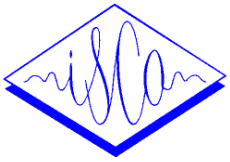

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

# PARTIAL CHANGE PHONE MODELS FOR PRONUNCIATION VARIATIONS IN SPONTANEOUS MANDARIN SPEECH

*LIU Yi, Pascale Fung*

Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology, Hong Kong
{eelyx, pascale@ee.ust.hk}

## ABSTRACT

Modeling pronunciation variations is a critical part of spontaneous Mandarin speech recognition. Such variations include both complete changes and partial changes. Complete pronunciation changes can usually be modeled by using an alternative phone to replace the canonical phoneme. Partial changes are variations within the phoneme and include diacritics, which cannot be modeled by conventional methods. In this paper, we propose using *partial change phone models* to represent such changes. The pre-trained acoustic model is reconstructed by sharing Gaussian mixtures between canonical phone models and partial change phone models at the state level. We improve the resolution of the acoustic model to accommodate partial changes. The effectiveness of this approach is evaluated on the Hub4NE Mandarin Broadcast News Corpus. The syllable accuracy increased 2.59% absolutely with respect to the baseline.

## 1. INTRODUCTION

Spontaneous speech has large variations in pronunciation, which can be caused by a speaker's accent, speaking style, speaking mode and speaking rate [1]. The variations include phonetic shifts, reduction and assimilation, duration changes, etc. All pronunciation variations can be classified into two types: complete changes and partial changes [2,5]. Complete changes - phone changes are the replacement of a canonical phoneme by another alternate phone. Partial changes - sound changes are the variations within the phoneme. A large number of variations in spontaneous Mandarin speech are partial changes. For example, Chinese initials are very flexible and around 30% of the variations are sound changes [5].

Partial changes have always been ignored in pronunciation modeling since they are hard to capture. Most of the current work on pronunciation modeling attempts to improve the recognition accuracy by predicting pronunciation variations, so that each word is allowed to have alternative phonetic representations [1,3,4]. This approach can only model complete changes but not partial changes. In particular, conventional phoneme units cannot be appropriately used to represent partial changes since partial changes always occur within the phoneme. Neither the canonical phoneme nor the alternative phone can represent partial changes very well.

In our previous work at the Johns Hopkins Summer Workshop on Speech and Language Technologies 2000, we proposed using Generalized Initial/Final (GIF) to model partial changes [2]. GIF is an extended set of the common Initial/Final (IF), which includes the canonical phoneme set as well as the extended phone set. GIF set is then used by linguists at Chinese Academy of Social Sciences to label the surface form of a spontaneous Mandarin corpus CASS [5]. However, there are still challenges for partial change modeling using GIF models: first, it is difficult even for phoneticians to clearly identify which particular sound change has occurred. Second, it is very time consuming to generate hand-labeled transcription for acoustic model training. Finally, the sparse data problem still exists for acoustic model training since there are only a limited number of training samples for GIF models.

In [7], Saraclar used Gaussian mixture sharing between baseform model ( $P(X \mid B)$ ) and surface form models ( $P(X \mid S)$) for modeling pronunciation variations. However, in this approach, the surface form model is based on the whole phoneme units and therefore cannot clearly describe partial changes. In this paper, we propose modeling partial pronunciation changes by Gaussian mixture sharing between canonical phone models and *partial change phone models*. Instead of surface form models, partial change phone models are established from samples obtained through DP alignment between baseform and surface form transcriptions. These surface form transcriptions are generated from the forced alignment.

The paper is organized as follows: section 2 introduces partial pronunciation changes in spontaneous Mandarin speech. In section 3, we describe how to generate partial change phone models. The HMM model reconstruction mechanism by Gaussian mixture sharing between the canonical baseform models and partial changes phone models at the state level is shown in section 4. Finally syllable recognition experiments are presented in section 5. We conclude in section 6.

## 2. PARTIAL PRONUNCIATION CHANGES

The recognition performance of current ASR systems on spontaneous speech is relatively low compared to that on carefully read speech. This is because spontaneous speech has a higher degree of pronunciation variability. Linguistic knowledge and empirical results show that pronunciation variations in Mandarin can be classified into two types: phone changes and sound changes [2,5]. Phone changes are the replacement of a canonical phoneme by another alternate phone, such as 'b' being pronounced as 'p'. Sound changes are variations within the same phoneme, such as nasalization, centralization, voiceless, voiced, rounding, etc. For example, 'ts' can change into 'ts_v' (voiced) and 'b' to 'b_m' (nasalization) or 'b_h' (voiceless). In Li's work [5], they defined 10 types of sound changes in spontaneous Mandarin speech.

