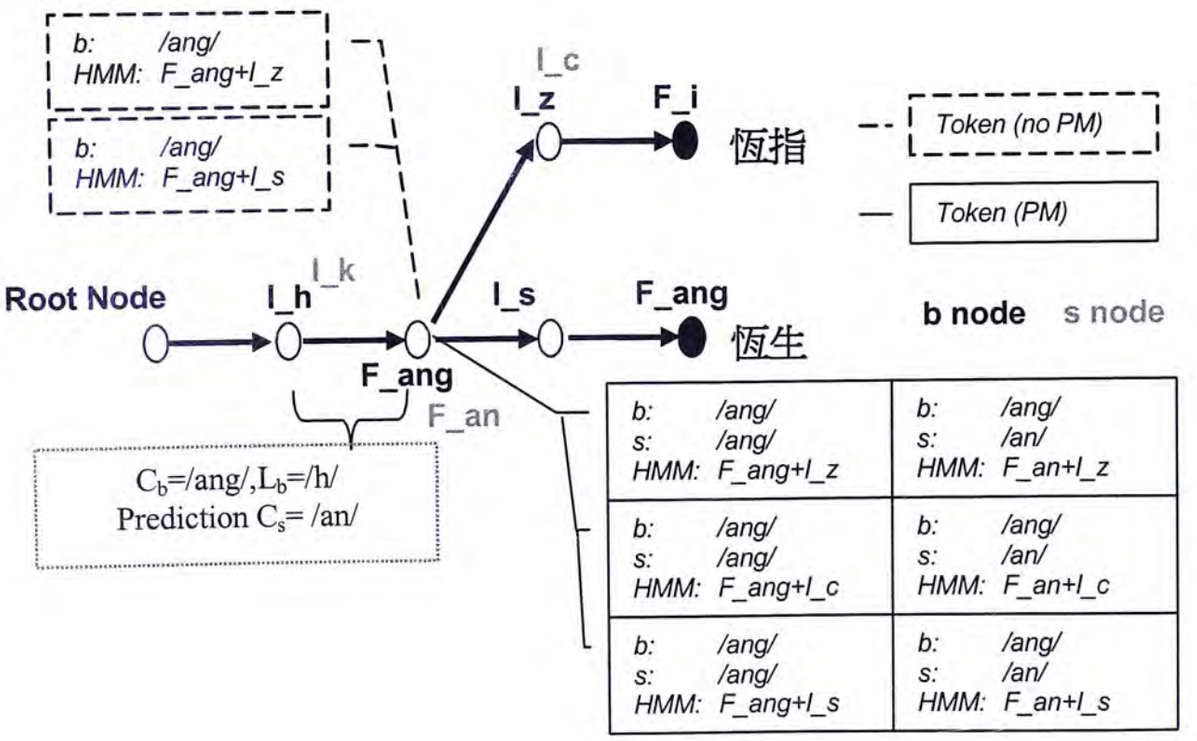


Figure 7.2: Token expansion with the incorporation of PM.

Figure 7.2 explains the operation of token expansion in the modified search process. In the baseline system, there are two nodes (**I_z** and **I_s**) connected to the

node **F_ang**. Therefore, two bi-IF HMMs are stored in this node and thus two alive tokens are recorded at the node **F_ang**.

A set of context-dependent decision tree pronunciation models (DTPM) as described in Section 4.2.2 is used to predict the surfaceform IF (s) from the baseform IF (b) and its context during decoding. It should be noted that the right context is not yet known in the forward Viterbi search. Therefore, *left context-dependent decision tree (LCDDT)* and *left phonetic class decision tree (LPCDT)* are used.

With the DTPM, predictions can be made with the prior knowledge of the current baseform IF (C_b) and the left baseform context (L_b). For example, in Figure 7.2, given the contextual information (C_b = /ang/, L_b = /h/), a surfaceform /an/ is predicted for the baseform node **F_ang**. The nodes **I_h**, **F_ang** and **I_z** have the surfaceforms /k/, /an/ and /c/ respectively.

The incorporation of PM increases the number of alive tokens. Apart from the original tokens carrying the baseform information, additional tokens are created to carry the surfaceform information. For example, two tokens at node **F_ang** are expanded to six tokens. With these additional tokens, each bi-IF connection is modified to allow the paths propagate to alternative pronunciations.

The search process is modified to find the word sequence W that maximizes the probability contributed by the surfaceform acoustic likelihood $P(O|S_k)$, variation probability $P(S_k|B)$ and the language model $P(W)$, i.e.

$$W^* = \arg\max_W P(W)P(O \mid S_k)P(S_k \mid B)P(B \mid W) \qquad (7.1)$$

where S_k is the k-th pronunciation variant sequence for the baseform sub-word sequence B. The probability $P(S_k|B)$ is obtained from the prediction using LCDDT or LPCDT.

7.3 State-Level Output Probability Modification

To deal with sound change, pronunciation modeling must be applied at a sub-model level. Acoustic model is usually trained with only assuming the baseform pronunciation without considering alternative pronunciations. In the previous chapter, we have discussed the refinement of the acoustic model by taking into account realistic pronunciations such that they can better represent the variations of speech sounds. However, those approaches can only handle context-independent sound change. In the following, the method "State-Level Output Probability Modification" is investigated to deal with context-dependent sound change.

The incorporation of PM in the search process neither change the original search space nor increase the number of alive tokens to carry the information of pronunciation variations. Instead, it is the way of computing acoustic score to be modified to take into account the easily confused surfaceform states.

During the search process, the acoustic score is given by the state output probability $p_j(o_t)$, which is a mixture of Gaussian distributions as follows

$$p_j(o_t) = \sum_{m=1}^{M} w_{jm} N(o_t; \mu_{jm}, \Sigma_{jm}) \qquad (7.2)$$

With the DTPM, predictions can be made with the prior knowledge of current baseform IF (C_b) and left baseform context (L_b). Given the predicted surfaceforms, the way of computing the state output probability is modified as

$$p_j'(o_t) = P(s_k = b \mid b) \cdot p_j(o_t) + \sum_{\substack{k=1 \\ s_k \neq b}}^{K} P(s_k \mid b) \cdot q_{kj}(o_t) \qquad (7.3)$$

where K is the number of surfaceform pronunciations for a particular baseform pronunciation b, $P(s_k=b|b)$ and $P(s_k|b)$ are the VPs obtained from the DTPM given the left context. $q_{kj}(o_t)$ is the state output probability of the k-th predicted surfaceform s_k state j.

This method is very similar to the method of "Sharing of Mixture Components" as described in Section 6.2. The only difference is that PM is dynamically added during the decoding process with the context information. Different context may give rise to different predictions in this method.

7.4 Recognition Experiments

7.4.1 Experiment 1 — Model-Level Search Space Expansion

In this experiment, the method "Model-Level Search Space Expansion" is evaluated in the LVCSR task. The search engine is modified such that DTPM is added dynamically during the search process to handle phone change.

Experimental Results:

Table 7.1 shows the recognition results of "Model-Level Search Space Expansion" using LCDDT and LPCDT with VP Th = 0.05 and 0.2. It can be seen that the incorporation of PM in decoding process improves the recognition performance. As the surfaceform path does not exist in the baseline system, the correct word sequence can never be retrieved for the utterance with phone change. This method generates the surfaceform paths by using LCDDT and LPCDT. The correct word sequence can be retrieved and a better performance can be obtained.

	Baseline	LCDDT VP Th 0.05	LCDDT VP Th 0.2	LPCDT VP Th 0.05	LPCDT VP Th 0.2
WER (%)	25.34	23.53	23.27	23.66	23.29
Relative WER Reduction (%)		7.14	8.17	6.63	8.09

Table 7.1: WER(%) of LVCSR task with "Model-Level Search Space Expansion" using LCDDT/LPCDT.

Unlike the results attained with the lexical-level approach, better performance is observed for VP Th = 0.2 than VP Th = 0.05. This is because the

pronunciation model used is context-dependent DTPM. The number of parameters needed for training a DTPM is more than that for training a context-independent IF confusion matrix (CM). This suggests that the amount of PM training data from CUSENT is probably enough to train a reliable CM but may not be adequate to train a precise DTPM. Rare predictions from DTPM should be pruned to improve the reliability. This account for the better performance with VP Th = 0.2.

