Computational Linguistics and Chinese Language Processing
Vol. 11, No. 1, March 2006, pp. 17-36

Modeling Cantonese Pronunciation Variations for Large-Vocabulary Continuous Speech Recognition

Tan Lee*, Patgi Kam* and Frank K. Soong*+

Abstract

This paper presents different methods of handling pronunciation variations in Cantonese large-vocabulary continuous speech recognition. In an LVCSR system, three knowledge sources are involved: a pronunciation lexicon, acoustic models and language models. In addition, a decoding algorithm is used to search for the most likely word sequence. Pronunciation variation can be handled by explicitly modifying the knowledge sources or improving the decoding method. Two types of pronunciation variations are defined, namely, phone changes and sound changes. Phone change means that one phoneme is realized as another phoneme. A sound change happens when the acoustic realization is ambiguous between two phonemes. Phone changes are handled by constructing a pronunciation variation dictionary to include alternative pronunciations at the lexical level or dynamically expanding the search space to include those pronunciation variants. Sound changes are handled by adjusting the acoustic models through sharing or adaptation of the Gaussian mixture components. Experimental results show that the use of a pronunciation variation dictionary and the method of dynamic search space expansion can improve speech recognition performance substantially. The methods of acoustic model refinement were found to be relatively less effective in our experiments.

Keywords: Automatic Speech Recognition, Pronunciation Variation, Cantonese

1. Introduction

Given a speech input, automatic speech recognition (ASR) is a process of generating possible hypotheses for the underlying word sequence. This can be done by establishing a mapping between the acoustic features and the yet to be determined linguistic representations. Given

* Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong Tel: 852-26098267 Fax: 852-26035558
E-mail: tanlee@ee.cuhk.edu.hk
The author for correspondence is Tan Lee.
+ Micrpsoft Research Asia, 5th Floor, Sigma Center, 49 Zhichun Road, Haidian, Beijing 100080, China

[Received June 16, 2005; Revised January 20, 2006; Accepted January 23, 2006]

the high variability of human speech, such mapping is in general not one-to-one. Different linguistic symbols can give rise to similar speech sounds, while the same linguistic symbol may also be realized in different pronunciations. The variability is due to co-articulation, regional accents, speaking rate, speaking style, etc. Pronunciation modeling is aimed at providing an effective mechanism by which ASR systems can be adapted to pronunciation variability.

Pronunciation variations can be divided into two types: phone change and sound change [Kam 2003] [Liu and Fung 2003]. In [Saraçlar and Khudanpur 2000] [Liu 2002], they are also referred to as complete change and partial change, respectively. A phone change happens when a *baseform* (canonical) phoneme is realized as another phoneme, which is referred to as its *surface-form*. The baseform pronunciation is considered to be the "standard" pronunciation that the speaker is supposed to use. Surface-form pronunciations are the actual pronunciations that different speakers may use. A sound change can be described as variation in phonetic properties, such as nasalization, centralization, voicing, etc. Acoustically, the variant sound is considered to be neither the baseform nor any surface-form phoneme. In other words, we cannot find an appropriate unit in the language's phoneme inventory to represent the sound. In terms of the scope of such variations, pronunciation variations can be divided into word-internal and cross-word variations [Strik and Cucchiarini 1999].

There have been many studies on modeling pronunciation variations for improving ASR performance. They are focused mainly on two problems: 1) prediction of the pronunciation variants, and 2) effective use of pronunciation variation information in the recognition process [Strik and Cucchiarini 1999]. Knowledge-based approaches use findings from linguistic studies, existing pronunciation dictionaries, and phonological rules to predict the pronunciation variations that could be encountered in ASR [Aubert and Dugast 1995] [Kessens *et al*. 1999]. Data-driven approaches attempt to discover the pronunciation variants and the underlying rules from acoustic signals. This is done by performing automatic phone recognition and aligning the recognized phone sequences with reference transcriptions to find out the surface forms [Saraçlar *et al*. 2000] [Wester 2003]. Some studies used hand-labelled corpora [Riley *et al*. 1999].

The key components of a large-vocabulary continuous speech recognition system are the acoustic models, the pronunciation lexicon and the language models [Huang *et al.* 2001]. The acoustic models are a set of hidden Markov models (HMM) that characterize the statistical variations of input speech. Each HMM represents a specific sub-word unit, e.g. a phoneme. The pronunciation lexicon and the language models are used to define and constrain the ways sub-word units can be concatenated to form words and sentences. They are used to define a search space from which the most likely word string(s) can be determined with a computationally efficient decoding algorithm. Within such a framework, pronunciation

variations can be handled by modifying one or more of the knowledge sources or improving the decoding algorithm. Phone changes can be handled by replacing the baseform transcription with surface-form transcriptions, i.e. the actual pronunciations observed. In an LVCSR system, this can be done by either augmenting the baseform lexicon with the additional pronunciation variants [Kessens *et al.* 1999] [Liu *et al*. 2000] [Byrne *et al.* 2001], or expanding the search space during the decoding process to include those variants [Kam and Lee 2002]. In order to deal with sound changes, pronunciation modeling must be applied at a lower level, for example, on the individual states of a hidden Markov model (HMM) [Saraçlar *et al*. 2000]. In general, acoustic models are trained solely with baseform transcriptions. It is assumed that all training utterances follow exactly the canonical pronunciations. This convenient, but apparently unrealistic, assumption renders the acoustic models inadequate in representing the variations of speech sounds. To alleviate this problem, various methods of acoustic model refinement were proposed [Saraçlar *et al*. 2000] [Venkataramani and Byrne 2001] [Liu 2002].