Sound changes are very common in spontaneous Mandarin speech. When sound changes occur, a phone is not completely substituted, deleted or inserted. An analysis of semi-syllable tier transcriptions of CASS corpus shows that the average transcriber agreement is around 84.23% and the majority of the disagreement is caused by partial changes [5]. This suggests that when partial changes occur, the surface form cannot be clearly identified. In addition, a detailed annotation in CASS transcription at the semi-syllable level shows that the percentage of partial changes relative to standard pronunciation is around 20% [5]. For Chinese initials, which have very flexible pronunciations in spontaneous speech, the percentage reaches 27.46%.

Partial changes are a little more difficult to be modeled as this diacritic set of phones can only be trained by the samples labeled initially by humans. GIF set is a solution for partial change modeling. Table 1 gives an example for IFs and their partial changes (GIF). However, using GIF model also has its obstacles for partial change modeling – the inventory of HMM units is enlarged, lack of training data for acoustic model training, the variations are heavily dependent on the hand-labeled transcriptions.

| IF (Pinyin) | GIF | Comments |
|---|---|---|
| z | /ts/ | Canonical |
| z | /ts_v/ | Voiced |
| z | /ts`/ | Changed to '*zh*' |
| z | /ts`_v/ | Changed to voiced '*zh*' |
| e | /7/ | Canonical |
| e | /7`/ | Retroflexed, or changed to '*er*' |
| e | /@/ | Changed to /@/ (a GIF) |

Table.1: An example of IFs and their partial changes (GIF)

Clearly, partial changes should be considered for pronunciation modeling in spontaneous speech. Simply enlarging the HMM unit set from hand-labeled transcriptions and attempting to use an optimal phone-level symbol to represent partial change are insufficient. In order to model partial changes, it requires that the uncertainty of the surface form representation caused by partial changes should be taken into account. We need to improve the resolution of the acoustic model to cover partial changes, and the identity of the model cannot be sacrificed at the same time.

## 3. PARTIAL CHANGE PHONE MODELS

### 3.1 Basic Idea

Our goal is to generate *partial change phone models* to accommodate partial pronunciation variations.

Let $B$ be the baseform sequence, $S$ the surface form sequence and $X$ the input speech vector. The decoding formula is

$$B^* = \arg\max_B P(B)P(X \mid B) \tag{1}$$

If pronunciations were always same, there is no need to consider pronunciation variations. The decoding would be relatively easy as shown in Eq.1. However, since pronunciations are always different in spontaneous speech, Eq.1 needs to be rewritten by taking pronunciation model into consideration

$$B^* = \arg\max_B P(B)\sum_S P(X \mid B,S)P(S \mid B) \tag{2}$$

Note that $P(B)$ is the language model, $P(X \mid B,S)$ is the acoustic model and $P(S \mid B)$ is the pronunciation model. In a general acoustic model training procedure, we assume that

$$P(X \mid B,S) \approx P(X \mid B) \tag{3}$$

This means that the acoustic model is trained with baseform transcriptions. If surface form transcriptions are available, the acoustic model training can be expressed as

$$P(X \mid B,S) \approx P(X \mid S) \tag{4}$$

It is obvious that both Eq.3 and Eq.4 are sub-optimal acoustic models. In fact, estimating acoustic model either from baseform or from surface form transcriptions is an approximation. Ideally, as shown in Eq.2, we should take both the baseform and surface form into consideration for acoustic model estimation when pronunciation variations are considered. Thus, partial change phone model ( $P(X \mid B,S)$ ) and the relevant transcriptions in terms of baseform/surface form phone pairs are required.

The disagreement of surface form transcription between linguist transcribers reveals that when partial changes occur, the identity of the surface form cannot be clearly defined. Using conventional phoneme set and estimating acoustic model either from the baseform or from the surface form transcription cannot differentiate partial changes. However, partial change phone models depending on both the baseform and surface form can efficiently differentiate partial variations. For example, partial change phone models 'b_d', 'b_f' and 'b_m' may accommodate different partial variations representing centralization, voiceless and nasalization, respectively, with respect to the baseform model 'b'.

