MODEL PARTIAL PRONUNCIATION VARIATIONS FOR SPONTANEOUS MANDARIN SPEECH RECOGNITION

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 changes can usually be modeled by using an alternate phone to replace the canonical phone. Partial changes, which cannot be modeled by conventional methods, are variations within the phoneme and include diacritics. In this paper, we propose using partial change phone model (PCPM) as well as auxiliary decision tree to model partial changes. A detailed but robust model can be achieved by merging canonical model with PCPMs through Gaussian distribution reconstruction. The effectiveness of this approach was evaluated on the Hub4NE Mandarin Broadcast News Corpus. The syllable error rate decreased 2.39% absolutely with respect to the baseline.

1. INTRODUCTION
Spontaneous speech has large variations in pronunciation, which may 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 [2,5,7]. 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].

Most of the current work on pronunciation modeling attempts to improve recognition accuracy by predicting the pronunciation variations, so that each word is allowed to have alternate phonetic representations [1,3,4]. This approach can only model complete pronunciation changes but not partial pronunciation changes. In particular, partial changes occur within the phoneme and cannot be successfully represented at the phone level. Very recently, state level pronunciation modeling was proposed to model partial changes [6,7]. However, in current state level pronunciation modeling scheme, the unit set of surface models is based on the same phoneme set and therefore cannot represent partial changes very well. Furthermore, since partial changes always occur within the phoneme, neither the canonical nor the alternate phone can accurately represent such changes. In addition, although state level pronunciation modeling can improve the resolution of the baseform models, however, the model confusion may be introduced when sharing Gaussian densities between the canonical model and surface model. At the same time, such improvement is based on the improving of the Gaussian mixture numbers in HMMs, which inflates HMMs and introduces more parameters.

In this paper we propose modeling partial pronunciation changes by generating partial change phone model (PCPM) and auxiliary decision tree. Instead of baseform or surface form models, partial change phone models take both baseform and surface form into consideration and are generated from samples through DP alignment. Auxiliary decision trees, established for triphones of PCPM, are not used independently but are merged into conventional decision trees of original baseform models. That is, we reconstruct the baseform models and improve their resolution to cover partial changes accommodated by PCPMs, yet do not sacrifice their identity.

The paper is organized as follows: in section 2 we introduce partial pronunciation changes in spontaneous Mandarin speech. In section 3, we describe how to generate PCPMs. In section 4 we report the generation of auxiliary decision trees and the mechanism of tree merging. The improved performance in spontaneous Mandarin speech is discussed in section 5. Finally, we conclude in section 6.

2. PARTIAL VARIATIONS IN SPONTANEOUS MANDARIN SPEECH

Linguistic knowledge and empirical results show that pronunciation variations in Mandarin can be classified into two types: phone changes and sound changes. 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, gernalization, voiceless, voiced, rounding, etc. In Li’s work [5], they defined 10 types of sound changes in spontaneous Mandarin speech. Sound changes are very flexible 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 database [2,5] shows that the average transcription agreement is around 84.23% and the majority of the disagreement is caused by partial pronunciation changes. This suggests that when partial changes occur, the surface form cannot be clearly identified.

The following example gives more details about the difference between phone changes and sound changes. The Chinese word ‘这种’ (‘this kind’) which has a standard
pronunciation /zhae zhung/ can be pronounced as [z e zh ong] if spoken in the Wu dialect (Surface form transcriptions are set off using brackets, whereas for baseform we use slashes). The first initial /zha/ is realized as [z]. The retroflexed affricative becomes the dental-velar. Such variation can be represented at different levels. If the acoustics of /zha/ is sufficiently similar to that of /jə/, the variation is considered as phone changes. Phone changes can be modeled by adding an extra transcription [z e zh ong] of the word /keitʃə/ into the dictionary. However, if the acoustics of /zha/ is between /jə/ and /j/, the variation is regarded as a sound change since no phoneme unit can be used to accurately represent it. In this case, different transcribers may generate different surface form due to the ambiguity between /zha/ and /jə/.

It has been shown that enlarging the phone set and using more alternative symbols to represent partial variations [2] and attempting to use more refined acoustic models trained from accurate surface form transcriptions [6] are of little benefit. On the other hand, in order to model partial changes the pronunciation model cannot be operated at the phone level, and it requires the uncertainty of the surface form should be taken into account. The acoustic model should be robust enough to cover the possible partial changes which cannot be represented using conventional phoneme units.

