

# Modeling Cantonese Pronunciation Variation by Acoustic Model Refinement

Patgi KAM<sup>1</sup>, Tan LEE<sup>1</sup> and Frank K. SOONG<sup>2</sup>

<sup>1</sup>Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong,

<sup>2</sup>Spoken Language Translation Labs, ATR, Kyoto, Japan  
{pgkam, tanlee}@ee.cuhk.edu.hk, frank.soong@atr.co.jp

## Abstract

Pronunciation variations can be roughly classified into two types: a phone change or a sound change [1][2]. A phone change happens when a canonical phone is produced as a different phone. Such a change can be modeled by converting the *baseform* (standard) phone to a *surfaceform* (actual) phone. A sound change happens at a lower, phonetic or subphonetic level within a phone and it cannot be modeled well by either the baseform or the surfaceform phone alone. We propose here to refine the acoustic models to cope with sound changes by (1) sharing the Gaussian mixture components of HMM states in the baseform and the surfaceform models; (2) adapting the mixture components of the baseform models towards those of the surfaceform models; (3) selectively reconstructing new acoustic models through sharing or adapting. The proposed pronunciation modeling algorithms are generic and can, in principle, be applied to different languages. Specifically, they were tested in a Cantonese speech recognition database. Relative word error rate reductions of 5.45%, 2.53%, and 3.04% have been achieved using the three approaches, respectively.

## 1. Introduction

Given a speech input, speech recognition is to produce highly probable hypotheses of the underlying word sequence. This can be done by establishing a probabilistic mapping between the observed acoustic feature vector sequence and the unknown, yet to be determined, linguistic representations. Given 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 a linguistic symbol may also be realized in different acoustic pronunciations. The variability is due to co-articulation, regional accent, speaking rate, speaking style, etc. Pronunciation modeling (PM) for automatic speech recognition (ASR) is to provide an effective mechanism to cope with such variability.

Pronunciation variations can be roughly classified into two types: a phone change and a sound change. While a phone change is the realization of a standard canonical *baseform* phone by a different *surfaceform* phone, such as replacing /b/ by /d/, a sound change is a structural variation within a phone, such as nasalization, centralization, voiceless, voiced, etc., which cannot be modeled by simply replacing the canonical phone with another (surfaceform) phone.

Phone change can be handled by replacing the baseform transcription by the actual pronunciation observed in an acoustic token, or the surfaceform transcription. This is

accomplished by augmenting the standard baseform lexicon with additional pronunciation variants for word entries or expanding the search space during sentence decoding to include those variations [3].

To handle a sound change, pronunciation modeling must be applied at a lower level, for example, at a state or Gaussian mixture component level. In most acoustic model training scenario, acoustic models are trained with only the baseform pronunciations and no alternative pronunciations are considered. This convenient but apparently wrong assumption of standard baseform pronunciations renders the acoustic models thus trained to be inadequate to represent the variations of speech sounds. It would then be useful to refine the acoustic models of a sound by taking into account its realistic pronunciations.

In this paper, we investigate three algorithms to refine the acoustic models in order to cope with sound changes. The algorithms were tested specifically in a Cantonese speech recognition database. The merit of each approach is discussed and the corresponding recognition improvements are compared.

## 2. Background

### 2.1. The Cantonese dialect

Cantonese is the dominant dialect used in Southern China, Hong Kong and many regions overseas. Like Mandarin, Cantonese is both monosyllabic, where each Chinese character (essentially a morpheme) is pronounced as a monosyllable, and tonal. While most (more than 70%) Chinese words are bisyllabic, monosyllabic words form a substantial set of frequently used words.

A Cantonese syllable has a structured form of a beginning initial (I) followed by a final (F) [4]. Altogether there are 20 initials and 53 finals. Initials and finals are combined under certain phonological constraints to form over 600 legitimate initial-final (I-F) combinations, referred here as the base syllables.

### 2.2. Acoustic model for Cantonese ASR

For Cantonese ASR, context-dependent initial and final models are usually used as the basic units for constructing acoustic Hidden Markov Models (HMMs). In this research, the acoustic models for large vocabulary, continuous speech recognition (LVCSR) are cross-word bi-IF HMMs trained with 20 hours of continuous speech database in the CUSENT corpus collected by the Chinese University of Hong Kong (CUHK) [5].

### 2.3. Pronunciation model for Cantonese LVCSR

Pronunciation models are used to derive or to predict surfaceform transcriptions from baseform transcriptions. In this paper, we denote the baseform and the surfaceform transcriptions at initial-final level as  $B$  and  $S$ , respectively.