### 3.2 Generation of Partial Change Phone Models

Partial change phone models are based on baseform/surface form phone pair transcriptions. The transcription in terms of phone pairs is generated from baseform and surface form alignment. In general, hand-labeled phone transcription is required as surface form. However, the amount of available hand-labeled transcriptions is limited and insufficient for acoustic model and pronunciation model training. Our method, described in [4], uses the limited hand-labeled data as bootstrap material. First, an initial set of pronunciation models trained on the CASS hand-labeled transcriptions will generalize well enough so that they contain pronunciation variability in the Mandarin Broadcast News domain [4]. The pronunciation model is then used to generate pronunciation networks and the most likely phone sequence of the utterance in training set of Mandarin Broadcast News domain can be achieved by the forced Viterbi alignment as shown in Eq.5:

$$S^* = \arg\max_S P_{AM}(X \mid S)P_{PM}(S \mid B) \tag{5}$$

Phone pair transcriptions then can be generated through DP alignment between baseform and surface form transcriptions. In addition, this alignment yields the inventory of partial change phone models, such as 'b_p' and 'b_f', and context information can be applied to accommodate partial changes caused by co-articulation, which is shown as follows.

Let $B_i$ be the current baseform phoneme and $B_{i-1}$ the left-context baseform phoneme. The phone level pronunciation model can be expressed as

$$P(S \mid B_i) = \sum_{B_{i-1}} P(S \mid B_i, B_{i-1}) P(B_{i-1} \mid B_i) \quad (6)$$

$P(S \mid B_i, B_{i-1})$ is the variation probability given the current and left-context phoneme. The term $P(B_{i-1} \mid B_i)$ is similar to bigram or the transition probability between phone units. Therefore, the context-dependent phone pair unit can be defined as $(S, B_i, B_{i-1})$. In order to limit the complexity of the model and avoid the sparse data problem, we ascribe different $B_{i-1}$ with similar acoustic representation to the same class. The definition of the phone class in Mandarin speech can be found in [2]. If the samples for partial change phone model are still insufficient for acoustic model training using the phone classes, the next approximation is applied:

$$\sum_{B_{i-1}} P(S \mid B_i, B_{i-1}) P(B_{i-1} \mid B_i) \approx \max_{B_{i-1}} P(S \mid B_i, B_{i-1}) P(B_{i-1} \mid B_i) \quad (7)$$

Combined with Eq.6, the acoustic model expressed in Eq.1 can be rewritten as

$$P(X \mid B_i) = \sum_S P(X \mid B_i, S) \sum_{B_{i-1}} P(S \mid B_i, B_{i-1}) P(B_{i-1} \mid B_i) \quad (8)$$

If the approximation shown in Eq.7 is introduced, the acoustic model is

$$P(X \mid B_i) \approx \sum_S P(X \mid B_i, S) \max_{B_{i-1}} P(S \mid B_i, B_{i-1}) P(B_{i-1} \mid B_i) \quad (9)$$

In Eq.9, the first term of the right hand is partial change phone model and the second term is the pronunciation model. Initial parameters of partial change phone model can be cloned from its relevant baseform model and re-estimated using BW algorithm with generated phone pair transcriptions.

The summation of all alternative surface forms represented in Eq.9 suggests that we should take all possible surface forms into baseform models to accommodate partial changes. It leads to a novel approach for partial change modeling: we reconstruct the pre-trained baseform HMM models by Gaussian mixtures sharing between partial change phone models and the canonical models, improve the resolution of the baseform acoustic model so as to cover partial changes. Partial change phone model is not regarded as an independent model but a hidden variable.

## 4. ACOUSTIC MODEL RECONSTRUCTION

### 4.1 Acoustic Model Reconstruction

We reconstruct the pre-trained baseform HMM model by sharing Gaussian mixtures between the baseform model and partial change phone models and enable the baseform model to acquire the ability from partial change phone models to cover partial changes.

Suppose the focus is on continuous density HMMs. Let $x$, $b$ and $s$ be input vector, baseform state and surface form state, respectively. $P(x \mid b)$ is the output distribution of $b$ and $P(s \mid b)$ is the pronunciation model.