As described above, the amount of training data to train LCDDT is inadequate. Intuitively, LPCDT should perform better than LCDDT. Nevertheless, it is found that the performance of LPCDT and LCDDT does not show a notable difference. This contradicting result can be explained by the aid of the acoustic model and the language model. "Model-Level Search Space Expansion" only expands the search space according to the DTPM to include more paths leading to different surfaceforms. However, the choice of these additional paths also depends on the AM and LM. Though LCDDT may produce more unreliable paths than LPCDT, as long as the true path is there, the true path is still possible to be selected by the decoder with the aids of AM and LM. Therefore, LPCDT only achieves a similar performance as LCDDT.

7.4.2 Experiment 2 — State-Level Output Probability Modification

In this experiment, the method "State-Level Output Probability Modification" is evaluated in the LVCSR task. The search process is modified in such a way that the calculation of the state output probability takes into account the predicted surfaceform.

Experimental Results:

Table 7.2 shows the recognition results of "State-Level Output Probability Modification" using LCDDT and LPCDT with VP Th = 0.05 and 0.2. It is found that using both LCDDT and LPCDT for this method does not affect the recognition

accuracy that much. It even deteriorates the recognition performance for VP Th = 0.05.

	Baseline	LCDDT VP Th 0.05	LCDDT VP Th 0.20	LPCDT VP Th 0.05	LPCDT VP Th 0.20
WER (%)	25.34	27.21	25.07	27.13	25.05
Relative WER Reduction (%)		-7.38	1.07	-7.06	1.14

Table 7.2: WER(%) of LVCSR task with "State-Level Output Probability Modification" using LCDDT/LPCDT.

The performance of this approach is not comparable with that obtained in Chapter 6, in which the method "Sharing of Mixture Components" can be regarded as its context-independent counterpart. The observation can be explained mainly by two reasons, the amount of PM training data and the contribution of the far-away Gaussian mixtures in the surfaceform HMM.

As we mentioned in the previous section, the amount of training data may not be enough to train a set of reliable context-dependent DTPM. Therefore, not all the predicted surfaceform HMMs are appropriate to be included in the computation of the state output probability for the baseform model. Unlike "Model-Level Search Space Expansion" where only one of the predicted surfaceform is applied for each token, all the Gaussian mixtures of all the predicted surfaceform HMMs are used in calculating the state output probability. The unreliable predictions may lead to a smaller state output probability to the baseform model, which will result in incorrect recognition. Therefore, the performance obtained with VP Th = 0.2 is better than VP Th = 0.05 as more unreliable predictions are filtered. The result is not comparable to the one achieved in "Sharing of Mixture Components" in Section 6.2. It suggests that the context-independent CM is more reliable than the context-dependent DTPM. As we use the same training data for both, the one with fewer parameters, i.e. context-independent CM would be more reliable.

Moreover, in "Sharing of Mixture Components", a KLD threshold is imposed so that the Gaussian mixtures of the surfaceform model which are far away from that of the baseform model are not considered in the calculation of the state output probability. However, we did not filter these "far-away" Gaussian mixtures in

the method described in the previous section. These mixtures will lower the state output probability. Therefore, this method only shows little improvement.

7.5 Summary

In this chapter, we discussed the incorporation of PM at decoding level in order to handle context-dependent, inter-word phone change and context-dependent sound change. We proposed the method "Model-Level Search Space Expansion" to handle phone change. The search space is expanded dynamically by increasing the number of alive tokens to contain variation information during sentence decoding. To deal with sound change, we proposed the method, "State-Level Output Probability Modification". The calculation of the state output probability in the search process is modified to consider also the surfaceform information.

Reference

[1] W.N. Choi, "An Efficient Decoding Method for Continuous Speech Recognition Based on a Tree-Structured Lexicon", *M.Phil. Thesis,* The Chinese University of Hong Kong, 2001.

[2] T. Slobada *et al*, "Dictionary learning for spontaneous speech recognition", in *Proceedings of ICSLP-96*, Vol.4, pp.2328-2331, Philadelphia, 1996.

[3] T. Shinozaki *et al*, "A New Lexicon Optimization Method for LVCSR Based on Linguistic and Acoustic Characteristics of Words", in *Proceedings of ICSLP-02*, Vol.1, pp.717-720, Denver, 2002.

[4] P. Kam *et al*, "Modeling Pronunciation Variation for Cantonese Speech Recognition", in *Proceedings of PMLA-02*, pp.12-17, Denver, 2002.

Chapter 8

Conclusions and Suggestions for Future Work

8.1 Conclusions

The work described in this thesis contributes to the study of pronunciation variations in continuous Cantonese speech. Different approaches of incorporating pronunciation models (PM) into Cantonese ASR system are investigated and analyzed with comprehensive experimental results.

Phone change and sound change are the two major types of pronunciation variations that we have considered. Phone change is the complete change of a baseform phoneme to another surfaceform phoneme. Phone change can be modeled by providing the baseform phoneme together with the surfaceform realization either at lexical level or decoding level. Sound change is caused by pronunciation ambiguity between the baseform phoneme and the surfaceform phoneme. Sound change can be handled by modifying the parameters of baseform models at acoustic model level or by including the surfaceform parameters in the computation of baseform state output probability at decoding level.

In our study, pronunciation model is a set of phone level pronunciation models (PLPM) trained from a large speech corpus. The PLPMs include context-independent IF confusion matrix (CM) and context-dependent decision tree pronunciation model (DTPM). PLPMs are integrated with different knowledge

sources, including the lexicon and the acoustic model, and the decoder in the ASR system.

The baseform lexicon is augmented with variation information obtained from CM to build a pronunciation variation dictionary (PVD). By replacing the baseform transcription of each word in the lexicon by the surfaceform transcriptions, phone change can be handled. It is found that the recognition performance can be further improved by pruning the PVD with word unigram information. This verifies that words with a small unigram tend to have fewer variations in their pronunciations. Performance analysis is performed in order to gain insight into the process of pronunciation modeling. The analysis shows that although incorporating pronunciation model leads to improvement on recognition accuracy, it may also deteriorate the recognition performance in some cases. This is due to the increased confusability when the number of characters represented by a pronunciation is increased. It is found that the inclusion of only a few common variations would lead to significant performance improvement. These variations are /ang/→/an/, /ng/→/m/, /gw/→/g/, /n/→/l/ and /ng/→/null/. It can be said that such a simple modification is already fairly effective to deal with phone change in Cantonese read speech.

The acoustic model is refined by CM to deal with sound change. In our work, the approaches of sharing or adaptation of Gaussian mixture components are investigated. Sharing of mixture components supplements the baseform model to better represent the realistic pronunciations. However, both model size and computation are increased. Including mixture components of similar surfaceform model in the modified baseform model is also redundant. Adaptation of mixture components effectively adjusts the baseform model to be acoustically more accurate. Performance improvement is observed in the case that the Gaussian mixture components of the baseform and the surfaceform models are close to each other. Then, we combine the two approaches in such a way that the surfaceform components that are close to the baseform are used for adaptation, and relatively distant components are used for sharing. Again, it is found that the refinement of only a few models can give improvement. They are /ang/~/an/, /ng/~/m/, /gw/~/g/, /n/~/l/, /ng/~/null/. However, the incorporation of PM may also cause confusion when the surfaceform is pronounced.

The decoding algorithm is modified such that dynamic search space expansion is allowed during the search process to handle phone change. The calculation of state output probability could take into account the variation information to handle sound change. When dealing with phone change using search space expansion, the recognition accuracy is increased as the search paths with surfaceform pronunciations are allowed. In addition, cross-word context-dependent DTPM can be applied at decoding level. More sophisticated PM can be used in this level as opposed to lexical level or acoustic model level. However, only minimal improvement is obtained in handling sound change at decoding level. This suggests that context-dependent PM may not be as good as context independent PM in dealing with sound change when the number of PM training data is not enough.

In general, the improvement for dealing with phone change is higher than that for sound change. Indeed, it is very difficult to accurately define sound change. The causes leading to sound change are not easily identified. It is difficult to tackle sound change effectively. The methods proposed in this thesis may not be effective to deal with the so-called sound change.

8.2 Suggestions for Future Work

Pronunciation modeling is a large research topic. Though pronunciation modeling at different levels of Cantonese LVCSR is investigated in this research, many have to be done in order to handle other types of variation in speech. Here, we present some suggestions to improve the current system.

