TRIPHONE MODEL RECONSTRUCTION FOR MANDARIN PRONUNCIATION VARIATIONS

Pascale Fung, Liu Yi

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

ABSTRACT

The high error rate of recognition accuracy in spontaneous speech is due in part to the poor modeling of pronunciations. In this paper, we propose modeling pronunciation variations through triphone model reconstruction. We first generate partial change phone model (PCPM) to differentiate pronunciation variations. In order to improve the resolution of triphone models, PCPM is used as a hidden model and merged into the pre-trained acoustic model through model reconstruction. To avoid model confusion, auxiliary decision trees are established for triphone PCPMs. The acoustic model reconstruction on triphones is equivalent to decision tree merging. The effectiveness of this approach is evaluated on the 1997 Hub4NE Mandarin Broadcast News Corpus (1997 MBN) with different styles of speech. It gives a significant 2.39% absolute syllable error rate reduction in spontaneous speech.

1. INTRODUCTION

An analysis of pronunciation variations at the acoustic level reveals that pronunciation variations include both complete changes and partial changes [2,4,6]. Complete changes are the replacement of a canonical phoneme by another alternative phone, such as 'b' being pronounced as 'p'. Partial changes are variations within the phoneme and include nasalization, centralization, voiceless, voiced, etc. Most of the current work in pronunciation modeling attempts to represent pronunciation variations either by alternative phonetic representations or by the concatenation of subphone units at the state level [1,3,5]. This approach can only model complete changes but not partial changes. It has been shown that partial changes are very flexible and a lot less clear-cut than previously assumed and cannot be modeled by mere representation in alternate or concatenation of phone units [2,5].

Recently, state level pronunciation modeling (SLPM) has been proposed to model partial changes [5,6]. However, there are still challenges and the improvement coming from this approach is limited. Current SLPM scheme uses the same phoneme unit inventory to represent partial changes. This approach may introduce model confusion. For example, based on SLPM, the modified HMM for 'b' utilizes the output densities of 'd', while the modified HMM for 'd' may also utilize the output densities of 'b' due to their pronunciation variations. If the variation probability between 'b' and 'd', 'd' and 'b' is close, the confusion between the modified HMMs 'b' and 'd' is increased. In other words, although SLPM improves the resolution of the acoustic model, it may introduce more model confusion. Moreover, the improvement from SLPM is based on augmenting the Gaussian mixture number of HMMs, which inflates the number of parameters and costs more computation time for training and decoding.

In this paper, we propose modeling pronunciation variations through triphone model reconstruction. Partial change phone model (PCPM) is first generated to represent the acoustic realizations between the canonical and alternative pronunciations. Instead of phone models, PCPMs are established from the samples obtained through the alignment between the baseform and surface transcriptions. Furthermore, in order to improve the resolution of the reconstructed triphone models, PCPM is treated as a hidden model and merged with relevant pre-trained acoustic model through decision tree merging. One auxiliary decision tree only maps to one standard decision tree of the pre-trained triphone models. Hence, compared to SLPM, the model resolution is improved to capture pronunciation variations, while no model confusion is introduced. Only the parameter size of the reconstructed model has a small inflation. Compared with phone level pronunciation modeling methods [1,3], our acoustic level approach models pronunciation variations with a higher resolution.

The paper is organized as follows. Section 2 describes the motivation and the mechanism of generating PCPMs. Section 3 describes the method of acoustic model reconstruction through decision tree merging on triphones. In section 4, experimental results on the 1997 MBN corpus are described. Finally, we conclude in section 5.

2. PARTIAL CHANGE PHONE MODELS

2.1. Motivation of Using PCPMs

Spontaneous Mandarin speech includes both complete changes and partial changes. For example, Chinese initials are very flexible and around 30% of the variations are partial changes [4].

When partial changes occur, a phone is not completely substituted, deleted or inserted. Therefore, the transcriber agreement on spontaneous speech is much lower than that on read speech. An analysis of phone level transcriptions of a spontaneous Mandarin speech corpus – CASS corpus [4] shows that the average transcriber agreement is 84.23%. Whereas the agreement on a read speech corpus – 863 corpus, is around 98%. The different transcriber agreement rate suggests that when partial changes occur, the transcribers who are forced to use a categorical label from the limited phonetic inventory may end up choosing different labels for phone level representation. That is, given a continuous acoustic signal whose relevant pronunciation is different from the canonical pronunciation, it is very difficult even for phoneticians to clearly identify the exact pronunciation changes.