In this paper, the pronunciation variations in continuous Cantonese speech are studied. The linguistic and acoustic properties of spoken Cantonese are considered in the analysis of pronunciation variations and, subsequently, the design of pronunciation modeling techniques for LVCSR. Like in most conventional approaches, phone changes are anticipated by using an augmented pronunciation lexicon. The lexicon includes the most frequently occurring alternative pronunciations that are derived from training data. We also describe a novel method of dynamically expanding the search space during decoding to include pronunciation variants that are predicted with context-dependent pronunciation models. For sound changes, we propose to measure the similarities between confused baseform and surface-form models at the Gaussian mixture component level and, accordingly, refine the models through sharing and adaptation of the relevant mixture components.

In the next section, the properties of spoken Cantonese are described and the fundamentals of Cantonese LVCSR are explained. In Section 3, different methods of modeling pronunciation variations at the lexical level are presented in detail and experimental results are given. The techniques for handling sound changes through acoustic model refinement are described in Section 4. Conclusions are given in Section 5.

2. Cantonese LVCSR

2.1 About Cantonese

Cantonese is one of the major Chinese dialects. It is the mother tongue of over 60 million people in Southern China and Hong Kong [Grimes *et al.* 2000]. The basic unit of written Cantonese is a Chinese character [Chao 1965]. Chinese characters are ideographic, meaning that they contain no information about pronunciation. There are more than ten thousand

distinctive characters. In Cantonese, each of them is pronounced as a single syllable that carries a specific tone. A sentence is spoken as a string of monosyllabic sounds. A character may have multiple pronunciations, and a syllable typically corresponds to a number of different characters.

A Cantonese syllable is formed by concatenating two types of phonological units: the *Initial* and the *Final*, as shown in Figure 1 [Hashimoto 1972]. There are 20 Initials (including the null Initial) and 53 Finals in Cantonese, in contrast to 23 Initials and 37 Finals in Mandarin. Table 1 and Table 2 list the Initials and Finals of Cantonese. They are labeled using *Jyut Ping*, a phonemic transcription scheme proposed by the Linguistic Society of Hong Kong [LSHK 1997]. In terms of the manner of articulation, the 20 Initials can be categorized into seven classes: null, plosive, affricate, fricative, glide, liquid, and nasal. The 53 Finals can be divided into five categories: vowel (long), diphthong, vowel with nasal coda, vowel with stop coda, and syllabic nasal. Except for [m] and [ng], each Final contains at least one vowel element. The stop codas, i.e., -p, -t and -k, are unreleased. In Cantonese, there are more than 600 legitimate Initial-Final combinations, which are referred to as *base syllables*.

BASE SYLLABLE		
Initial	*Final*	
[Onset]	Nucleus	[Coda]

Figure 1. The composition of a Cantonese syllable.* [] *means optional.

Table 1.The Cantonese Initials

Jyut Ping symbols	Manner of Articulation	Place of Articulation
[b]	Plosive, unaspirated	Labial
[d]	Plosive, unaspirated	Alveolar
[g]	Plosive, unaspirated	Velar
[p]	Plosive, aspirated	Labial
[t]	Plosive, aspirated	Alveolar
[k]	Plosive, aspirated	Velar
[gw]	Plosive, unaspirated, lip-rounded	Velar, labial
[kw]	Plosive, aspirated, lip-rounded	Velar, labial
[z]	Affricate, unaspirated	Alveolar
[c]	Affricate, aspirated	Alveolar
[s]	Fricative	Alveolar
[f]	Fricative	Dental-labial
[h]	Fricative	Vocal
[j]	Glide	Alveolar
[w]	Glide	Labial
[l]	Liquid	Lateral
[m]	Nasal	Labial
[n]	Nasal	Alveolar
[ng]	Nasal	Velar

Table 2.The 53 Cantonese Finals

		CODA								
		Nil	-i	-u	-p	-t	-k	-m	-n	-ng
NUCLEUS	-aa-	[aa]	[aai]	[aau]	[aap]	[aat]	[aak]	[aam]	[aan]	[aang]
	-a-		[ai]	[au]	[ap]	[at]	[ak]	[am]	[an]	[ang]
	-e-	[e]	[ei]				[ek]			[eng]
	-i-	[i]		[iu]	[ip]	[it]	[ik]	[im]	[in]	[ing]
	-o-	[o]	[oi]	[ou]		[ot]	[ok]		[on]	[ong]
	-u-	[u]	[ui]			[ut]	[uk]		[un]	[ung]
	-yu-	[yu]				[yut]			[yun]	
	-oe-	[oe]	[eoi]			[eot]	[oek]		[eon]	[oeng]
								[m]		[ng]

From phonological points of view, Cantonese has nine tones that are featured by differently stylized pitch patterns. They are divided into two categories: entering tones and non-entering tones. The entering tones occur exclusively with syllables ending in a stop coda (-p, -t, or –k). They are contrastively shorter in duration than the non-entering tones. In terms of pitch level, each entering tone coincides roughly with a non-entering counterpart. In many transcription schemes, only six distinctive tone categories are defined. They are labeled as Tone 1 to Tone 6 in the *Jyu Ping* system. If tonal difference is considered, the total number of distinctive *tonal syllables* is about 1,800.

Table 3 gives an example of a Chinese word and its spoken form in Cantonese. The word 我們 (meaning “we”) is pronounced as two syllables. The first syllable is formed from the Initial [ng] and the Final [o], with Tone 5. The second syllable is formed from the Initial [m] and the Final [un], with Tone 4.