1. Accent Speech and Spontaneous Speech

In this research, the framework for the incorporation of PM in different components of the ASR was discussed. Also, the evaluation techniques are developed such that we can analysis the contribution of every variation. In this stage, only read speech is under testing. Spontaneous speech and accented speech will contain even more variations than read speech. Spontaneous speech being uttered with less concern in a less formal condition will contain more variations. Speakers would tend to preserve

only the most informative words [1][2]. Function word will be uttered with more variations. Also, fast speaking rates tend to coincide with significant phonological reduction [1]-[3]. Accented speech contains a lot of variations as native and non-native Cantonese speakers use different phone sets when they speak. Non-native speakers either do not know the exact phone should be used or they use phone belonging to the phone set of their mother tongue language to utter a Cantonese word. Therefore, both phone substitution and sound change will occur frequently.

Being provided the framework, it is possible to extend our works to deal with these kinds of variations.

2. Other Variations such as Deletion and Insertion

In the proposed framework, only the variations due to substitution of phoneme are handled. As mentioned in the previous paragraph, deletion and insertion are also found in spontaneous speech. Towards a practical system, it is crucial to handle the variations caused by deletions and insertions.

At lexical level, deletion/insertion could be handled by a PVD with word entries having more/fewer phonemes than the baseform pronunciation. Deletion may also be considered at acoustic model level by modifying the baseform model such that some of the HMM states could be skipped. At decoding level, deletion/insertion is handled by providing more paths in the search space.

3. Integration of Pronunciation Modeling at Different Levels

In this thesis, pronunciation modeling at lexical level, acoustic model level and decoding level are presented and compared. All of these methods show contributions to recognition accuracy. However, these methods are applied independently to the ASR system. It would be nice to have all the proposed methods be integrated in a single system. Intuitively, the system having all these features in pronunciation modeling could perform better as PM at each level shows improvement. On the other hand, some points should be noted in building such a system: 1) PM applied to

different levels may show redundancy, 2) dealing with phone change and sound change simultaneously may not be appropriate, etc.

Reference

[1] M.Y. Tsai *et al*, "Pronunciation Variation Analysis with respect to Various Linguistic Levels and Contextual Conditions for Mandarin Chinese", in *Proceedings of Eurospeech-01*, Vol.2, pp.1445-1448, Alborg, 2001.

[2] R.S. Bauer *et al*, *Trends in Linguistics, Studies and Monographs 102, Modern Cantonese Phonology*, Mouton de Gruyter, Berlin, New York, 1997.

[3] E. Fosler-Lussier *et al*, "Effects of speaking rate and word frequency on pronunciations in conversational speech", *Speech Communication*, Vol.29, pp.137-158, 1999.

Appendix I Base Syllable Table

Final	Initial																			
		b	C	d	f	g	gw	h	j	k	kw	l	m	n	ng	p	s	t	w	z
aa	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
aai	*	*	*	*	*	*	*	*				*	*	*	*	*	*	*	*	*
aak	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*		*	*
aam	*		*	*		*	*	*				*		*	*		*	*		*
aan	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
aang	*		*	*		*		*			*	*	*		*	*	*		*	*
aap	*	*	*	*		*		*				*		*	*		*	*		*
aat	*	*	*	*	*	*	*	*		*		*	*		*	*	*	*	*	*
aau	*	*	*		*	*		*	*	*		*	*	*	*	*	*	*		*
ai	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ak	*	*	*	*		*		*	*			*	*		*		*			*
am	*	*	*	*		*		*	*	*		*		*	*		*	*		*
an	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ang	*	*	*	*	*	*	*	*		*		*	*	*	*	*	*	*	*	*
ap			*	*		*		*	*	*		*		*			*	*		*
at	*	*	*	*	*	*	*	*	*	*		*	*		*	*	*		*	*
au	*		*	*	*	*		*	*	*		*	*	*	*		*	*		*
e		*	*	*	*	*			*	*		*	*	*		*	*	*		*
ei		*		*	*	*		*		*		*	*	*		*	*	*	*	
ek			*	*		*		*		*		*				*	*	*		*
eng		*	*	*		*		*	*			*	*	*		*	*	*		*
eoi			*	*		*		*	*	*		*		*			*	*		*
eon			*	*					*			*		*			*	*		*
eot			*									*					*			*
i			*	*					*			*	*	*			*	*		*
ik		*	*	*		*			*	*	*	*	*	*		*	*	*	*	*
im			*	*		*		*	*	*		*		*			*	*		*
in		*	*	*		*		*	*	*		*	*	*		*	*	*		*
ing		*	*	*		*	*	*	*	*		*	*	*		*	*	*	*	*
ip			*	*		*		*	*			*		*			*	*		*
it		*	*	*		*		*	*	*		*	*	*		*	*	*		*
iu		*	*	*		*		*	*	*		*	*	*		*	*	*		*
m	*																			
ng	*																			
o	*	*	*	*	*	*	*	*		*		*	*	*	*	*	*	*	*	*
oe			*	*		*		*	*	*		*		*			*	*		*
oek			*	*		*		*	*	*		*		*			*			*
oeng			*	*		*		*	*	*		*		*			*			*
oi	*		*	*		*		*		*		*		*	*		*	*		*
ok	*	*		*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
on	*		*	*	*	*	*	*		*		*	*		*	*	*	*	*	*
ong	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
ot						*		*												
ou	*	*	*	*		*		*		*		*	*	*	*	*	*	*		*
u		*			*	*				*		*	*						*	
ui		*			*					*			*			*			*	
uk	*	*	*	*	*	*		*	*	*		*	*	*	*	*	*	*		*
un		*		*	*	*		*					*			*			*	
ung		*	*	*	*	*		*	*	*		*	*	*		*	*	*		*
ut		*			*					*			*			*		*	*	
yu			*						*								*			*
yun			*	*		*		*	*	*		*		*			*	*		*
yut			*	*				*	*	*		*		*			*	*		*

Table I: Legitimate Initial/Final combinations for Cantonese base syllables.

Appendix II Cantonese Initials and Finals

LSHK	IPA	LSHK	IPA
b	p	j	j
d	t	m	m
g	k	n	n
gw	k^w	ng	ŋ
p	p^h	s	s
t	t^h	f	f
k	k^h	h	h
kw	k^{wh}	z	ts
l	l	c	ts^h
w	w		

Table II: List of Cantonese Initials.

LSHK	IPA	LSHK	IPA	LSHK	IPA
i	i	am	ɐm	ip	ip
yu	y	an	ɐn	it	it
u	u	ang	ɐŋ	ik	ɪk
e	ɛ	aam	am	yut	yt
oe	œ	aan	an	ut	ut
o	ɔ	aang	aŋ	uk	ʊk
aa	a	m	m	ek	ɛk
im	im	ng	ŋ	eot	ɵt
in	in	ui	ui	oek	œk
ing	ɪŋ	ei	ei	ot	ɔt
yun	yn	eoi	ɵy	ok	ɔk
un	un	oi	ɔi	ap	ɐp
ung	ʊŋ	ai	ɐi	at	ɐt
eng	ɛŋ	aai	ai	ak	ɐk
eon	ɵn	iu	iu	aap	ap
oeng	œŋ	ou	ɔu	aat	at
on	ɔn	au	ɐu	aak	ak
ong	ɔŋ	aau	au		

Table III: List of Cantonese Finals.