In this study, the pronunciation model is trained as a confusion matrix which characterizes the probabilistic mappings between baseform I-F's and surfaceform I-F's. The confusion matrix is obtained by the following procedure:

1. The baseform transcription of CUSENT is obtained from the baseform dictionary with the standard pronunciation transcriptions of Cantonese words.
2. The surfaceform transcription is obtained from continuous phone recognition output.
3. Surfaceform transcriptions are then aligned with the corresponding baseform transcriptions using the dynamic programming to produce a confusion matrix. Variation probability (VP) is defined as the mapping frequency between baseform and surfaceform I-F's.
4. A threshold is set to filter out infrequent surfaceform pronunciations.

As a result, a number of possible surfaceform I-Fs is found for a given baseform I-F.

### 3. Sharing of Mixture Components

Since acoustic models are usually trained with only the baseform pronunciations, the HMMs thus trained under this convenient but incorrect assumption are not very effective to model surfaceform variations. It is then necessary to modify the baseform acoustic models to capture the variation of pronunciations. First, the surfaceform mixture components can be used to enrich the baseform models such that they have a better coverage of mixture components acoustically [6]. Second, the baseform mixture components can be adapted by incorporating the surfaceform mixture components.

In the first approach we first align the states of the baseform and surfaceform models such that each baseform state includes a mixture of surfaceform states. The output observation probability density function (pdf) is now modified by considering the contributions of the surfaceform output observation pdf's in addition to the one in the original baseform. If the output pdf of the original baseform state  $b_j$  is

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

where  $M$  is the number of Gaussian mixture components in the  $j$ -th state of the baseform, and  $w_{jm}$ , the weight for  $m$ -th mixture component of state  $j$ .

The output pdf of the modified baseform is then,

$$b_j'(O_t) = P(S=B|B) \cdot b_j(O_t) + \sum_{\substack{n=1 \\ S_n \neq B}}^N P(S_n|B) \cdot s_{nj}(O_t) \quad (2)$$

where  $N$  is the total number of surfaceform pronunciations for a particular baseform pronunciation,  $B$ ,  $s_{nj}$  is the output pdf of the  $j$ -th state of the  $n$ -th surfaceform  $S_n$ ,  $P(S=B|B)$  is the VP for a baseform realized as itself, and  $P(S_n|B)$  is the VP for a baseform realized as  $S_n$ .

The number of mixtures of the resultant baseform model depends on the number of surfaceform pronunciations. More surfaceform pronunciations will therefore bring in more

mixture components to the modified baseform output pdf. As the number of mixture components of each state is changed after the sharing process, re-estimation of mixture weights can be performed.

### 4. Mixture Component Adaptation

Although sharing mixture components yields an acoustically richer model, it also increases the model size. Hence, more memory space and higher computation complexities are needed. Moreover, if the baseform and surfaceform mixture components are very similar, including them all in the modified baseform can be unnecessarily superfluous.

We propose here to refine acoustic models by adapting the Gaussian mixtures of the baseform states. The states of the baseform and surfaceform models are aligned first to form corresponding state pairs between a baseform and associated surfaceforms. The baseform mixture component pdf's are adapted towards the nearest surfaceform mixture component pdf's. Within each state pair, we need to find the nearest neighbor correspondence between the baseform and surfaceform mixture component pdf's.

The "distance" between two pdf's can be calculated by the Kullback-Leibler divergence (KLD) [7] which is an information-theoretic measure for measuring the similarity between them. Specifically, when the two given pdf's,  $f$  and  $g$ , are multivariate Gaussian, the symmetric KLD has a closed form as

$$d(f, g) = \frac{1}{2} \text{trace}\{(\Sigma_f^{-1} + \Sigma_g^{-1})(\mu_f - \mu_g)(\mu_f - \mu_g)^T + \Sigma_f \Sigma_g^{-1} + \Sigma_g \Sigma_f^{-1} - 2I\} \quad (3)$$

where  $\mu$  and  $\Sigma$  are the mean vectors and the covariance 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. We compute the KLDs between all the possible pairs,  $(m_B(i), m_S(j))$ . Each surfaceform mixture component is paired up with the nearest baseform mixture component in KLD. Equivalently, for each  $m_S(j)$ , we find

$$\hat{i} = \arg \min_{m_B(i)} d(m_B(i), m_S(j)) \quad (4)$$

As a result, a particular baseform mixture component  $m_B(i)$  will be associated with  $k$  surfaceform components. To modify  $m_B(i)$ , first we have to find the centroid mixture component  $c_S$  of the  $k$  surfaceform mixture components. If the baseform  $B$  has  $n$  such surfaceforms,  $n$  centroids need to be generated. For each  $m_B(i)$ , we have the  $n$  associated centroids,  $c_{S_1}, \dots, c_{S_n}$ . All centroids are then combined with the corresponding baseform components to form a centroid weighted by VP. This VP weighted centroid becomes the modified baseform mixture  $m_B(i)'$ .