Table 3.An example Chinese word and its Cantonese pronunciations

Word	Chinese characters	Base syllables	Initial & Final	Tone
我們	我	ngo	[ng] [o]	5
	們	mun	[m] [un]	4

2.2 Linguistic Studies on Pronunciation Variations in Cantonese

Over the past twenty years, there have been sociolinguistic studies on how phonetic variations in Cantonese are related with social characteristics of speakers such as sex, age, and educational background. They have revealed some systematic patterns underlying the phonetic variations [Bauer and Benedict 1997] [Bourgerie 1990] [Ho 1994]. Table 4 gives a summary of the major observations in these studies.

Table 4. Major phonetic variations in Cantonese observed by sociolinguistic studies

Initial consonants	[n] ~ [l]	Inter-change between nasal and lateral Initials
	[ng]~ null	Inter-change between velar nasal and null Initial.
	[gw] → [g]	Change from labialized velar to delabialized velar before back-round vowel [o]
Syllabic nasal	[ng] → [m]	Change from velar nasal to bilabial nasal
Final consonants	-ng → -n	Change from velar nasal coda to dental nasal coda
	-k ~ -t -k ~ -p	Inter-change between velar stop coda and dental or glottal stop coda

It was found that [n]→[l], [ng]→null, and [gw]→[g] correlate with the sex and age of a speaker [Bourgerie 1990]. Older people make these substitutions much less frequently than younger generations. Female speakers tend to substitute [n] with [l], and delete [ng] more frequently than males. A correlation with the formality of the speech situation was also observed [Bourgerie 1990]. In casual speech, [l], null Initial, and [g] occur more frequently. According to [Bauer and Benedict 1997], the variations are also related to the development of neighboring dialects in the Pearl River Delta.

When the preceding syllable ends with a nasal coda, there is a tendency to substitute the Initial [l] of the succeeding syllable with [n] [Ho 1994]. Labial dissimilation is probably the cause of the change [gw]→[g], when the right context is -o, for example "gwok" 國 (country), pronounced as "gok" 角 (corner). The sequence of the two lip-rounded segments -w- and -o- become redundant or unnecessary with the second one driving out the first. The change [ng]→[m] is due to the fact that when [ng] occurs in the presence of a bilabial coda, its place of articulation changes to bilabial. For example, "sap ng" 十五 (fifteen) becomes "sap m" through the perseverance of the bilabial closure of the coda -p into the articulation of the following syllabic nasal. This is referred to as perseveratory assimilation [Bauer and Benedict 1997].

Other pronunciation variations are due to the dialectal accents of non-native speakers, who may have difficulties mastering some of the Cantonese pronunciations. They sometimes use the pronunciation of their mother tongue to pronounce a Cantonese word, for example, "ngo" 我 (me) is pronounced as "wo" by a Mandarin speaker.

2.3 Cantonese LVCSR: the Baseline System

Figure 2 gives the functional block diagram of a typical LVCSR system. At the front-end processing module, the input speech is analyzed and converted into a sequence of acoustic feature vectors, denoted by O. The goal of speech recognition is to determine the most probable word sequence W, given the observation O. With the Bayes' formula, the decision

can be made as

$$W^* = \arg\max_W P(W \mid O) = \arg\max_W P(O \mid W)P(W) . \quad (1)$$

Usually the acoustic models are built at the sub-word level. Let B be the sub-word sequence that represents W. Eq. (1) can be written as

$$W^* = \arg\max_W P(O \mid B)P(B \mid W)P(W) , \quad (2)$$

where $P(O \mid B)$ and $P(W)$ are referred to as the (sub-word level) acoustic models and the language models, respectively. $P(B \mid W)$ is given by a pronunciation lexicon.

In the case of Chinese speech recognition, the sub-word units can be either syllables, Initials and Finals, or phone-like units. The recognition output is typically represented as a sequence of Chinese characters. The details of our baseline system for Cantonese LVCSR are given below.

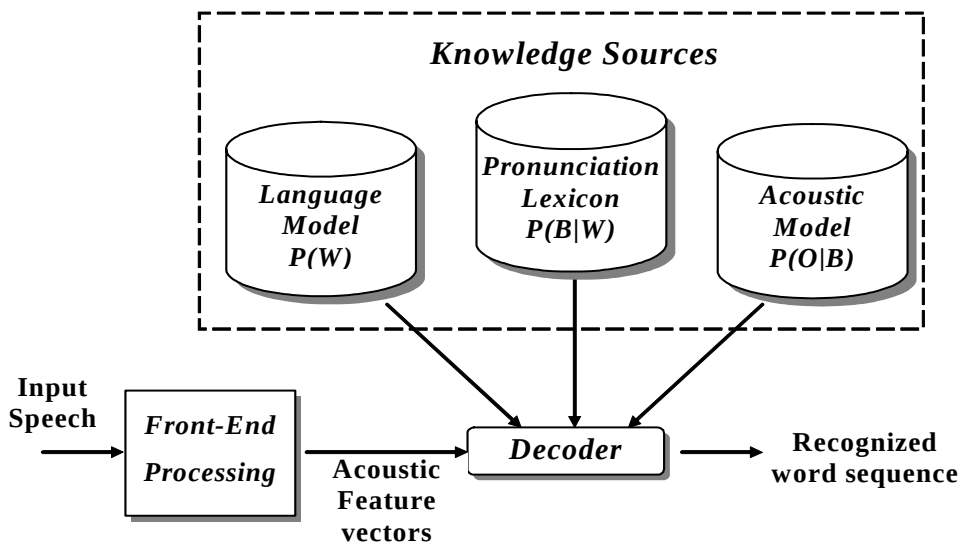

Figure 2. A typical LVCSR system

Front-end processing

Acoustic feature vectors are computed every 10 msec. Each feature vector is composed of 39 elements, which includes 12 Mel-frequency cepstral coefficients, log energy, and their first-order and second-order derivatives. The analysis window size is 25 msec.