Appendix III IF confusion matrix

	b	c	d	f	g	gw	h	j	k	kw	l	m	n	ng	null	p	s	t	w	z
b	**94.66**	0.00	1.21	0.29	0.11	0.01	0.02	0.03	0.02	0.00	0.22	0.48	0.06	0.02	0.12	1.83	0.01	0.12	0.57	0.02
c	0.00	**93.87**	0.00	0.05	0.07	0.01	0.02	0.06	0.04	0.00	0.03	0.00	0.01	0.01	0.00	0.00	1.62	0.35	0.01	3.63
d	1.19	0.06	**93.31**	0.06	0.71	0.00	0.02	0.19	0.12	0.01	1.11	0.08	0.47	0.07	0.06	0.08	0.04	1.52	0.01	0.53
f	0.63	0.09	0.20	**97.69**	0.03	0.03	0.16	0.00	0.00	0.01	0.09	0.03	0.01	0.05	0.08	0.18	0.23	0.21	0.15	0.06
g	0.12	0.06	0.94	0.02	**93.64**	1.38	0.14	0.35	1.57	0.13	0.27	0.01	0.12	0.19	0.15	0.11	0.00	0.19	0.12	0.20
gw	0.18	0.00	0.21	0.18	28.85	**67.10**	0.03	0.00	0.40	0.91	0.09	0.00	0.06	0.12	0.21	0.09	0.00	0.15	0.85	0.06
h	0.01	0.08	0.03	0.06	0.09	0.02	**95.90**	0.39	0.41	0.17	0.29	0.15	0.08	0.27	0.55	0.39	0.03	0.39	0.15	0.01
j	0.06	0.19	0.14	0.03	0.36	0.00	0.10	**96.66**	0.09	0.00	0.57	0.07	0.17	0.10	0.12	0.01	0.05	0.14	0.08	0.32
k	0.01	0.32	0.18	0.04	2.51	0.09	0.33	0.10	**92.91**	1.02	0.06	0.00	0.05	0.08	0.03	0.21	0.03	1.60	0.09	0.14
kw	0.00	0.00	0.00	0.13	0.54	1.07	0.27	0.13	14.21	**82.71**	0.00	0.00	0.00	0.00	0.13	0.27	0.00	0.27	0.13	0.00
l	0.18	0.01	0.43	0.05	0.12	0.00	0.05	0.33	0.00	0.00	**96.06**	0.36	1.25	0.23	0.28	0.07	0.06	0.07	0.12	0.06
m	0.40	0.00	0.02	0.02	0.02	0.01	0.03	0.02	0.00	0.02	0.44	**96.90**	0.12	0.12	0.19	0.08	0.01	0.00	1.37	0.00
n	0.19	0.00	0.51	0.02	0.07	0.03	0.10	0.53	0.03	0.00	73.08	0.82	**22.07**	0.89	0.67	0.03	0.02	0.03	0.27	0.03
ng	0.19	0.02	0.13	0.13	0.42	0.13	1.45	0.88	0.40	0.04	1.77	2.00	0.55	**35.99**	52.83	0.10	0.02	0.04	1.43	0.02
null	0.15	0.18	0.06	0.00	0.18	0.00	0.49	0.58	0.12	0.12	0.80	0.80	0.31	6.59	**85.41**	0.06	0.03	0.00	0.98	0.06
p	3.43	0.06	0.14	0.42	0.14	0.00	0.30	0.00	0.26	0.12	0.10	0.28	0.06	0.04	0.06	**92.47**	0.02	1.66	0.26	0.02
s	0.00	0.70	0.05	0.31	0.02	0.00	0.02	0.02	0.00	0.00	0.05	0.00	0.01	0.00	0.02	0.01	**97.74**	0.03	0.01	0.89
t	0.06	0.61	2.04	0.02	0.26	0.00	0.36	0.04	1.07	0.01	0.41	0.02	0.22	0.08	0.09	0.82	0.10	**93.22**	0.04	0.32
w	0.57	0.01	0.01	0.07	4.35	0.64	0.09	0.02	0.19	0.15	0.20	1.67	0.01	0.16	0.22	0.14	0.02	0.02	**91.10**	0.05
z	0.00	2.89	0.27	0.05	0.15	0.00	0.01	0.71	0.03	0.00	0.08	0.01	0.03	0.01	0.01	0.00	1.55	0.13	0.01	**93.90**

Table IV: Confusion matrix for Initials

am	ak	ai	aau	aat	aap	aang	aan	aam	aak	aai	aa	
0.08	0.91	0.11	3.37	8.70	2.58	1.83	2.00	1.29	9.59	2.06	**90.39**	**aa**
0.00	0.00	1.27	0.00	0.88	0.00	0.46	0.46	0.00	0.82	**92.27**	1.12	**aai**
0.00	4.33	0.04	0.08	2.77	4.74	1.14	0.34	0.26	**71.91**	0.29	1.39	**aak**
1.26	0.00	0.00	0.12	0.21	1.16	4.79	1.10	**90.02**	0.26	0.00	0.43	**aam**
0.16	0.28	0.03	0.16	1.26	0.08	17.81	**90.82**	3.24	0.97	0.60	1.06	**aan**
0.04	0.00	0.00	0.00	0.08	0.00	**56.85**	1.00	0.70	0.07	0.05	0.08	**aang**
0.08	0.46	0.01	0.12	0.46	**78.97**	0.00	0.04	0.44	1.99	0.07	0.12	**aap**
0.00	1.48	0.05	0.04	**78.74**	1.91	0.00	0.26	0.15	7.49	0.22	1.63	**aat**
0.00	0.17	0.00	**90.76**	0.13	0.25	0.00	0.06	0.07	0.41	0.02	0.40	**aau**
0.00	0.17	**93.00**	0.00	0.17	0.00	0.00	0.04	0.00	0.19	1.75	0.07	**ai**
0.04	**68.17**	0.05	0.04	0.17	0.33	0.00	0.02	0.00	0.79	0.05	0.13	**ak**
94.81	0.06	0.00	0.00	0.04	0.00	0.68	0.02	1.36	0.07	0.00	0.05	**am**
0.83	0.40	0.05	0.00	0.08	0.08	5.48	1.34	0.26	0.04	0.05	0.04	**an**
0.51	0.06	0.00	0.08	0.08	0.00	6.62	0.16	0.26	0.07	0.00	0.01	**ang**
0.28	2.33	0.01	0.04	0.17	5.82	0.00	0.00	0.07	0.37	0.00	0.13	**ap**
0.08	14.24	0.30	0.20	2.35	0.83	0.00	0.08	0.04	1.69	0.02	0.45	**at**
0.20	0.34	0.01	3.45	0.21	0.58	0.23	0.04	0.48	0.11	0.00	0.25	**au**
0.00	0.23	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.02	**e**
0.00	0.11	0.82	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	**ei**
0.00	0.40	0.07	0.00	0.04	0.00	0.00	0.02	0.00	0.04	0.02	0.01	**ek**
0.00	0.06	0.04	0.00	0.00	0.00	0.23	0.02	0.00	0.11	0.00	0.01	**eng**
0.00	0.34	1.78	0.04	0.00	0.00	0.00	0.02	0.00	0.00	0.22	0.01	**eoi**
0.20	0.11	0.00	0.00	0.00	0.00	0.46	0.10	0.00	0.00	0.00	0.00	**eon**
0.00	1.94	0.03	0.00	0.17	0.17	0.00	0.00	0.00	0.11	0.02	0.05	**eot**
0.00	0.00	0.11	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	**i**
0.04	0.68	0.20	0.00	0.00	0.00	0.23	0.00	0.00	0.04	0.02	0.02	**ik**
0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.04	0.00	0.02	0.00	**im**
0.00	0.00	0.03	0.04	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	**in**
0.00	0.11	0.09	0.00	0.00	0.00	0.68	0.02	0.00	0.00	0.00	0.00	**ing**
0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**ip**
0.00	0.00	0.05	0.00	0.76	0.00	0.00	0.00	0.00	0.00	0.10	0.01	**it**
0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**iu**
0.20	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.04	0.00	0.00	0.00	**m**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	**ng**
0.12	0.23	0.01	0.16	0.55	0.25	0.23	0.10	0.07	0.19	0.05	0.76	**o**
0.00	0.00	0.01	0.00	0.00	0.08	0.00	0.02	0.00	0.00	0.00	0.00	**oe**
0.00	0.68	0.04	0.00	0.25	0.08	0.00	0.00	0.00	0.30	0.07	0.05	**oek**
0.28	0.17	0.17	0.12	0.17	0.17	0.68	0.48	0.48	0.45	0.22	0.20	**oeng**
0.00	0.06	0.48	0.20	0.08	0.00	0.00	0.14	0.00	0.11	0.98	0.27	**oi**
0.08	0.40	0.01	0.04	0.67	1.08	0.00	0.00	0.00	1.05	0.02	0.11	**ok**
0.00	0.00	0.01	0.00	0.00	0.00	0.46	0.40	0.04	0.04	0.00	0.00	**on**
0.20	0.06	0.00	0.12	0.08	0.17	0.23	0.30	0.44	0.11	0.00	0.13	**ong**
0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**ot**
0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**ou**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	**u**
0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**ui**
0.00	0.57	0.00	0.08	0.04	0.00	0.00	0.00	0.00	0.04	0.00	0.00	**uk**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	**un**
0.35	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0.04	0.00	0.00	0.00	**ung**
0.00	0.00	0.01	0.04	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	**ut**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	**yu**
0.00	0.00	0.03	0.04	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.01	**yun**
0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**yut**