The state output distribution is the mixture of Gaussians

$$P(x \mid b) = \sum_j w_{jb} f(x, \mu_j, \sum_j) \quad (10)$$

where $w_{jb}$ is the mixture weight of the *jth* mixture component. In the following equations, we use $f_j(\cdot)$ to represent $f_j(x, \mu_j, \sum_j)$ for simplification.

Let $P'(x \mid b)$ be the new output distribution of the reconstructed HMM model. Taking partial change phone model into consideration, we have

$$P'(x \mid b) = \lambda P(x \mid b) + (1-\lambda) P(x \mid b, s) P(s \mid b) \quad (11)$$

Since one canonical phoneme relates to multiple baseform/surface form phone pairs, Eq.11 can be expressed as

$$P'(x \mid b) = \lambda P(x \mid b) + (1-\lambda) \sum_k P(x \mid b, s_k) P(s_k \mid b) \quad (12)$$

where $k = 1, 2, \cdots, M$, $M$ is the total number of partial change phone models corresponding to one baseform state model. Then we have

$$\begin{aligned} P'(x \mid b) &= \lambda \sum_j w_{jb} f_j(\cdot) + (1-\lambda) \sum_k \sum_i w_{i,(b,s_k)} f_i(\cdot) P(s_k \mid b) \\ &= \sum_j w'_{jb} f_j(\cdot) + \sum_k \sum_i w'_{i,(b,s_k)} f_i(\cdot) \end{aligned} \quad (13)$$

In order to keep the total mixture number of the reconstructed acoustic model at a practical level without losing partial variation information, we use a recently introduced method [6] to select the dominant Gaussians from partial change phone models for model reconstruction. Let $g^i_{(b,s_k)}$ be the *ith* dominant Gaussian selected from the partial change phone model, we have

$$P'(x \mid b) = \sum_j w'_{jb} f_j(\cdot) + \sum_k w'_{i,(b,s_k)} g^i_{(b,s_k)} \quad (14)$$

$w'_{jb}$ and $w'_{i,(b,s_k)}$ are new mixture weights of state $b$ in the reconstructed model, they are

$$\begin{aligned} w'_{jb} &= \lambda \cdot w_{jb} \\ w'_{i,(b,s_k)} &= P(s_k \mid b) \cdot (1-\lambda) \cdot w_{i,(b,s_k)} \end{aligned} \quad (15)$$

$\lambda$ is the normalized rate to make sure that the sum of new mixture weight equals to 1. Eq.14 shows that after acoustic reconstruction, the distribution of the new constructed model includes both the canonical and alternative realizations. In addition, the weight of shared mixture components is governed by pronunciation model.

### 4.2 Acoustic Model Re-estimation

Parameters of the reconstructed acoustic model can be re-estimated using the conventional Baum-Welch algorithm. All

configurations for re-training strictly follow those used in pre-trained procedures. Thus, the pure effect given by the partial change modeling can be evaluated. Mixture weight of the reconstructed acoustic model can be initialized with Eq.15 and reestimated in the following iterations using Baum-Welch algorithm.

## 5. SYLLABLE RECOGNITION EXPERIMENTS

The first two CDs of Hub4NE 1997 Broadcast News Corpus were used to evaluate the effectiveness of our approach. HTK toolkit was used to train context-dependent (CD) initials and context-independent (CI) finals acoustic model. The HMM topology was three-states, left-to-right without skips. The total number of CD-initials and CI-finals was 139. In addition, 415 standard Chinese syllable without tone were used in the system. The acoustic features were $13MFCC$ , $13\Delta MFCC$ and $13\Delta\Delta MFCC$ . The acoustic training set consisted of 10 hours of speech (10,483 utterances) and the testing set was 724 utterances apart from the training set. Most of the training and testing utterances were spontaneous and conversational speech.