Acoustic models

The acoustic models are right-context-dependent cross-word Initials and Finals models [Wong 2000]. The number of HMM states for Initial and Final units are 3 and 5, respectively. Each state is represented by a mixture of 16 Gaussian components. The decision tree based state clustering approach is used to allow the sharing of parameters among models.

Pronunciation lexicons and language models

The lexicon contains about 6,500 entries, among which 60% are multi-character words and the others are single-character words [Wong 2000]. These words were selected from a newspaper text corpus of 98 million Chinese characters. The out-of-vocabulary percentage is about 1% [Wong 2000]. For each word entry, the canonical pronunciation(s) is specified in the form of Initials and Finals [CUPDICT 2003]. The language models are word bi-grams that were trained with the same text corpus described above.

Decoder

The search space is formed from lexical trees that are derived from the pronunciation lexicon. One-pass Viterbi search is used to determine the most probable word sequence [Choi 2001]. The acoustic models were trained using CUSENT, which is a read speech corpus of continuous Cantonese sentences collected at the Chinese University of Hong Kong [Lee *et al.* 2002]. There are over 20,000 gender-balanced training utterances. The test data in CUSENT consists of 1,200 utterances from 6 male and 6 female speakers. The performance of the LVCSR system is measured in terms of word error rate (WER) for the 1,200 test utterances. The baseline WER is 25.34%.

3. Handling Phone Change with Pronunciation Models

The pronunciation lexicon used in the baseline system provides only the baseform pronunciation for each of the word entries. In real speech, the baseform pronunciations are realized differently, depending on the speakers, speaking styles, etc. Phone change means that the pronunciation variation can be considered as one or more Initial or Final (IF) unit in the baseform pronunciation being substituted by another IF unit. Note that the substituting surface-form unit is also one of the legitimate IF units, as listed in Tables 1 and 2.

A pronunciation model (PM) is a descriptive and predictive model by which the surface-form pronunciation(s) can be derived from the baseform one. There have been three different types of models proposed by previous studies. They are: 1) phonological rules for generating pronunciation variations [Wester 2003] [Kessens *et al*. 2003], 2) a pronunciation variation dictionary (PVD) that explicitly lists alternative pronunciations [Aubert and Dugast 1995] [Kessens *et al.* 1999] [Liu *et al*. 2000], and 3) statistical decision trees that predict pronunciation variations according to phonetic context [Riley *et al*. 1999] [Fosler-Lussier

1999] [Saraçlar *et al*. 2000]. In this study, two different approaches to handling phone changes in Cantonse ASR are formulated and evaluated. The first approach uses a probabilistic PVD to augment the baseform lexicon. This is a straightforward and commonly used method that has been proven effective for various tasks and languages [Strik and Cucchiarini 1999]. In the second approach, pronunciation variation information is introduced during the decoding process. Decision tree based PMs are used to dynamically expand the search space. In [Saraçlar *et al*. 2000], a similar idea was presented. Decision tree based PMs were applied to a word lattice to construct a recognition network that includes surface-form realizations.

3.1 Use of a Pronunciation Variation Dictionary (PVD)

In this study, the information about Cantonese pronunciation variations is obtained through the data-driven approach. This is done by aligning the baseform transcriptions with the recognized surface-form IF sequences for all training utterances. For each training utterance, the surface-form IF sequence is obtained through phoneme recognition with the acoustic models as described in Section 2.3. To reflect the syllable structure of Cantonese, the recognition output is constrained to be a sequence of Initial-Final pairs. With this approach, only substitutions at the IF level are considered pronunciation variations. Partial change of an IF unit and the deletion of an entire Initial or Final are not reflected in the surface-form IF sequences.

The surface-form phoneme sequence is then aligned with the baseform transcription. This gives a phoneme accuracy of 90.33%. The recognition errors are due, at least partially, to phoneme-level pronunciation variation. For a particular baseform phoneme *b* and a surface-form phoneme *s*, the probability of *b* being pronounced as *s* is computed based on the number of times that *b* is recognized as *s*. This probability is referred to as the variation probability (VP). As a result, each pair of IF units is described with a probability of being confused. This is also referred to as a confusion matrix [Liu *et al*. 2000]. It is assumed that systematic phone change can be detected by a relatively high VP, while a low VP is more likely due to recognition errors. A VP threshold is used to prune those less frequent surface-form pronunciations. As a result, for each baseform IF unit, we can find a certain number of surface-form units, each with a pre-computed VP.

A straightforward way of handling pronunciation variation is to augment the basic pronunciation lexicon with alternative pronunciations [Strik and Cucchiarini 1999]. Such an augmented lexicon is named a pronunciation variation dictionary (PVD). In the PVD, each word can have multiple pronunciations, each being assigned a word-level variation probability (VP). The PVD can be obtained from the IF confusion matrix. The word-level VP is given by multiplying the phone-level VPs of all the individual phonemes in the surface-form pronunciation. With the use of the PVD, the goal of speech recognition is essentially to search

for the most probable word sequence by considering all possible surface-form realizations. This can be conceptually illustrated by modifying Eq. (2) as

$$W^* = \arg\max_{W,k} P(O \mid S_{W,k}) P(S_{W,k} \mid W) P(W), \tag{3}$$

where $S_{W,k}$ denotes one of the surface-forms realizations of W. $P(S_{W,k} \mid W)$ are obtained from the word-level VPs.