eot	eon	eoi	eng	ek	ei	e	au	at	ap	ang	an	
0.06	0.00	0.01	0.00	0.00	0.01	0.12	0.53	0.40	0.21	0.21	0.07	**aa**
0.00	0.00	0.04	0.00	0.16	0.01	0.06	0.02	0.02	0.03	0.00	0.04	**aai**
0.06	0.00	0.00	0.00	0.48	0.00	0.00	0.02	0.36	0.24	0.06	0.04	**aak**
0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.38	0.17	**aam**
0.00	0.16	0.03	0.15	0.32	0.01	0.00	0.02	0.12	0.03	2.65	1.27	**aan**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	2.83	0.03	**aang**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	1.31	0.06	0.00	**aap**
0.06	0.00	0.00	0.00	0.16	0.00	0.06	0.05	0.83	0.10	0.03	0.04	**aat**
0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.95	0.02	0.08	0.06	0.00	**aau**
0.00	0.00	0.61	0.00	0.80	0.57	0.18	0.06	0.20	0.00	0.09	0.01	**ai**
0.06	0.00	0.00	0.00	0.32	0.00	0.02	0.09	1.74	0.39	0.21	0.10	**ak**
0.00	0.05	0.01	0.00	0.00	0.00	0.00	0.06	0.13	0.63	1.15	0.73	**am**
0.11	1.98	0.01	0.31	0.00	0.00	0.00	0.01	0.55	0.08	22.89	**90.56**	**an**
0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.05	0.03	0.05	**61.56**	2.17	**ang**
0.23	0.00	0.00	0.00	0.00	0.03	0.00	0.22	1.34	**91.56**	0.12	0.03	**ap**
1.54	0.16	0.06	0.00	1.27	0.02	0.14	0.34	**89.12**	1.91	0.27	0.55	**at**
0.00	0.00	0.09	0.00	0.00	0.00	0.00	**93.93**	0.23	0.84	0.35	0.08	**au**
0.29	0.05	0.36	2.00	10.51	1.20	**94.11**	0.08	0.38	0.03	0.00	0.03	**e**
0.17	0.05	0.33	0.00	0.16	**96.26**	1.73	0.00	0.13	0.05	0.00	0.01	**ei**
0.06	0.00	0.01	0.15	**80.73**	0.02	0.16	0.00	0.18	0.08	0.00	0.07	**ek**
0.00	0.00	0.01	**83.82**	0.64	0.03	0.16	0.00	0.02	0.00	0.18	0.14	**eng**
1.37	0.43	**95.79**	0.00	0.16	0.37	0.35	0.09	0.22	0.03	0.00	0.06	**eoi**
0.46	**91.29**	0.04	0.31	0.00	0.00	0.02	0.00	0.02	0.03	1.12	1.19	**eon**
91.50	0.11	0.29	0.00	0.16	0.00	0.02	0.03	0.83	0.29	0.00	0.04	**eot**
0.00	0.00	0.03	0.00	0.00	0.25	1.65	0.00	0.15	0.00	0.00	0.00	**i**
0.40	0.05	0.07	0.15	2.07	0.20	0.14	0.00	0.95	0.05	0.00	0.04	**ik**
0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	**im**
0.00	0.00	0.04	0.00	0.16	0.08	0.02	0.01	0.02	0.00	0.03	0.04	**in**
0.00	0.16	0.01	9.24	0.16	0.09	0.04	0.00	0.02	0.00	0.41	0.32	**ing**
0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	**ip**
0.00	0.00	0.01	0.00	0.00	0.16	0.04	0.00	0.00	0.00	0.00	0.01	**it**
0.00	0.00	0.06	0.00	0.00	0.01	0.06	0.09	0.07	0.03	0.03	0.01	**iu**
0.06	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.09	0.10	**m**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	**ng**
0.00	0.05	0.06	0.00	0.00	0.00	0.00	0.48	0.18	0.08	0.00	0.03	**o**
0.11	0.00	0.10	0.00	0.00	0.00	0.06	0.02	0.00	0.00	0.00	0.00	**oe**
0.80	0.05	0.12	0.00	0.96	0.00	0.06	0.08	0.40	0.18	0.06	0.04	**oek**
0.57	3.21	0.32	1.85	0.48	0.02	0.12	0.15	0.22	0.21	2.42	1.13	**oeng**
0.17	0.05	0.50	0.00	0.00	0.01	0.00	0.13	0.07	0.03	0.00	0.04	**oi**
0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.07	0.08	0.37	0.00	0.03	**ok**
0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.10	**on**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.05	0.00	1.56	0.22	**ong**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**ot**
0.00	0.00	0.04	0.00	0.00	0.02	0.00	1.41	0.02	0.08	0.00	0.00	**ou**
0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	**u**
0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	**ui**
0.80	0.00	0.04	0.00	0.00	0.01	0.02	0.17	0.08	0.50	0.03	0.03	**uk**
0.00	0.11	0.06	0.00	0.00	0.00	0.00	0.06	0.02	0.00	0.03	0.10	**un**
0.00	0.91	0.03	0.00	0.00	0.00	0.00	0.06	0.00	0.03	0.38	0.01	**ung**
0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.07	0.00	0.00	0.00	**ut**
0.06	0.00	0.36	0.00	0.00	0.11	0.08	0.00	0.02	0.00	0.00	0.00	**yu**
0.23	0.37	0.04	0.00	0.00	0.04	0.04	0.00	0.02	0.00	0.00	0.00	**yun**
0.34	0.00	0.06	0.00	0.00	0.03	0.10	0.00	0.02	0.00	0.00	0.01	**yut**

oe	o	ng	m	iu	it	ip	ing	in	im	ik	i	
1.61	0.62	0.00	0.00	0.02	0.06	0.00	0.00	0.00	0.00	0.00	0.00	aa
0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	aai
0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	aak
0.00	0.04	0.11	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	aam
0.00	0.01	0.00	0.70	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	aan
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	aang
1.61	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	aap
0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	aat
0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	aau
0.00	0.03	0.00	0.00	0.00	0.11	0.00	0.09	0.00	0.00	0.10	0.19	ai
0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	ak
0.00	0.04	0.50	0.94	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	am
0.00	0.00	0.11	0.23	0.02	0.00	0.00	0.03	0.04	0.00	0.03	0.01	an
0.00	0.01	0.11	0.00	0.02	0.00	0.00	0.05	0.00	0.00	0.03	0.00	ang
0.00	0.06	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.00	ap
0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.21	0.01	at
0.00	0.31	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	au
0.00	0.37	0.06	0.00	0.17	0.11	0.09	0.12	0.01	0.00	0.50	0.11	e
0.00	0.09	0.00	0.00	0.15	0.86	0.38	0.40	0.35	0.00	0.99	0.32	ei
0.00	0.00	0.06	0.00	0.02	0.00	0.00	0.04	0.01	0.00	0.30	0.00	ek
0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.28	0.01	0.00	0.07	0.00	eng
1.61	0.10	0.00	0.00	0.07	0.06	0.09	0.08	0.01	0.00	0.43	0.03	eoi
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.01	0.00	0.00	0.00	eon
0.00	0.10	0.06	0.00	0.02	0.00	0.00	0.01	0.01	0.00	0.16	0.00	eot
0.00	0.00	0.00	0.00	0.25	5.27	0.38	0.04	0.80	0.54	0.42	**96.41**	i
0.00	0.00	0.06	0.00	0.07	0.29	0.28	0.50	0.05	0.00	**93.39**	0.09	ik
0.00	0.00	0.83	0.00	0.15	0.11	0.19	0.11	1.24	**91.67**	0.03	0.12	im
0.00	0.03	0.11	0.23	0.15	1.03	0.47	0.67	**95.55**	5.89	0.04	0.61	in
0.00	0.00	0.06	0.00	0.02	0.06	0.00	**95.24**	0.24	0.00	0.87	0.04	ing
0.00	0.00	0.11	0.00	0.12	1.55	**92.55**	0.00	0.07	0.42	0.31	0.13	ip
0.00	0.00	0.06	0.00	0.10	**89.40**	3.58	0.04	0.29	0.12	0.77	0.79	it
0.00	0.03	0.00	0.00	**97.69**	0.06	0.75	0.07	0.16	0.42	0.09	0.15	iu
0.00	0.00	77.82	**85.25**	0.00	0.00	0.00	0.02	0.04	0.06	0.00	0.00	m
0.00	0.01	**17.64**	3.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	ng
0.00	**89.47**	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	o
93.55	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	oe
0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.04	0.00	oek
1.61	0.04	0.06	0.47	0.00	0.00	0.00	0.18	0.00	0.06	0.04	0.00	oeng
0.00	0.78	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.03	0.05	oi
0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	ok
0.00	0.13	0.06	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	on
0.00	0.84	0.00	1.41	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	ong
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	ot
0.00	3.82	0.00	0.23	0.07	0.00	0.00	0.01	0.00	0.00	0.00	0.03	ou
0.00	0.27	0.06	0.00	0.05	0.00	0.00	0.03	0.00	0.00	0.03	0.00	u
0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	ui
0.00	0.19	0.00	0.00	0.05	0.00	0.00	0.02	0.01	0.00	0.06	0.00	uk
0.00	0.00	0.06	0.23	0.02	0.00	0.00	0.02	0.01	0.00	0.00	0.00	un
0.00	0.10	0.33	1.17	0.00	0.00	0.00	0.04	0.03	0.00	0.00	0.00	ung
0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	ut
0.00	0.00	0.00	0.00	0.37	0.46	0.38	0.00	0.05	0.06	0.03	0.33	yu
0.00	0.00	0.06	0.00	0.07	0.06	0.00	0.35	0.72	0.42	0.04	0.01	yun
0.00	0.05	0.00	0.00	0.07	0.34	0.47	0.04	0.05	0.00	0.21	0.01	yut