The baseform/surface form phone pair transcriptions were generated using the flexible DP alignment [2]. After filtering the sparse phone pairs, the number of selected phone pairs is 437. Therefore the total number of partial change phone models is 437, which cover the majority of partial pronunciation variations in spontaneous Mandarin speech. The baseline system is 12 Gaussians per state HMM. The reconstructed model starts from 6 Gaussians HMM. After sharing the Gaussian mixtures between the canonical baseform model and partial change phone models, the original 2502 mixture components are increased to 4754. There are 11.4 Gaussian mixtures per state of constructed HMMs on average, which is comparable to the baseline. Table 2 gives a comparison of the syllable recognition accuracy using different acoustic models. In Table 2, each item under acoustic model means:

- $P(X \mid B)$ is estimated from the baseform transcriptions, baseline acoustic model
- $P(X \mid S)$ is estimated from the surface form transcriptions, surface form trained model
- $P(X \mid B,S)$ is estimated from phone pair transcriptions, partial change phone model
- $P(X \mid B)+P(X \mid S)$ baseline model is merged with the surface form trained model, accounting for complete changes [7]
- $P(X \mid B)+P(X \mid B,S)$ baseline model is merged with partial change phone models, accounting for partial changes
- $P(X \mid B)+P(X \mid B,S)+P(X \mid S)$ baseline model is merged with both the surface form trained model as well as partial change phone models, accounting for complete changes and partial changes

In Table 2, it shows that after HMM reconstruction by sharing Gaussian mixtures between the canonical baseform model and partial change phone models, the syllable accuracy improves 2.59% absolutely with respect to the baseline, while the syllable accuracy improves 1.03% for only considering complete changes. If both complete changes and partial changes are modeled, the syllable accuracy improves 2.84%.

| Acoustic model | Syllable accuracy |
|---|---|
| $P(X \mid B)$ | 66.18% |
| $P(X \mid S)$ | 64.53% |
| $P(X \mid B)+P(X \mid S)$ | 67.21% |
| $P(X \mid B)+P(X \mid B,S)$ | 68.77% |
| $P(X \mid B)+P(X \mid B,S)+P(X \mid S)$ | 69.02% |

Table.2: modeling partial changes is better than modeling complete change

We find that modeling partial changes achieves a higher recognition performance compared with those only modeling complete changes. The reason lies in the fact that partial change phone models cover not only the partial changes but also some of the complete changes. In some cases, partial changes can be very close to complete changes. For instance, a particular partial change between 'zh' and 'z' (its partial change phone model is 'zh_z') can be very close to a complete change 'zh' in some actual pronunciations. Note that pronunciation model techniques described in [4,7] can only model complete changes but not partial changes.

## 6. CONCLUSION

Partial change phone models were used to represent partial variations in spontaneous Mandarin speech. We reconstructed the pre-trained baseform model by sharing Gaussian mixtures between the canonical baseform model and partial change phone models at the state level to accommodate partial changes. The experimental results have shown that using partial change phone models to improve the resolution of the acoustic model through model reconstruction is an efficient way to accommodate partial changes. At the same time, the sparse data problem can be avoided by using the hand-labeled transcriptions as the bootstrap material. Our approach can be easily extended to other languages although it is applied in spontaneous Mandarin speech.

## 7. REFERENCES


[1] M.Finke, J.Fritsch, D.Koll and A.Waibel, "Modeling and Efficient Decoding of Large Vocabulary Conversational Speech", *Proc.Eurospeech99*, 1999

[2] P.Fung, W.Byrne, et.al, "Pronunciation Modeling of Mandarin Casual Speech", Final report at the ws00 of Johns Hopkins summer workshop, 2000

[3] W.Byrne, et.al, "Pronunciation Modelling Using a Hand-Labelled Corpus for Conversational Speech Recognition", *Proc. ICASSP98*, 1998

[4] W.Byrne, et al., "Automatic Generation of Pronunciation Lexicons for Mandarin Spontaneous Speech", *Proc. ICASSP01*, 2001

[5] A.Li, et al., "CASS: A Phonetically Transcribed Corpus of Mandarin Spontaneous Speech", *Proc. ICSLP00*, 2000

[6] Y.Liu and P.Fung, "Estimating Pronunciation Variations from Acoustic Likelihood Score for HMM Reconstruction", *Proc. Eurospeech01*, 2001

[7] M.Saraclar, H.Nock and S.Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models", *Computer Speech and Language*, (2000) 14, 137-160

[8] M.Riley and A.Ljolje, "Automatic generation of detailed pronunciation lexicons", *Automatic Speech and Speaker Recognition: Advanced Topics*, chapter 12 pages 285-302. Kluwer Academic Press, 1995