3. PARTIAL CHANGE PHONE MODELS

3.1 Motivation for using partial change phone models

We start from the recognition formulae and deduce the representation of PCPMs. Let $X^N = x_1, x_2, \ldots, x_T$ be the input speech vector, $B^N = b_1, b_2, \ldots, b_N$ the baseform sequence and $S^N = s_1, s_2, \ldots, s_N$ the surface form sequence, $N$ is the number of phonemes in the utterance. The decoding formula is

$$\hat{B}^N = \arg \max_{B^N} P\left( X^T | B^N \right) P\left( B^N \right)$$

(1)

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

$$\hat{B}^N = \arg \max_{B^N} P\left( B^N \right) \sum_{S^N} P\left( X^T | B^N, S^N \right) P\left( S^N | B^N \right)$$

(2)

Note that in Eq. (2), $P\left( B^N \right)$ is the language model, $P\left( X^T | B^N, S^N \right)$ is the acoustic model and $P\left( S^N | B^N \right)$ is the pronunciation model. In general acoustic model training procedure it is assumed that

$$P\left( X^T | B^N, S^N \right) = P\left( X^T | B^N \right)$$

(3)

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

$$P\left( X^T | B^N, S^N \right) = P\left( X^T | S^N \right)$$

(4)

Obviously, both $P\left( X^T | B^N \right)$ and $P\left( X^T | S^N \right)$ are sub-optimal acoustic models if pronunciation variations are considered. In fact, estimating acoustic model either from the baseform or from the surface form transcriptions is an approximation. Ideally, as shown in Eq. (2), both the baseform and surface form should be taken into consideration for acoustic model estimation. Thus, PCPM and related transcriptions in terms of baseform/surface phone pairs are required.

The whole acoustic probability $P\left( X^T | B^N, S^N \right)$ can be factorized into successive contributions. Let $t_i$ and $t_i-1$ denote the start and end time of the realization of each unit,

$$P\left( X^T | B^N, S^N \right) = \prod_{i=1}^{N} P\left( X_{t_i-1}^{t_i} | b_i, s_i \right)$$

(5)

In Eq. (5), $P\left( X_{t_i-1}^{t_i} | b_i, s_i \right)$ is the partial change phone model, which takes both the baseform and the surface form into consideration. Therefore, using PCPMs which depend on both the baseform and surface form can clearly differentiate partial changes. For example, PCPMs ‘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 Partial change phone models generation

PCPMs are based on the baseform/surface form phone pair transcriptions. The transcriptions in terms of phone pairs are generated from the baseform and surface form alignment. The procedures are shown as follows:

Generate baseform phonemic transcriptions. The baseform phonemic transcriptions are obtained by looking up a canonical word-to-phoneme pronunciation dictionary.

Generate surface form phonetic transcriptions. Generally, hand-labeled phonetic transcriptions are used as the surface form. However, the amount of available hand-labeled transcriptions is always insufficient for acoustic and pronunciation model training. Our method described in [4], uses the hand-labeled data as bootstrap and automatically generates phonetic transcriptions from phone recognition.

Align baseform and surface form transcriptions. A flexible dynamic programming (DP) tool [2] is used for phoneme-to-phone alignment.

Obtain the inventory of PCPMs. The inventory of PCPM is based on the mapped baseform/surface form phone pair through DP alignment. In order to avoid the sparse data problem, the phone pair occurrences below the threshold are discarded.

Generate baseform/surface form phone pair transcriptions. If the phoneme in baseform transcriptions has a related alternate phone in surface form transcriptions, and such
A phoneme-phone pair can be found in the inventory of PCPM at the same time, the phoneme in the baseform transcription is replaced by the phone pair unit.

**Train acoustic model of PCPMs.** The initial parameters of PCPM are cloned from related baseform models and re-estimated using the Baum-Welch (BW) algorithm with phone pair transcriptions.