un	uk	ui	u	ou	ot	ong	on	ok	oi	oeng	oek	
0.00	0.00	0.00	0.00	0.02	0.00	0.17	0.18	0.74	0.69	0.17	0.09	**aa**
0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.00	1.06	0.00	0.09	**aai**
0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.74	0.03	0.00	0.09	**aak**
0.00	0.03	0.00	0.00	0.02	0.00	0.12	0.24	0.03	0.00	0.04	0.00	**aam**
0.75	0.00	0.00	0.00	0.00	0.00	0.44	0.72	0.00	0.12	0.68	0.00	**aan**
0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.18	0.00	0.00	0.13	0.00	**aang**
0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.11	0.02	0.00	0.00	**aap**
0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.21	0.02	0.01	0.18	**aat**
0.00	0.00	0.00	0.00	0.07	0.00	0.10	0.06	0.18	0.02	0.00	0.00	**aau**
0.03	0.00	0.13	0.00	0.00	0.00	0.00	0.06	0.03	0.26	0.03	0.00	**ai**
0.00	0.08	0.00	0.00	0.02	0.00	0.03	0.00	0.13	0.02	0.00	0.27	**ak**
0.03	0.00	0.00	0.00	0.02	0.00	0.05	0.00	0.00	0.00	0.01	0.09	**am**
0.23	0.00	0.00	0.00	0.00	0.00	0.19	0.12	0.03	0.02	0.21	0.00	**an**
0.03	0.00	0.00	0.00	0.01	0.00	0.38	0.18	0.00	0.03	0.38	0.00	**ang**
0.00	0.18	0.00	0.00	0.02	0.00	0.00	0.06	0.18	0.03	0.00	0.00	**ap**
0.00	0.05	0.00	0.00	0.02	0.00	0.02	0.00	0.40	0.19	0.13	2.84	**at**
0.00	0.43	0.10	0.00	1.44	0.00	0.15	0.06	0.47	0.10	0.12	0.46	**au**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.17	0.73	**e**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.09	**ei**
0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.27	**ek**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.09	**eng**
0.06	0.00	0.20	0.08	0.09	0.00	0.05	0.00	0.00	0.71	0.30	0.55	**eoi**
0.12	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.03	0.66	0.00	**eon**
0.00	0.54	0.00	0.00	0.12	0.00	0.00	0.00	0.11	0.21	0.09	0.92	**eot**
0.00	0.03	0.03	0.04	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	**i**
0.00	0.05	0.03	0.00	0.01	0.00	0.00	0.00	0.03	0.07	0.00	0.37	**ik**
0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	**im**
0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.24	0.03	0.02	0.01	0.00	**in**
0.06	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.05	0.00	**ing**
0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	**ip**
0.00	0.00	0.00	0.00	0.01	1.39	0.00	0.00	0.00	0.05	0.00	0.00	**it**
0.00	0.03	0.03	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.09	**iu**
0.06	0.00	0.00	0.00	0.01	0.00	0.14	0.12	0.00	0.00	0.00	0.00	**m**
0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**ng**
0.09	0.61	0.10	0.41	1.49	6.94	2.53	1.01	7.30	1.48	0.01	0.09	**o**
0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	**oe**
0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.31	0.14	**89.18**	**oek**
0.03	0.05	0.00	0.00	0.02	0.00	0.19	0.18	0.03	0.50	**95.33**	2.29	**oeng**
0.00	0.03	0.00	0.04	0.06	0.00	0.14	0.77	1.03	**92.58**	0.07	0.00	**oi**
0.00	1.61	0.00	0.08	0.05	13.89	0.63	0.18	**84.10**	0.16	0.00	0.09	**ok**
0.12	0.00	0.00	0.00	0.02	1.39	2.41	**87.01**	0.05	0.12	0.01	0.00	**on**
0.35	0.10	0.00	0.00	0.24	0.00	**89.99**	7.03	0.95	0.16	0.08	0.00	**ong**
0.00	0.00	0.00	0.00	0.00	**76.39**	0.00	0.00	0.00	0.00	0.00	0.00	**ot**
0.14	1.84	0.07	0.23	**94.44**	0.00	0.38	0.12	0.47	0.05	0.01	0.00	**ou**
0.58	0.15	0.23	**97.22**	0.36	0.00	0.00	0.00	0.05	0.00	0.00	0.00	**u**
0.40	0.00	**97.86**	0.68	0.11	0.00	0.00	0.00	0.00	0.16	0.00	0.00	**ui**
0.03	**92.59**	0.00	0.15	0.40	0.00	0.07	0.12	1.79	0.00	0.01	0.55	**uk**
95.11	0.05	0.34	0.41	0.03	0.00	0.12	0.48	0.00	0.05	0.01	0.00	**un**
0.81	0.89	0.03	0.19	0.32	0.00	0.96	0.06	0.05	0.07	0.04	0.00	**ung**
0.46	0.08	0.64	0.30	0.12	0.00	0.00	0.00	0.03	0.03	0.00	0.00	**ut**
0.00	0.00	0.00	0.00	0.02	0.00	0.05	0.00	0.00	0.00	0.00	0.00	**yu**
0.09	0.03	0.00	0.00	0.00	0.00	0.02	0.06	0.00	0.02	0.07	0.09	**yun**
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.00	**yut**