Pronunciation variations can be represented at different levels. If the variation is large enough and can be identified at the phone level as represented by another phone, then pronunciation modeling can be used at the phone level [1,3] to handle this variation. If the variation is small enough, then using more Gaussians within the acoustic model [5] can solve this problem. However, if the variation is at an intermediate level, the above approaches cannot differentiate and deal with this deviation. Therefore, a more powerful model is required to account for the ambiguity of acoustic representations caused by partial changes. The acoustic model for spontaneous speech should be different from that of read speech – it should have a strong ability to cover partial changes as well as complete changes.

2.2. Representation of PCPMs

In this section, we start from the recognition formulae and deduce the representations of PCPMs. In current ASR systems, the decoding formula is

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

where B is the baseform sequence in terms of phoneme representations, and X is the input speech vectors. If words are 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 variations into consideration. Suppose a word can be pronounced in several alternative ways, and assuming S is one possible sequence of a pronunciation, the surface form, in terms of phone representations, the decoder formula becomes

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

where $P(B)$ is the language model, $P(X \mid B,S)$ is the acoustic model, and $P(S \mid B)$ is the pronunciation model. In general, the acoustic model training procedures assumes that

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

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

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

Obviously, both $P(X \mid B)$ and $P(X \mid S)$ 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, both the baseform and surface form should be taken into account for acoustic model estimation. In Eq.2, $P(X \mid B,S)$ is called the *partial change phone model* (PCPM).

Compared with the conventional predefined phone symbol, PCPM is represented using the phoneme/phone pair which is automatically generated from the baseform and surface form alignment. If the phoneme in the baseform has a different phone representation in the surface form, this phoneme and phone will be combined to form a phoneme/phone pair for PCPM representation. PCPMs can be considered as an extended model set in regard to the conventional phone model set. For example, the baseform model is 'b' and its related PCPMs could be 'b_p', 'b_f' and 'b_d' due to different types of pronunciation variations.

2.3. Acoustics Represented by PCPMs

In order to investigate the characteristics of PCPMs, we analyze the acoustic features represented by PCPMs. Suppose 'b', 'd' and 'b_d' are the baseform, surface form and PCPM representation, respectively. The acoustic realization for 'b_d' is the acoustic samples which are labeled as 'b' in the baseform but transcribed as 'd' in the surface form. For 'b', it is the acoustic samples which are labeled 'b' both in the baseform and surface form. Similarly, acoustic realization for 'd' can be determined. We first calculate the global mean μ_b and μ_d for the acoustic samples of 'b' and 'd', then plot the acoustic realization of 'b_d' according to its relative distance to μ_b and μ_d in a normalized acoustic space. In this space, the mean distance between μ_b and μ_d is normalized to 1.

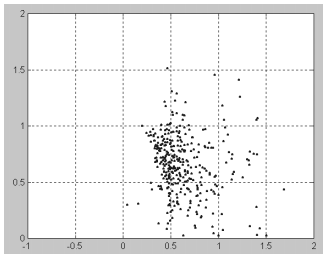

Fig.1: Acoustic realization of a PCPM 'b_d' in a normalized acoustic space

In Fig.1, the x-axis and y-axis are μ_b and μ_d, the coordinates of the points in this normalized acoustic space are the relative distance of acoustic realization of 'b_d' to μ_b and

μ_d . It has been shown that the points representing the acoustic realizations of PCPM 'b_d' fall mostly within the area between (0,0) and (1,1). This means that the acoustics of a phoneme, e.g., 'b' when realized as a phone, e.g., 'd', lies between the average realization of the phoneme and the average realization of the phone. Neither the phoneme nor the phone unit can accurately represent this type of variation. However, during the generation of PCPMs, both the baseform and surface form representations are taken into account. PCPMs, e.g., 'b_d' can be efficiently used to model the acoustic realizations at the intermediate level. In addition, using PCPM is easy to differentiate pronunciation variations. Acoustic samples belonging to 'b_d' are different from those of 'd_b' since the tendency of pronunciation variation from 'b' to 'd' is different from that of 'd' to 'b'. Now we will discuss how to use PCPMs.

3. ACOUSTIC MODEL RECONSTRUCTION

Previously, we have shown that increasing the phone model set to model pronunciation variations gives no significant improvement [2]. In this work, we propose using PCPM as a hidden model and merging them into the pre-trained baseform model to improve the model's resolution. This approach aims at making the pre-trained model acquire the ability from PCPMs to accommodate pronunciation variations.