3.2 Prediction of Pronunciation Variation during Decoding

The PVD includes both context-independent and context-dependent phone changes. Since each word is treated individually, the phonetic context being considered is limited to within the word. To deal with cross-word context-dependent phone changes, we propose applying pronunciation models at the decoding level. Our baseline system uses a one-pass search algorithm [Choi 2001]. The search space is structured as lexical trees. Each node on a tree corresponds to a baseform IF unit. The search is token based. Each token represents a path that reaches a particular lexical node. The propagation of tokens follows the lexical trees, which cover only the legitimate phoneme sequences as specified by the pronunciation lexicon. The search algorithm can be modified in a way that the number of alive tokens is increased to account for pronunciation variations. When a path extends from a particular IF node, its destination node can be either the legitimate node (baseform pronunciation) or any of the predicted surface-form nodes. In other words, the search space is dynamically expanded during the search process.

In this approach, a context-dependent pronunciation model is needed to predict the surface-form phoneme given the baseform phoneme and its context. It is implemented using the decision tree clustering technique, following the approaches described in [Riley *et al.* 1999] [Fosler-Lussier 1999]. Each baseform phoneme is described using a decision tree. Given a baseform phoneme, as well as its left context (the right context is not available in a forward Viterbi search), the respective decision-tree pronunciation model (DTPM) gives all possible surface-form realizations and their corresponding VPs [Kam and Lee 2002].

Like the confusion matrix, the DTPM is trained with the phoneme recognition outputs for the CUSENT training utterances. The training involves an optimization process by which the surface-form phonemes are clustered based on phonetic context. At a particular node of the tree, a set of "yes/no" questions about the phonetic context are evaluated. Each question leads to a different partition of the training data. The question that minimizes the overall conditional entropy of the surface-form realizations is selected for that node. The node-splitting process stops when there are too few training data [Kam 2003].

3.3 Experimental Results and Discussion

Table 5 gives the recognition results with the use of PVDs that are constructed with different values of the VP threshold. The baseline system uses the basic pronunciation lexicon that contains 6,451 words. The size of the PVD increases as the VP threshold decreases. It is obvious that the introduction of pronunciation variants improves recognition performance. The best performance is attained with a VP threshold of 0.05. In this case, the PVD contains 8,568 pronunciations for the 6,451 words, i.e. 1.33 pronunciation variants per word. With a very small value for the VP threshold, e.g. 0.02, the recognition performance is not good because there are too many pronunciation variants being included and some of them do not really represent pronunciation variation.

Table 5. Recognition results of using a PVD with different VP thresholds

	Baseline	VP threshold				
		0.02	**0.05**	0.10	0.15	0.20
Word error rate (%)	25.34	23.91	**23.49**	23.70	23.64	23.58
No. of word entries in the PVD	6,451	20,840	**8,568**	7,356	7,210	7,171

Table 6 shows the recognition results attained by using the DTPM for dynamic search space expansion. It appears that this approach is as effective as the PVD. Unlike the results for the PVD, the performance with a VP threshold of 0.2 is better than that with a threshold of 0.05. This means that the predictions made by the DTPM should be pruned more stringently than the IF confusion matrix. Because of its context-dependent nature, the DTPM has relatively less training data, and the variation probabilities cannot be reliably estimated. It is preferable not to include those unreliably predicted pronunciation variants.

Table 6. Recognition results by dynamic search space expansion

	Baseline	VP threshold	
		0.05	**0.2**
Word error rate (%)	25.34	23.53	**23.27**

By analyzing the recognition results in detail, it is observed that many errors are corrected by allowing the following pronunciation variations:

Initials: [gw]→[g], [n]→[l], [ng]→null

Finals: [ang]→[an], [ng]→[m] (syllabic nasal)

These observations match well with the findings in sociolinguistic studies on Cantonese phonology (Section 2.2).

4. Handling Sound Change by Acoustic Model Refinement

Unlike phone changes, a sound change cannot be described as a simple substitution of one phoneme for another. It is regarded as a partial change from the baseform phoneme to a surface-form phoneme [Liu and Fung 2003]. Our approaches presented below attempt to refine the acoustic models to handle the acoustic variation caused by sound changes. The acoustic models are continuous-density HMMs. The output probability density function (pdf) at each HMM state is a mixture of Gaussian distributions. The use of multiple mixture components is intended to describe complex acoustic variabilities. The acoustic models trained only according to the baseform pronunciations are referred to as baseform models. Each baseform phoneme may have different surface-form realizations. The acoustic models representing these surface-form phonemes are referred to as surface-form models. A baseform model doesnot reflect the acoustic properties of the relevant surface-form phonemes. One way of dealing with this deficiency is through the sharing of Gaussian mixture components among the baseform and surface-form models. In [Saraçlar *et al.* 2000], a state-level pronunciation model (SLPM) was proposed. It allows the HMM states of a baseform model to share the output densities of its surface-form phonemes. A state-to-state alignment was obtained from decision-tree PMs, and the most frequently confused state pairs were involved in parameter sharing. In [Liu and Fung 2004], the method of phonetic mixtures tying was applied to deal with sound changes. A set of so-called extended phone units were derived from acoustic training data to describe the most prominent phonetic confusion. These units were then modeled by mixture tying with the baseform models. In this study, we investigate both the sharing and adaptation of the acoustic model parameters at the mixture level [Kam *et al.* 2003].