Table V: Confusion matrix for Finals

yut	yun	yu	ut	ung	
0.00	0.00	0.00	0.00	0.00	**aa**
0.00	0.00	0.00	0.00	0.00	**aai**
0.08	0.00	0.00	0.00	0.00	**aak**
0.00	0.00	0.00	0.00	0.00	**aam**
0.00	0.00	0.00	0.00	0.00	**aan**
0.00	0.00	0.00	0.00	0.01	**aang**
0.00	0.00	0.00	0.00	0.01	**aap**
0.00	0.00	0.00	0.00	0.00	**aat**
0.00	0.00	0.00	0.00	0.01	**aau**
0.08	0.00	0.00	0.00	0.00	**ai**
0.00	0.00	0.00	0.00	0.01	**ak**
0.00	0.00	0.00	0.00	0.06	**am**
0.00	0.06	0.00	0.00	0.04	**an**
0.00	0.02	0.00	0.00	0.11	**ang**
0.00	0.00	0.00	0.00	0.00	**ap**
0.17	0.00	0.09	0.00	0.00	**at**
0.00	0.00	0.00	0.18	0.06	**au**
0.17	0.11	0.09	0.00	0.01	**e**
0.58	0.06	0.03	0.00	0.01	**ei**
0.08	0.00	0.00	0.00	0.00	**ek**
0.00	0.00	0.00	0.00	0.00	**eng**
1.25	0.13	0.21	0.00	0.06	**eoi**
0.00	0.06	0.00	0.00	0.20	**eon**
0.25	0.02	0.12	0.09	0.00	**eot**
0.50	0.06	0.59	0.00	0.00	**i**
0.91	0.02	0.09	0.00	0.00	**ik**
0.08	0.04	0.06	0.00	0.01	**im**
0.00	0.71	0.03	0.00	0.01	**in**
0.08	0.32	0.00	0.00	0.09	**ing**
0.58	0.06	0.09	0.00	0.00	**ip**
0.58	0.02	0.23	0.00	0.00	**it**
0.33	0.04	0.06	0.00	0.05	**iu**
0.00	0.11	0.03	0.00	0.05	**m**
0.00	0.00	0.00	0.00	0.06	**ng**
0.00	0.00	0.03	0.27	0.12	**o**
0.08	0.00	0.03	0.00	0.00	**oe**
0.33	0.00	2.55	0.09	0.01	**oek**
0.00	0.11	0.15	0.00	0.28	**oeng**
0.00	0.02	0.00	0.00	0.00	**oi**
0.00	0.00	0.00	0.00	0.00	**ok**
0.00	0.00	0.00	0.00	0.02	**on**
0.00	0.02	0.00	0.00	1.10	**ong**
0.00	0.00	0.00	0.00	0.00	**ot**
0.00	0.00	0.00	0.54	0.60	**ou**
0.08	0.00	0.00	1.07	0.02	**u**
0.00	0.00	0.00	1.07	0.00	**ui**
0.00	0.00	0.00	0.45	0.23	**uk**
0.00	0.09	0.00	0.72	0.04	**un**
0.00	0.00	0.00	0.00	**96.52**	**ung**
0.00	0.00	0.00	**95.53**	0.00	**ut**
3.41	0.52	**93.10**	0.00	0.00	**yu**
1.41	**96.97**	1.09	0.00	0.02	**yun**
88.45	0.19	1.03	0.00	0.00	**yut**

Appendix IV Phonetic Question Set

A phonetic question used for the decision-tree based clustering essentially classifies all phonetic contexts into two groups. A set of Initial/Final contexts is defined for each phonetic question. Any IF having a context belongs to this set of IF contexts is equivalent to an answer "yes" to the question.

There are four major types of phonetic questions defined in this work: 1) manner of articulation of the adjacent onset/coda, 2) place of articulation of the adjacent onset/coda, 3) the vowel identity of the adjacent nucleus, and 4) the adjacent IF identity. All the questions are listed in the following tables.

Question	**context set**	**Question**	**context set**[3]
left = silence?	{ silence }	*left = i?*	{ *i }
left = labial onset?	{ b, p, m, w, m }	*left = u?*	{ u, aau, au, iu, ou }
left = alveolar onset?	{ d, t, n, c, z, s, j }	*left = nasal coda?*	{ *m, *n, *ng }
left = velar onset?	{ g, k , ng, gw, kw, ng }	*left = stop coda?*	{ *t, *k }
left = lip-round onset?	{ w, gw, kw }	*left = labial coda?*	{ *m, *p }
left = lateral onset?	{ l }	*left = alveolar coda?*	{ *n, *t }
left = vocal onset?	{ h }	*left = velar coda?*	{ *ng, *k }
left = dental-labial onset?	{ f }	*left = m?*	{ *m }
left = plosive onset?	{ b, d, g, gw, p, t, k, kw }	*left = n?*	{ *n }
left = nasal onset?	{ m, n, ng }	*left = ng?*	{ *ng }
left = fricative onset?	{ s, f, h }	*left = p?*	{ *p }
left = affricate onset?	{ z, c }	*left = t?*	{ *t }
left = pan fricative onset?	{ s, f, h, z, c }	*left = k?*	{ *k }
left = approximate onset?	{ l, w, j }	*left = labial?*	{ b, p, m, w, *m, *p }
left = glide onset?	{ w, j }	*left = alveolar?*	{ d, t, n, c, z, s, j, *n, *t }
left = aspirated onset?	{ p, t, k, kw, c }	*left = velar?*	{ g, k , ng, gw, kw, *ng, *k }
left = unaspirated onset?	{ b, d, g, gw, z }	*left = nasal?*	{ m, n, ng, *m, *n, *ng }
+ 72 questions on the left INTIAL/FINAL identity			

Table VI: Phonetic questions on left context.

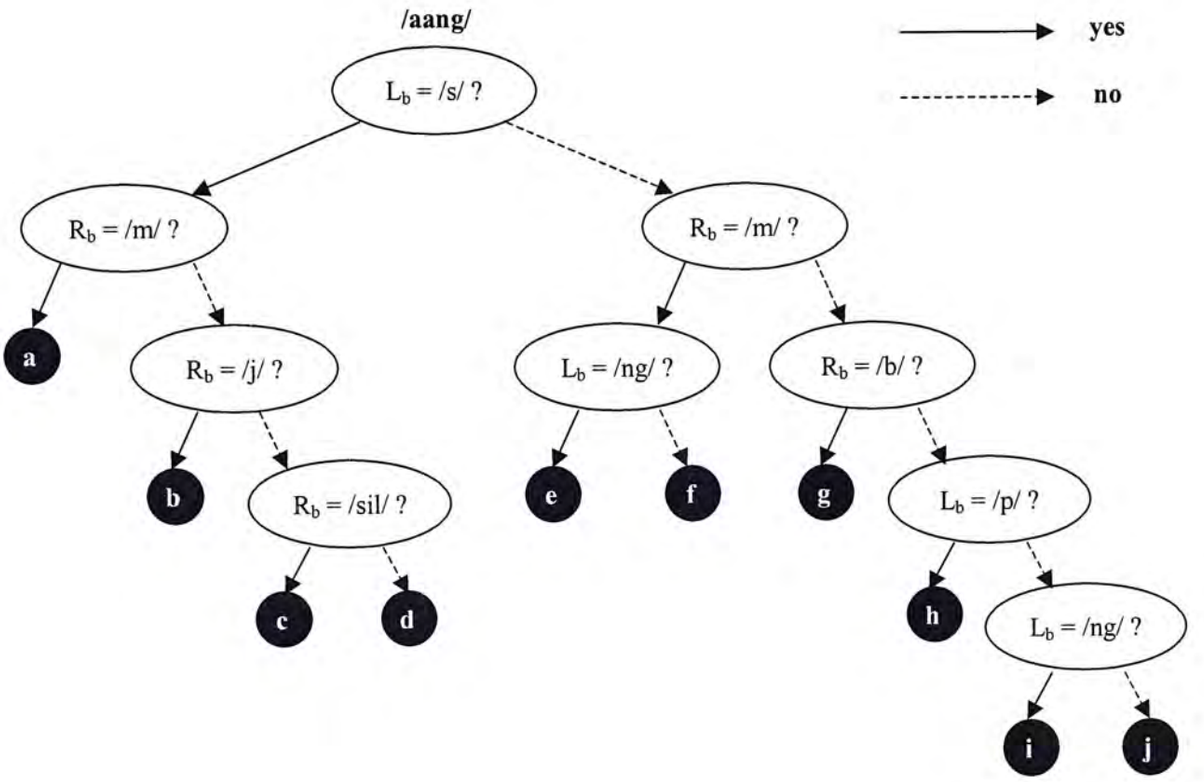
[3] The context starts with a * refers to any Final with the specified coda.

Question	context set	Question	context set[4]
right = silence?	{ silence }	*right = front vowel?*	{ i*, ei, eng, e, ek }
right = labial?	{ b, p, m, w, m }	*right = middle vowel? (front-back)*	{ a* , aa* }
right = alveolar?	{ d, t, n, c, z, s, j }	*right = back vowel?*	{ oi, ou, on, ong, o, ot, ok, u* }
right = velar?	{ g, k , ng, gw, kw, ng }	*right = round vowel?*	{ u*, o*, oe*, eo*, yu* }
right = lip-round?	{ w, gw, kw }	*right = unround vowel?*	{ a*, aa*, i*, ei, eng, e, ek }
right = lateral?	{ l }	*right = high vowel?*	{ i*, u*, yu* }
right = vocal?	{ h }	*right = middle vowel? (high-low)*	{ ai, au, am, an, ang, ap, at, ak, e*, oe* }
right = dental-labial?	{ f }	*right = low vowel?*	{ aa*, oi, ou, on, ong, o, ot, ok }
right = plosive?	{ b, d, g, gw, p, t, k, kw }	*right = aa?*	{ aa* }
right = nasal?	{ m, n, ng }	*right = a?*	{ a* }
right = fricative?	{ s, f, h }	*right = i?*	{ i* }
right = affricate?	{ z, c }	*right = yu?*	{ yu* }
right = pan fricative?	{ s, f, h, z, c }	*right = u?*	{ u* }
right = approximant?	{ l, w, j }	*right = e?*	{ ei, eng, e, ek }
right = glide?	{ w, j }	*right = oe?*	{ oe*, eo* }
right = aspirated?	{ p, t, k, kw, c }	*right = o?*	{ o* }
right = unaspirated?	{ b, d, g, gw, z }		
+ 72 questions on the right INITIAL/FINAL identity			

TableVII: Phonetic questions on right context.