3.1. Generating Auxiliary Decision Trees for PCPMs

Current ASR systems always use context-dependent triphone model. In order to limit the complexity of triphone models and avoid the sparse data problem in acoustic model training, decision tree based state clustering is commonly used [7]. In our system, triphones for PCPM are similar to conventional triphones except for the central phone. The former is a phone pair and the latter is a phoneme or phone. The trees for PCPMs are called *auxiliary decision trees*, while trees for standard triphone models are called *standard decision trees*. The structure of auxiliary decision trees is similar to that of standard decision trees. However, auxiliary decision trees are only used during the state-tying procedure for PCPM triphone models, while not used in the acoustic model estimation and decoding. This is because each leaf node of decision tree represents a tied-state, and after acoustic model reconstruction, auxiliary decision trees will be merged into standard decision trees and will not appear in the following steps.

3.2. Triphone Model Reconstruction through Decision Tree Merging

Acoustic model reconstruction of the triphone model is more complicated than that of the monophone model. The triphone variants of the same central phone have several alternatives, the relation between baseform triphone models and PCPMs is *many-to-many* as shown in Fig.2.

Since the leaf node of decision tree represents a tied-state triphone unit in tree-based state tying system, therefore, acoustic model reconstruction equals to tree merging between auxiliary decision trees and standard decision trees. The mapping nodes between auxiliary trees and the relevant standard tree can be determined according to the Minimum Gaussian Distance Measure between two tied states as described in [7].

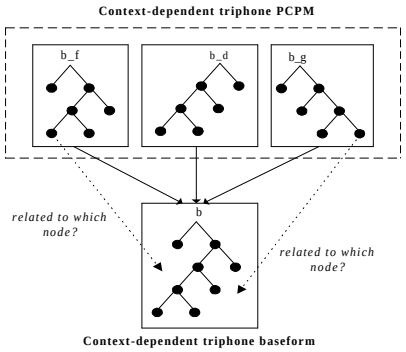

Fig.2: The relationship between tied-state triphones of PCPM and baseform model in acoustic model reconstruction

Determined by the minimum distance between tied states, leaf nodes of auxiliary decision trees are merged into the relevant nodes of standard decision trees as shown in Fig.3. According to this tree merging, the pre-trained baseform models are reconstructed and include Gaussian mixtures from its own as well as from the PCPMs to represent pronunciation variations. For example, in Fig.3, the leaf node, i.e., tied state 'ST_4_3' of the standard decision tree includes the nodes from different auxiliary decision trees in order to model different pronunciation changes, e.g., $b \rightarrow f$ and $b \rightarrow p$.

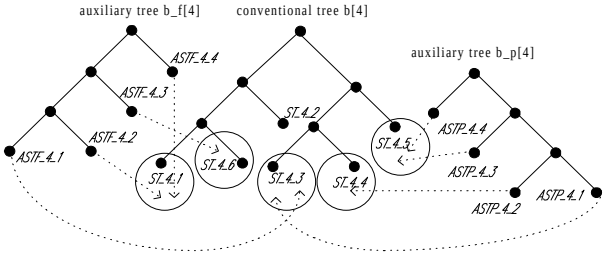

Fig.3: Auxiliary decision trees are merged into a standard decision tree

In this approach, different types of pronunciation changes are represented using different PCPMs and auxiliary decision trees. One auxiliary decision tree cannot be used by two different standard decision trees, so no model confusion is introduced. Using the Gaussians from PCPMs enables the reconstructed model to acquire the ability from PCPMs to model pronunciation variations. That is, without introducing the model confusion, the model resolution is improved.

4. RECOGNITION EXPERIMENTS

The acoustic training set consists of 10 hours of speech (10,483 utterances) selected from the first two CDs in the 1997 MBN corpus. The testing set consists of two parts: the first one (test_set1) includes 865 spontaneous utterances consisted of 11512 syllables in total. The second one (test_set2) is 1263 clean utterances (F0 condition) from the Hub4NE evaluation sets

[2,3], consists of 15535 syllables in total. HTK toolkit is used to train triphone models. The HMM topology is three-states, left-to-right without skips. The acoustic features are $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. The HTK flat-start procedure is used to build the 10 Gaussians model, state clustered HMMs with 2904 states.