The mean and covariance of the weighted centroid,  $f_c$ , of  $k$  mixture components can be found by minimizing the weighted divergence

$$\{\mu_c', \Sigma_c'\} = \arg \min_{\mu_c, \Sigma_c} \sum_{n=1}^k a_n d(f_c, f_n) \quad (5)$$

where  $a_n$  is the weighting coefficient of the  $n$ -th pdf,  $f_n$ ,  $a_n$  is the mixture weight in calculating the centroid mixture for the surfaceform mixtures.  $a_n$  becomes the VP when calculating the final VP weighted optimal centroid. Similar to the

derivation in [7] when the diagonal covariances are used, the  $i$ -th component of the centroid is

$$\mu_c'(i) = \frac{\sum_{n=1}^k a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i)) \mu_n(i)}{\sum_{n=1}^k a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i))} \quad (6)$$

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

## 5. Results and Analysis

The methods described above are evaluated in a Cantonese LVCSR task. The test set of CUSENT corpus contains 1200 sentences (about 1.1 hours) recorded by 6 male speakers and 6 female speakers. The acoustic models are cross-word bi-IFs trained by using 20 hours of the CUSENT corpus. The number of Gaussian mixture components for each state is 16. Each speech frame is represented by a 39 dimensional feature vector with 12 MFCCs, log energy, and their first and second order time derivatives.

	Baseline (32144)	Sharing (37505)	Adaptation (32144)
No retrain	25.34	24.38	24.70
Retrained	N/A	23.96	N/A

Table 1. WER(%) of using two different HMM refining methods. Figures inside ( ) are the numbers of mixture components of different model sets.

The results of the two different HMM refining methods are shown as in Table 1. ‘‘Sharing’’ refers to the HMM mixture component sharing discussed in section 3 and ‘‘adaptation’’ refers to HMM mixture component adaptation discussed in section 4. It can be seen that incorporating more detail information of pronunciations in acoustic models improves a better recognition performance.

Experimentally we found that it is appropriate to prune out variation probability (VP) of low frequency surfaceforms at 5%. The KLD thresholds for HMM sharing and adaptation are set at 300 and 50, respectively. By using these thresholds, relative WER reduction is 3.79% and 2.53% for ‘‘sharing’’ and ‘‘adaptation’’, respectively. The relative error reduction can be further improved to 5.45% when the models were re-estimated for ‘‘sharing’’. As no extra mixture components are included in the models for ‘‘adaptation’’, no re-estimation was done to prevent any loss of surfaceform information due to retraining.

From the results, it is found that including surfaceform mixture components in the acoustic model refinement 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 models. 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 adapting the baseform mixture components as some of them may represent irregular or idiosyncratic pronunciations. Therefore, we may want to choose selectively those useful mixture components for refining the acoustic models rather than merge them all with the baseform mixtures.

## 6. Combination of Mixture Component Sharing and Adaptation

Both mixture component sharing and adaptation are aimed to incorporate the surfaceform mixtures into the corresponding baseform models. In the case of adaptation, the pdf’s of the baseform mixture components are shifted towards the corresponding surfaceform pdf’s. If the surfaceform pdf is far away from the baseform pdf, the adapted pdf may become worse than the original baseform (before adaptation) in modeling those correctly pronounced canonical tokens and adaptation tends to produce more harm than help. On the other hand, including surfaceform mixture components which are very close to the baseform models may be superfluous. Thus, we propose to adapt those similar components in the baseform and surfaceforms, while keep all the distinctive components for acoustic resolution by sharing.