**4. ACOUSTIC MODEL RECONSTRUCTION**

Partial variations can be well represented by using PCPMs. However, due to the limitation of phone pair occurrence in the training data i.e. around 20% of the variations are partial changes, we cannot achieve robust acoustic models if PCPMs are used independently. Therefore, rather than direct use of PCPMs during recognition, the original canonical model is reconstructed by merging with PCPMs through Gaussian distribution reconstruction. Accordingly, the uncertainty of the surface form is taken into account, and PCPM is regarded as a hidden model but not an independent model.

**4.1 Auxiliary decision tree generation**

If context-independent models are used, the model reconstruction can be easily determined by coping each Gaussians in PCPM to the related Gaussian distribution of baseform model. However, for context-dependent models, Gaussians from PCPMs cannot be simply copied to the related baseform models because the relationship between triphones of PCPMs and baseform model is **multiple-to-one**, while for context-independent model it is **multiple-to-one**. Therefore, a decision tree based scheme is required to combine the triphone baseform models with triphone of PCPMs.

In our system, the trees for PCPM triphones are called **auxiliary decision trees**, while trees for the baseform triphones are called **conventional decision trees**. The structure of auxiliary decision trees is similar to that of conventional decision trees. They are phonetic binary trees in which a yes/no phonetic question is attached to each node. Such trees are built using a top-down sequential optimization process. The phonetic question set used for tree splitting is based on the phonetic knowledge of Mandarin. More details about these question set can be found in our previous work [2]. On the other hand, compared with conventional decision trees, auxiliary decision trees are only used during state-tying for PCPM triphone models, while not used in decoding since these trees are merged into conventional decision trees before decoding.

**4.2 Auxiliary decision tree merging**

Since PCPM is not used as an independent model, the auxiliary decision trees need to be merged into conventional decision trees before decoding. From this, the original baseform model can be reconstructed to accommodate the partial changes modeled by PCPMs.

The first step of tree merging is to find the mapping leaf nodes between the auxiliary decision tree and the related conventional decision tree. The criterion for determining the mapping node is based on Minimum Gaussian Distance between two tied states. Suppose we focus on continuous density HMMs, the distance is calculated as

$$d(i,j) = \frac{1}{S} \sum_{s=1}^{S} \left[ \frac{1}{V_s} \sum_{k,j} \left( \frac{\mu_{ik} - \mu_{jk}}{\sigma_{ik} \sigma_{jk}} \right)^2 \right]^{\frac{1}{2}}$$

where $V_s$ is the dimensionality of stream $s$, $\mu$ and $\sigma$ are mean and variance, respectively. $i$ denotes the leaf node of an auxiliary decision tree, and $j$ the leaf node of a conventional decision tree. According to Eq. (6), each node in the auxiliary decision tree can find its related node in the conventional decision tree based on the smallest distance value.

**Fig. 1 Merging of a conventional decision tree with auxiliary decision trees**

Determined by the mapping pairs, leaf nodes of a conventional decision tree are merged with those of auxiliary decision trees. Therefore, canonical models are reconstructed; they are robust and able to capture partial changes which are represented by the nodes i.e. tied-states in auxiliary decision trees. Note that not all leaf nodes of conventional decision trees have mapping nodes. In fact, some nodes may have several mapping nodes while some nodes may have none. The number of mapping nodes is based on the state confusion. In other words, the more confusible a state is, the more leaf nodes from the auxiliary decision tree it gets. The detailed effect of auxiliary decision tree merging is illustrated in Fig. 1. Readers can refer to our previous work [2] for more details about the formule of reconstructing the output distributions.

**5. RECOGNITION EXPERIMENTS**

The acoustic training set consisted of 10 hours of speech (10,483 utterances) selected from the first two CDs in the LDC Hub4NE 1997 Mandarin Broadcast News Corpus. The testing set consisted of two parts: the first one (test_set1) included 865 spontaneous utterances which were apart from training set selected from the first two CDs. The second one (test_set2) was 1263 clean utterances (F0 condition) from the 1997 and 1998 Hub4NE evaluation sets [2,4]. HTK toolkit was used to train context-dependent triphone models. The HMM topology was three-states, left-to-right without skips. The acoustic features were **MFCC** , **ΔMFCC** and **ΔΔMFCC** . The HTK flat-start procedure was used to build the 10 Gaussians model, state clustered HMMs with 2904 states.