415 toneless standard Chinese syllables are used in the experiments. 145 context-independent PCPMs are generated through DP alignment between the baseform and surface transcriptions in the training set. Using the decision tree based state-tying approach [7], 818 tied-states are generated for auxiliary decision trees of PCPM triphones. Through decision tree merging, the reconstructed acoustic model includes 37,220 ((2904+818)*10) Gaussians and each state has 12.8 Gaussians on average. Compared with the baseline model of 29040 Gaussians, this only gives a 28.2% increase in parameter size. Note that in SLPM discussed in [5], when two set of models are merged, the number of Gaussians is nearly doubled. In order to make a fair comparison, we generate an enhanced HMM which has 13 Gaussians per state. For SLPM system, each state has 13.1 Gaussians on average. The recognition performance is listed in Table 1.

	Syllable Error Rate (SER) %	
system	**Test_set1**	**Test_set2**
Baseline	42.23	30.92
Baseline HMMs & pronunciation dictionary	41.66	30.64
Enhanced HMMs	41.57	30.47
SLPM	41.29	30.05
Triphone model reconstruction using PCPMs	**39.84**	**29.68**

Table 1: Using triphone model reconstruction outperforms other pronunciation modeling approaches

In the second system described in Table.1, the pronunciation dictionary is established on our previous work [2,3]. It has been shown that only a very limited improvement is obtained by using multiple pronunciations. Note that pronunciation model technique shown here can only model complete changes but not partial changes. A comparison of the recognition performance of using triphone model reconstruction with baseline and SLPM by Gaussian mixture sharing discussed in [5] is presented in the last three rows. It shows that using the reconstructed models yields a significant 2.39% absolute improvement in SER on test_set1 compared with the baseline, and 1.73% with respect to using the enhanced HMMs. Furthermore, it gives an additional 1.45% absolute SER reduction in spontaneous speech compared with that of SLPM. The higher efficiency of pronunciation modeling through acoustic model reconstruction lies in the fact that (1) PCPMs can efficiently differentiate pronunciation changes at the model level; (2) no model confusion is introduced during acoustic model reconstruction by auxiliary decision tree merging.

The results in Table.1 shows that simply increasing the Gaussian numbers per state does not help much in terms of SER reduction, since some of the Gaussians are poorly estimated as the number of Gaussians increased. However, in our proposed method, the reconstructed model includes the Gaussians from PCPMs, which enable the "borrowed" Gaussians to cover the boundaries of the original probability distribution. More Gaussians in this region make it possible to model in detailed distributions. Fig.4 illustrates that the output distribution of the reconstructed model at the boundary, e.g., between -10 and -5, is more robust than that of the baseline model.

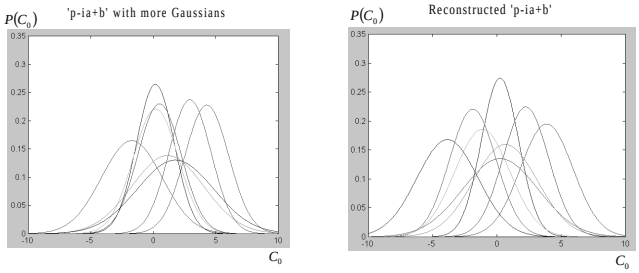

Fig.4: Reconstructed triphone model covers a wide Gaussian distribution compared with simply increasing Gaussian numbers

5. CONCLUSION

We have described an approach of triphone model reconstruction for modeling pronunciation variations. In order to improve the resolution of the reconstructed model, we propose PCPMs to differentiate pronunciation changes and merge them into the pre-trained baseform model. In addition, we generate auxiliary decision trees for triphone PCPMs, and use decision tree merging to perform acoustic model reconstruction. One auxiliary decision tree can only be used by one standard decision tree during model reconstruction, so no model confusion is introduced. It has been shown that this new pronunciation modeling approach provides a significant 2.39% absolute SER reduction for spontaneous speech. Our method is applied to spontaneous Mandarin speech but can be easily extended to other languages.

6. REFERENCE

[1] M. Finke, et.al., "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, Aug.2000

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

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

[5] M. Saraclar, et al., "Pronunciation modeling by sharing Gaussian densities across phonetic models", *Computer Speech and Language*, (2000) 14, 137-160

[6] M. Saraclar, et al., "Pronunciation ambiguity vs pronunciation variability in speech recognition", *Proc. ICASSP00*, 2000

[7] S.Young, et.al., The HTK book. Entropic Cambridge Research Laboratory, 1999