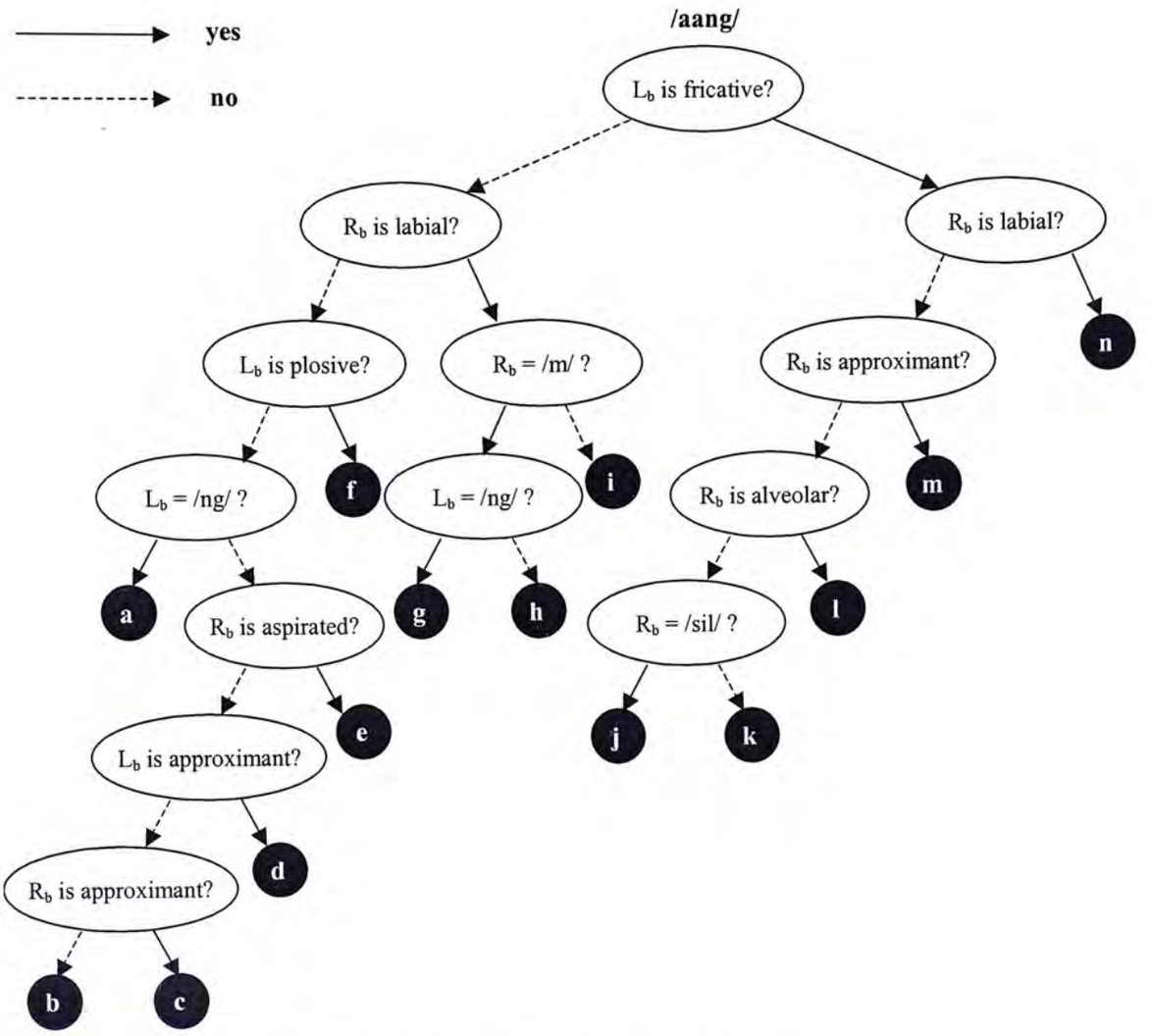
[4] The context ends with a * refers to any Final with the specified nucleus.

Appendix V CDDT and PCDT



a: { /aam/: 0.08, /aang/: 0.25, /am/: 0.25, /an/: 0.17, /ang/: 0.25 }
b: { /aan/: 0.11, /aang/: 0.53, /an/: 0.08, /ang/: 0.13, /ik/: 0.03, /ing/ 0.08, /oeng/: 0.05 }
c: { /aang/: 0.38, /an/: 0.44, /ang/: 0.13, /eon/: 0.06 }
d: { /aa/: 0.02, /aam/: 0.02, /aan/: 0.13, /aang/: 0.60, /an/: 0.08, /ang/: 0.13, /eon/: 0.01 }
e: { /aa/: 0.05, /aak/: 0.05, /aam/: 0.1, /aan/: 0.4, /aang/: 0.3, /an/: 0.05, /oeng/: 0.05 }
f: { /aa/: 0.12, /aam/: 0.35, /aan/: 0.23, /aang/: 0.23, /ang/: 0.04, /au/ 0.04 }
g: { /aam/: 0.17, /aan/: 0.08, /aang/: 0.42, /an/: 0.08, /eng/: 0.08, /m/: 0.08, /ong/: 0.08 }
h: { /aa/: 0.01, /aai/: 0.01, /aak/: 0.04, /aam/: 0.01, /aan/: 0.10, /aang/: 0.77, /ang/: 0.03, /o/: 0.01 }
i: { /aa/: 0.08, /aan/: 0.5, /aang/: 0.25, /ang/: 0.08, /on/: 0.08 }
j: { /aak/: 0.01, /aam/: 0.03, /aan/: 0.25, /aang/: 0.66, /an/: 0.01, /ang/: 0.03, /on/, 0.01 }

Figure I: CDDT for the Final /aang/.



a: { /aa/: 0.08, /aan/: 0.50, /aang/: 0.25, /ang/: 0.08, /on/: 0.08 }
b: { /aan/: 0.41, /aang/: 0.59 }
c: { /aan/: 0.58, /aang/: 0.42 }
d: { /aan/: 0.24, /aang/: 0.76 }
e: { /aam/: 0.17, /aan/: 0.17, /aang/: 0.67 }
f: { /aa/: 0.01, /aai/: 0.01, /aak/: 0.05, /aam/: 0.01, /aan/: 0.10, /aang/: 0.77, /ang/: 0.01, /o/: 0.01, /on/: 0.01 }
g: { /aa/: 0.05, /aak/: 0.05, /aam/: 0.10, /aan/: 0.40, /aang/: 0.30, /an/: 0.05, /oeng/: 0.05 }
h: { /aa/: 0.09, /aam/: 0.36, /aan/: 0.27, /aang/: 0.23, /au/: 0.05 }
i: { /aam/: 0.13, /aan/: 0.10, /aang/: 0.55, /an/: 0.03, /ang/: 0.10, /eng/: 0.03, /m/: 0.03, /ong/: 0.03 }
j: { /aang/: 0.38, /an/: 0.44, /ang/: 0.13, /eon/: 0.06 }
k: { /aa/: 0.07, /aang/: 0.73, /an/: 0.07, /ang/: 0.07, /eon/: 0.07 }
l: { /aa/: 0.02, /aan/: 0.11, /aang/: 0.56, /an/: 0.13, /ang/: 0.19 }
m: { /aan/: 0.14, /aang/: 0.52, /an/: 0.07, /ang/: 0.12, /ik/: 0.02, /ing/: 0.07, /oeng/: 0.05 }
n: { /aa/: 0.04, /aam/: 0.17, /aan/: 0.09, /aang/: 0.30, /am/: 0.13, /an/: 0.09, /ang/: 0.17 }

Figure II PCDT for the Final /aang/.