According to the procedures illustrated in section 3.2, 145 context-independent PCPMs were selected. Triphones were generated for 145 PCPMs, and 818 tied-states were for auxiliary decision trees after state-tying. These states needed to be
merged into 2904 states of the baseline model according to Fig. 1 for model reconstruction. Hence, the reconstructed model still had 2094 tied-states while 37,220 ((2904+818)*10) Gaussians. On average, each state had 12.8 Gaussians. In order to make a fair comparison, the Gaussian number per state of the baseline was increased from 10 to 13, we called such model as enhanced HMMs. The recognition performance was listed in Table 1.

<table>
<thead>
<tr>
<th>system</th>
<th>Test_set1</th>
<th>Test_set2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>42.23</td>
<td>30.92</td>
</tr>
<tr>
<td>Baseline HMMs &amp; pronunciation dictionary</td>
<td>41.66</td>
<td>30.64</td>
</tr>
<tr>
<td>Independent use of PCPMs</td>
<td>43.37</td>
<td>30.58</td>
</tr>
<tr>
<td>Enhanced HMMs</td>
<td>41.57</td>
<td>30.47</td>
</tr>
<tr>
<td>SLPM</td>
<td>41.29</td>
<td>30.05</td>
</tr>
<tr>
<td>Merging auxiliary trees to baseline</td>
<td>39.84</td>
<td>29.68</td>
</tr>
</tbody>
</table>

Table 1: Recognition performance of modeling partial changes compared with other pronunciation modeling methods.

First, the performance of modeling complete changes is indicated in the second row. The pronunciation dictionary was established on our previous work [2,4]. Pronunciation model was integrated in this dictionary which included alternative pronunciations of each syllable. Results show that there is a 0.57% improvement in SER for test_set1 and a slight 0.28% improvement for test_set2. Note that pronunciation model technique shown here can only model complete changes but not partial changes. Second, the third row shows that direct use of auxiliary tree-based PCPMs into the decoder to model pronunciation variations is a failure (SER increased 1.14% absolute on test_set1). Such degradation in spontaneous speech is caused by non-robust acoustic models and lexicon confusion.

A comparison of the recognition performance of modeling partial changes with baseline and SLPM by Gaussian mixture sharing discussed in [6] is presented in the last three rows. The results presented in the last row shows that the use of reconstructed models yields a significant 2.39% improvement in SER absolutely on test_set1 compared with the baseline, and 1.73% in SER reduction with respect to using the enhanced HMMs. Even in testing on planned and clean speech (F0 condition, test_set2), which has fewer pronunciation variations, the reduction of SER is still remarkable (1.24% absolutely). Furthermore, use of the proposed method yields an additional 1.45% SER reduction in spontaneous speech compared with that of SLPM. The higher efficiency of pronunciation modeling in the proposed method lies in the fact that (1) PCPMs are independent models and can efficiently represent partial changes at the model level; (2) no model confusion is introduced during acoustic model reconstruction by auxiliary decision tree merging. We increased the resolution of the original models yet do not sacrifice their identity.

In Fig. 2 we give a more detailed comparison between using the baseline HMMs and reconstructed HMMs in decoding. Note that use of the baseline HMMs, the acoustic likelihood score near the 135th frame drops significantly because of an initial ‘n’ is mis-recognized as ‘l’ due to pronunciation variations. However, use of the modified HMMs can restore such local model mismatch and gives a correct recognition result. The reason lies in the fact that the modified HMM of ‘n’ utilizes the output densities of PCPM ‘n_l’, which covers the variations between ‘n’ and ‘l’. Thus, such actual pronunciation with variation can obtain a higher acoustic likelihood score compared with using baseline model. Therefore, ‘n’ may become the winner during decoding.

![Fig. 2 Modified HMMs restore the local model mismatch](image)

6. CONCLUSION

Partial change phone models were introduced to represent the partial variations in spontaneous Mandarin speech. We reconstructed the original baseform models by auxiliary decision tree merging through the reconstruction of Gaussian distributions. The uncertainty of the surface form caused by partial pronunciation changes was described by phone pairs which could be automatically generated from the baseform and surface form alignment. The experimental results have shown that modeling partial changes achieved higher recognition performance than only modeling complete changes. At the same time, we do not need to augment the lexicon and modify the decoder. Our method can be easily extended to other languages although it is applied in spontaneous Mandarin speech.

7. REFERENCE