4.1 Sharing of Mixture Components

First of all, the states of the baseform and surface-form models are aligned. It is assumed that both models have the same number of states. Then, state j of the baseform model is aligned with state j of the surface-form model. Consider a baseform phoneme B. The output pdf at state j is given as

$$b_j(o_t) = \sum_{m=1}^{M} w_{jm} N(o_t; \mu_{jm}, \Sigma_{jm}), \tag{4}$$

where M is the number of Gaussian mixture components, and w_{jm} is the weight for the mth mixture component. The baseform output pdf can be modified to include the contributions from the surface-form states

$$b_j'(o_t) = VP(B,B)\cdot b_j(o_t) + \sum_{\substack{n=1 \\ S_n \neq B}}^{N} VP(S_n,B)\cdot q_{S_n,j}(o_t), \tag{5}$$

where S_n denotes the *n*th surface-form of B, N is the total number of surface-forms, $VP(S_n,B)$ is the variation probability of S_n with respect to baseform B, and $q_{S_n,j}(o_t)$ denotes the output pdf of state j of the *n*th surface-form model.

The number of mixture components in the resultant baseform model depends on N. More surface-form pronunciations bring in more mixture components to the modified baseform state. As the number of mixture components is changed, re-estimation of mixture weights is required.

4.2 Adaptation of Mixture Components

Although sharing mixture components yields an acoustically richer model, it also greatly increases the model size for which more memory space and higher computation complexities are required. Moreover, if the baseform and surface-form mixture components are very similar, including them all in the modified baseform is unnecessarily superfluous.

We propose to refine the baseform acoustic models through parameters adaptation. The total number of model parameters remains unchanged. Like in the approach of mixture sharing, the states of the baseform and surface-form models are aligned. The surface-forms are generated from the IF confusion matrix. Consider the aligned states of the baseform phoneme B and one of its surface-forms S. Let $m_B(i)$ and $m_S(j)$ denote the *i*th mixture component in the baseform state and the *j*th mixture component in the surface-form state, respectively, where $i, j = 1,2,\cdots,M$. The distances between all pairs $(m_B(i), m_S(j))$ are computed. Then each surface-form component is paired up with the nearest baseform component. That is, for each $m_S(j)$, we find

$$\hat{i} = \arg\min_{m_B(i)} d(m_B(i), m_S(j)). \tag{6}$$

The "distance" between two Gaussian distributions is calculated using the Kullback-Leibler divergence (KLD) [Myrvoll and Soong 2003]. Given two multivariate Gaussian distributions f and g, the symmetric KLD has the following closed form

$$d(f,g) = \frac{1}{2} trace\{(\Sigma_f^{-1} + \Sigma_g^{-1})(\mu_f - \mu_g)(\mu_f - \mu_g)^T + \Sigma_f \Sigma_g^{-1} + \Sigma_g \Sigma_f^{-1} - 2\mathrm{I}\}, \tag{7}$$

where μ and Σ denote the mean vectors and the covariance matrices of the two distributions, respectively, and I is the identity matrix.

As a result, for this pair of baseform and surface-form states, each Gaussian component $m_B(i)$ is associated with k surface-form components, as illustrated in Figure 3. The centroid of these k components is computed. If the baseform B has n surface forms, there will be n such centroids. These surface-form centroids and the corresponding baseform component are weighted with the VP, and together produce a new centroid that is taken as the adapted baseform component. In this way, the adapted model is expected to shift towards the surface-form phonemes. The extent of such a shift depends on the VP. The mean and covariance of the centroid of k weighted Gaussian components can be found by minimizing the following weighted divergence

$$\{\mu_c', \Sigma_c'\} = \arg\min_{\mu_c, \Sigma_c} \sum_{n=1}^{k} a_n d(f_c, f_n), \tag{8}$$

where f_n denotes the nth component and a_n is the respective weighting coefficient. Assuming diagonal covariances, the weighted centroid is given as [Myrvoll and Soong 2003]

$$\begin{aligned} \mu_c'(i) &= \frac{\Sigma_{n=1}^{k} a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i)) \mu_n(i)}{\Sigma_{n=1}^{k} a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i))} \\ \Sigma_c'(i) &= \sqrt{\frac{\Sigma_{n=1}^{k} a_n [\Sigma_n(i) + (\mu_c(i) - \mu_n(i))^2]}{\Sigma_{n=1}^{k} a_n \Sigma_n^{-1}(i)}} \end{aligned}. \tag{9}$$