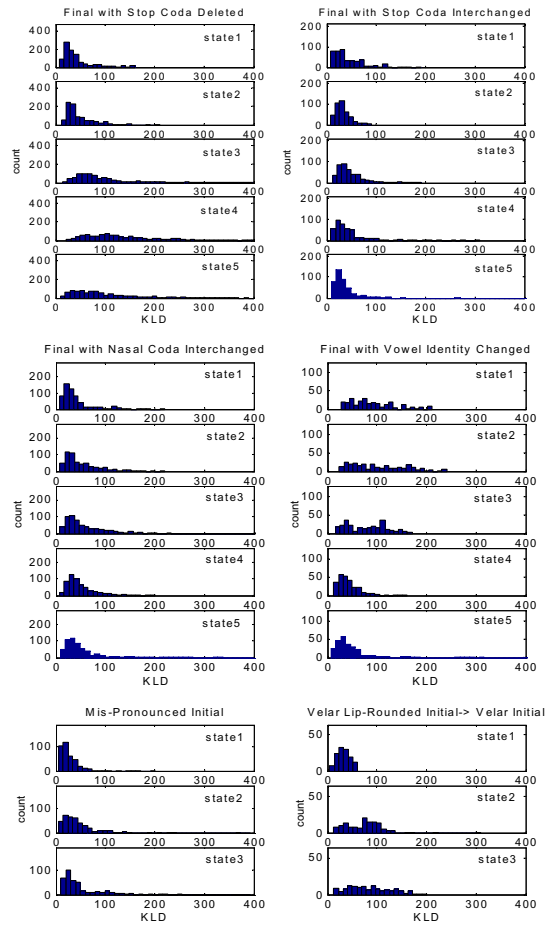


Figure 1. The distributions of KLD for different types of pronunciation variations

The distributions of KLD for different variations are shown in Figure 1. From these figures, we notice that there are two main types of distributions, one is showing consistently small KLD (<50), while the others showing a wider range of KLD distributions. For the former case, the mixture components of the base- and surfaceforms are similar, hence small KLDs. In this case, it is possible to adapt the baseform components towards the surfaceform components.

For the other case, the surfaceform components should not be used to adapt the baseform, but be included in the new baseform in order to characterize irregular pronunciations more effectively. We therefore combine the surfaceform with the baseform using either adaptation (A) or sharing (S) based upon their KLD distributions and the resultant choice is shown by the sound variations and the model state numbers in Table 2.

	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 2. Mixture combination in different states using adaptation or sharing for different pronunciation variations

## 7. Results and Analysis

The methods described above are evaluated in the same continuous Cantonese speech recognition task.

Combine (34042)	No retrain	Retrained
	24.87	24.57

Table 3. WER(%) of combining “sharing” and “adaptation”. Figure in () is the number of mixture components in the model set.

The result is shown in Table 3. Combining “sharing” and “adaptation” reduces WER by 1.85%. The relative error reduction can be further improved to 3.04% when the models were 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.

The two main types of distributions: consistently small KLD values distributed in a narrow range and a wide-range distribution of larger KLD values. They actually reflect the following pronunciation variations:

### 1) Small KLD

Small KLD is usually found when a vowel nucleus remains unchanged or a consonant initial/coda interchanged with another phone in the same phone class. For example, the first 2 states in the finals with coda deleted/interchanged, the last 3 states in the finals with stop/nasal coda interchanged, and the first few states in mis-pronounced initials, e.g. /n/→/l/.

### 2) Wide-Range KLD

These distributions are found when a vowel nucleus is changed or a stop coda deleted. For example, the first 3 states

in the finals with vowel identity changed, and the last 3 states in the finals with stop coda deleted. Such kind of distribution is also found in the last 2 states of velar lip-rounded initials mixed with velar initials, e.g. /gw/→/g/. These cases are somewhat similar to the effect of phone change, but occur at the state level.

## 8. Conclusions

In this paper, we refine acoustic models to cope with sound changes in pronunciations: sharing the baseform Gaussian mixture components with those of the surfaceform; adapting mixture components of the baseform towards those of the surfaceforms; selectively to share or to adapt the models by using the distributions information of the KLD between mixture component pair of base- and surfaceforms. Relative error reduction of 5.45%, 2.53% and 3.04% was achieved with the three approaches, respectively.

## 9. Acknowledgements

The project is partially supported by a Research Grant for the Hong Kong Research Grant Council (Ref. CUHK 4206/01E). The authors would like to give sincere thanks to Mr. Y.W. Wong, Ms. Y. Qian and Ms. C. Yang of the DSP Lab., CUHK, for their help and invaluable advices.

## 10. References

- [1] M. Saraclar and S. Khudanpur, “Pronunciation Ambiguity VS Pronunciation Variability in Speech Recognition”, in *Proceedings of ICASSP-00*, Vol.3, pp.1679-82, 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] P. Kam, “Modeling Pronunciation Variation for Cantonese Speech Recognition”, in *Proceedings of PMLA-02*, pp.1445-8, Denver, 2002.
- [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] W.K. Lo, T. Lee and P.C. Ching, “Development of Cantonese Spoken Language Corpora For Speech Applications”, in *Proceedings of ISCSLP-98*, pp.102-7, Singapore, 1998.
- [6] M. Saraclar, “Pronunciation Modeling by Sharing Gaussian Densities across Phonetic Models”, in *Proceedings of Eurospeech-99*, Vol.1, pp.515-8, Hungary, 1999.
- [7] T.A. Myrvoll and F. K. Soong, “Optimal Clustering of Multivariate Normal Distributions Using Divergence and its Application to HMM Adaptation”, in *Proceedings of ICASSP-03*, Vol.1, pp.552-5, Hong Kong, 2003.