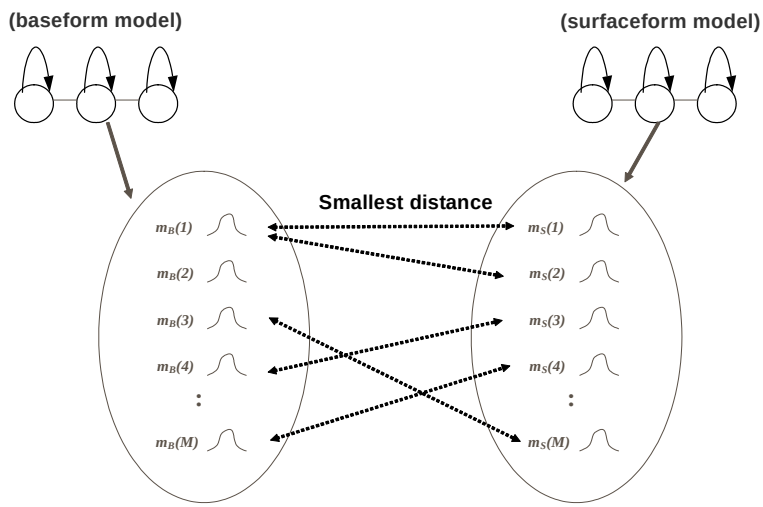

Figure 3. Mapping between baseform and surfaceform mixture components

4.3 Experimental Results and Discussion

Table 7 gives the recognition results attained with the two methods of acoustic model refinement. The VP threshold for surface-form prediction is set at 0.05. Apparently, both approaches improve recognition performance. The sharing of mixture components seems to be more effective than adaptation. However, this is at the cost of a substantial increase in model complexity. The baseline acoustic models have a total of 32,144 Gaussian components. The adaptation approach retains the same number of Gaussian components. The models obtained with the sharing approach have 37,505 components, 17% more than the baseline. If we use an equal number of components in the baseline acoustic models, the baseline word error rate will be reduced to 24.34%, and the benefit of sharing mixture components is only marginal.

Table 7. Recognition results with different methods of acoustic model refinement

	Baseline	Sharing	Adaptation
Word error rate (%)	25.34	23.96	24.70

With the adaptation approach, the baseform pdf is shifted towards the corresponding surface forms. If a surface-form pdf is far away from the baseform one, the extent of the modification will be substantial and, consequently, the modified pdf may fail to model the original baseform. On the other hand, the sharing approach has the problem of undesirably including redundant components in the baseform models. Thus we combine these two approaches. The idea is to perform adaptation using the surface-form components that are close to the baseform, and at the same time, to use those relatively distant components for sharing.

The values of the KLD between the baseform pdf and the nearest surface-form pdf have been analyzed. As illustrative examples, the histograms of the KLD at different states between [aak] (baseform) and [aa] (surface form), and between [aak] and [aat], are shown as in Figure 4. There are two main types of KLD distributions: 1) concentration around small values (e.g., states 1 and 2 of the pair "[aak]→[aa]"), and 2) a wide range of values (e.g., states 3 to 5 of the pair "[aak]→[aa]"). A small KLD means that the mixture components of the baseform and surface forms are similar. In this case, the baseform components adapt to the surface form. In the case of a widely distributed KLD, the surface-form components should not be used to adapt the baseform components, but rather should be kept along with the modified baseform model in order to explicitly characterize irregular pronunciations. In this way, a combined approach to baseform model refinement is formulated.

Despite the good intentions, the combined use of sharing and adaptation doesnot lead to favorable experimental results. With a total of 34,042 mixture components in the refined acoustic models, the word error rate is 24.57%. The baseline performance is 24.93% with the same model complexity.

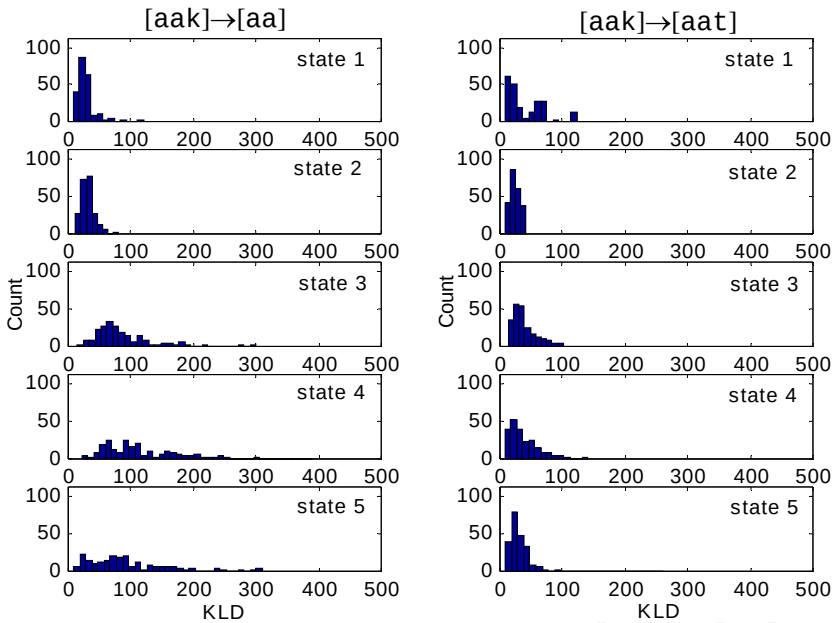

Figure 4. KLD distributions for variation pairs **[aak]→[aa]** ***and*** **[aak]→ [aat]**

5. Conclusions

In this study, we have classified pronunciation variations into phone changes and sound changes. However, these are not well defined classifications, especially for the sound changes. There is not a clear boundary that separates a phoneme substitution (phone change) from a phoneme modification (sound change). This may partially explain why the proposed techniques of handling sound change are not as effective as the methods for handling phone change.

The use of a PVD is intuitive and straightforward in implementation. It can reduce the word error rate noticeably. When constructing a PVD, the value of the VP threshold needs to be carefully determined. While a tight threshold obviously doesnot show any effect, a lax control of the PVD size leads to not only a long recognition time but also performance degradation. The method of dynamic search space expansion during decoding can bring about the same degree of performance improvement as the PVD. However, the training of context-dependent pronunciation prediction models requires a large amount of data.

The methods of acoustic model refinement donot improve recognition performance as much as we expected. Similar effect can be achieved by using more mixture components. Indeed, more mixture components can describe more complex acoustic variations, which include the variations caused by alternative pronunciations. The sharing of mixture

components is equivalent to having more mixture components right at the beginning of acoustic models training. Adaptation of mixture components is not as effective as increasing the number of mixture components.

For any of the above methods to be effective, the accurate and efficient acquisition of pronunciation variation information is most critical. Manual labeling is impractical. Automatic detection of pronunciation variations is still an open problem.

Acknowledgement

This research is partially supported by a Research Grant from the Hong Kong Research Grants Council (Ref: CUHK4206/01E).

References

Aubert, X., and C. Dugast, "Improved acoustic-phonetic modeling in Philips' dictation system by handling liaisons and multiple pronunciations," In *Proceedings of 1995 European Conference on Speech Communication and Technology*, pp.767 – 770.

Bauer, R.S., and P.K. Benedict, *Trends in Linguistics, Studies and Monographs 102, Modern Cantonese Phonology*, Mouton de Gruyter, Berlin, New York, 1997.

Bourgerie, D.S., *A Quantitative Study of Sociolinguistic Variation in Cantonese*, PhD Thesis, The Ohio State University, 1990.

Byrne, W., V. Venkataramani, T. Kamm, T.F. Zheng, Z. Song, P. Fung, Y. Liu and U. Ruhi, "Automatic generation of pronunciation lexicons for Mandarin spontaneous speech," In *Proceedings of the 2001 International Conference on Acoustics, Speech and Signal Processing*, l.1, pp.569 – 572.

Chao, Y.R., *A Grammar of Spoken Chinese*, University of California Press, 1965.

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

CUPDICT: Cantonese Pronunciation Dictionary (Electronic Version), Department of Electronic Engineering, The Chinese University of Hong Kong, http://dsp.ee.cuhk.edu.hk/speech/, 2003.

Fosler-Lussier, E., "Multi-level decision trees for static and dynamic pronunciation models," In *Proceedings of 1999 European Conference on Speech Communication and Technology*, pp.463 – 466.

Grimes, B.F. *et al.*, *Ethnologue, Languages of the World*, SIL International, 2000.

Hashimoto, O.-K. Y., *Studies in Yue Dialects 1: Phonology of Cantonese*, Cambridge University Press, 1972.

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

Huang, X., A. Acero, and H.W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR., 2001.

Kam, P., *Pronunciation Modeling for Cantonese Speech Recognition*, MPhil Thesis, The Chinese University of Hong Kong, 2003.

Kam, P., and T. Lee, "Modeling pronunciation variation for Cantonese speech recognition," In *Proceedings of ISCA ITR-Workshop on Pronunciation Modeling and Lexicon Adaptation* 2002, pp.12-17.

Kam, P., T. Lee and F. Soong, "Modeling Cantonese pronunciation variation by acoustic model refinement," In *Proceedings of 2003 European Conference on Speech Communication and Technology*, pp.1477 – 1480.

Kessens, J.M., M. Wester and H. Strik, "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation," *Speech Communication*, 29, pp.193 – 207, 1999.

Kessens, J.M., C. Cucchiarini and H. Strik, "A data driven method for modeling pronunciation variation," *Speech Communication*, 40, pp.517 – 534, 2003.

Lee, T., W.K. Lo, P.C. Ching and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, 36, No.3-4, pp.327-342, 2002

Linguistic Society of Hong Kong (LSHK), *Hong Kong Jyut Ping Characters Table (粵語拼音字表)*. Linguistic Society of Hong Kong Press (香港語言學會出版), 1997.

Liu, M., B. Xu, T. Huang, Y. Deng and C. Li, "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling," In *Proceedings of the 2000 International Conference on Acoustics, Speech and Signal Processing*, 2, pp.1025-1028.

Liu, Y., *Pronunciation Modeling for Spontaneous Mandarin Speech Recognition*, PhD Thesis, The Hong Kong University of Science and Technology, 2002.

Liu, Y. and P. Fung, "Modeling partial pronunciation variations for spontaneous Mandarin speech recognition," *Computer Speech and Language*, 17, 2003, pp.357 – 379.

Liu, Y. and P. Fung, "State-dependent phonetic tied mixtures with pronunication modeling for spontaneous speech recognition," *IEEE Trans. Speech and Audio Processing*, 12(4), 2004, pp.351 – 364.

Myrvoll, T.A. and F. Soong, "Optimal clustering of multivariate normal distributions using divergence and its application to HMM adaptation", In *Proceedings of the 2003 International Conference on Acoustics, Speech and Signal Processing*, 1, pp.552 - 555.

Riley, M., W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraçlar, C. Wooters and G. Zavaliagkos, "Stochastic pronunciation modeling from hand-labelled phonetic corpora," *Speech Communication*, 29, 1999, pp.209 – 224.

Saraçlar, M. and S. Khudanpur, "Pronunciation ambiguity vs. pronunciation variability in speech recognition," In *Proceedings of the 2000 International Conference on Acoustics, Speech and Signal Processing*, 3, pp.1679-1682.

Saraçlar, M., H. Nock and S. Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, 14, 2000, pp.137 – 160.

Strik, H. and C. Cucchiarini, "Modeling pronunciation variation for ASR: a survey of the literature," *Speech Communication*, 29, 1999, pp.255 – 246.

Venkataramani, V. and W. Byrne, "MLLR adaptation techniques for pronunciation modeling," In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* 2001, CD-ROM.

Wester, M., "Pronunciation modeling for ASR – knowledge-based and data-derived methods," *Computer Speech and Language*, 17, 2003, pp.69 – 85.

Wong, Y.W., *Large Vocabulary Continuous Speech Recognition for Cantonese*, MPhil Thesis, The Chinese University of Hong Kong, 2000.