Estimating Pronunciation Variations from Acoustic Likelihood Score for HMM Reconstruction

LIU Yi and 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

It is widely acknowledged that pronunciation modeling is an efficient way to improve recognition performance in spontaneous speech. In pronunciation modeling, almost all methods of generating variation probability are based on relative frequency counting from DP alignment. In this paper, we investigate the local model mismatching caused by pronunciation variations and propose to estimate variation probability from acoustic likelihood score. According to estimated probability, we present a method of reconstructing pre-trained HMM models to include alternate pronunciations by sharing optimal mixture components instead of distributions. Experimental results show that using reconstructed HMM set reduces syllable error rate by 2.03% absolutely compared to the baseline system, also the accuracy improvement gained from proposed method is almost double with respect to that from previous DP alignment.

1. Introduction

There are many attempts to incorporate pronunciation variations into the lexicon, the decoder or the training labels [5,6,7]. Among various approaches, augmenting lexicon with alternative pronunciations can lead to more recognition confusion and a larger lexicon requires more computation time for recognition. Training such a lexicon also requires a huge amount of data. Incorporating pronunciation variation into the decoder requires modification of the search algorithm [5]. In [1], it is suggested that sharing Gaussian mixtures between canonical and surface form states is beneficial. However, in this method, when “merged” two sets of acoustic models, it will copy all Gaussian mixture densities from surface state to base state, which will dramatically increase the mixture component numbers and leads to more training and decoding time. On the other hand, pronunciation variation probability is usually obtained from counting number of times in which state in base form transcriptions is aligned to a state in surface form transcription [1]. Previous variation probability estimating method is based on related phoneme pair frequency counting. If we get state-to-state alignment from a phoneme-to-phone alignment, it requires all the HMMs in the model set have the same number of states, and phone-to-phoneme alignment should rely on a knowledge based phone feature distance matrix, and deletion errors between phone and phoneme sequence is hard to handle.

In this paper, we propose a different method to estimate pronunciation variation probability from acoustic modeling mismatch likelihood score, and use it for restructuring Gaussian mixture density functions (pdfs) of pre-trained original HMMs. After HMM reconstruction, several acoustic phenomena and pronunciation variations can be modeled properly, the number of mixture components of reconstructed HMMs is just a little inflated by sharing new components from mixture pdfs of alternate realization states.

Our approach takes advantage of acoustic feathers for estimating variation probability, does not require all the HMMs in the model set have the same number of states, which should be flexible for HMM topology design. Since we consider frame-level error between baseform phoneme state and surface form phone state, the deletion error will not be involved. Further more, we includes alternate pronunciations in state mixture component level by HMM reconstruction, so that it avoids the problem of lexicon augmentation, and does not require any modification of the decoder and training labels. Finally, since we only share mixture components instead of mixture densities for HMM reconstruction, so the Gaussian mixture numbers will not dramatically be increased when we “merged” different alternate states.

The paper is organized as follows: section 2 introduces generating pronunciation variations from local acoustic modeling mismatch. In section 3, we describe how to reconstruct original HMM sets by variant probability. Finally speech recognition experiments which accommodate this approach are presented in section 4. We conclude in section 5.

2. Generating Pronunciation Variations

2.1 Overview of Previous Methods

Pronunciation variation probability is the basic factor for pronunciation modeling. Generally speaking, previous method for generating variant probability is showed as:

1. Starting with a canonical transcription of the acoustic training corpus and its surface form transcription.
2. Align baseform sequence with surface form sequence using dynamic programming (DP) alignment, the unit could be phone, syllable, state, etc.
3. Count the relative mapping frequency between phonemes and phones (or other units).
4. Filter rare events by using threshold. The threshold can be set as absolute counts or frequency probability.
5. Estimate variation probabilities with the frequency count, \( b_j \) is base form unit and \( s_i \) is surface form unit.

\[
p(s_i | b_j) = \sum_j P(s_i, b_j) / P(b_j)
\]

(1)
From above steps, it is obvious that the probability estimation heavily depends on the relative phone-pair occurrence. Here we focus on state level modeling, in order to convert the phoneme-to-phone alignment to a state-to-state alignment, we should assume all the HMMs must have the same number of states. However, in practice we always assign different state numbers to different HMMs according to their characters. Thus, when HMMs have different state number, it is very hard for us to estimate variation probability from above method. On the other hand, traditional alignment method only relies on a knowledge based phone feature distance matrix. The distance matrix is generated by human knowledge and related to a special language, for different languages, we need to describe different cost matrix. Another weakness of this approach is that in step 2, only substitution between base form and surface form sequences can be considered, deletion and insertion will be omitted because we cannot find its mapping phone or phoneme. So in the next section, we will show our new approach to estimate variation probability from frame-level error likelihood score, it will overcome the mentioned problems and easily to do state to state alignment, also it is convenient to expand to phoneme to phone and syllable level alignment.

2.2 Basic Ideas

The key technique of variation probability estimation from acoustic likelihood score is to take into account of frame-level errors caused by local acoustic modeling mismatch.

Let $B = b_1, b_2, \ldots, b_T$ be the time sequence of baseform state, and $S = s_1, s_2, \ldots, s_T$ the time sequence of surface form state, and $X = x_1, x_2, \ldots, x_T$ the input vector. Moreover, let $N$ denote the total number of HMM units and $M = \{m_1, m_2, \ldots, m_N\}$ is all pre-trained HMM set.

First, forced Viterbi alignment algorithm is used for generating baseform time state sequence. Here, a standard pronunciation dictionary [3] is used, the dictionary has an average of 1.6 pronunciations per syllable. When the syllable has more than one pronunciation, the recognizer considers all pronunciations for each syllable and outputs the pronunciation that best matches the acoustic data. During alignment, we keep track of full state, and at each frame $t$, output acoustic likelihood $L_{\text{base}}(b_t) = P(x_t \mid b_t)$ is saved for further usage.

The time sequence of surface state $S$ can be obtained by phone recognition. Since we only concern about acoustic likelihood, so during recognition we do not use phone ngrams. Also, in order to keep the consistency as baseform time state generation, we use same dictionary and free loop grammar in decoding. Same as before, given each HMM state at frame $t$, the output acoustic likelihood based on the acoustic vectors is $L_{\text{surface}}(s_t) = P(x_t \mid s_t)$, it will be kept for calculating frame level errors.

At frame $t$, $b_t$ and $s_t$ are frame-by-frame time state sequences. If $b_t \neq s_t$, which means:

$$b_t = \arg \max_{m_t \in M} \{ P(x_t \mid m_t) \}$$

(2)

Frame-level error occurs, it can be denoted as

$$E = \left| L_{\text{base}}(b_t) - L_{\text{surface}}(s_t) \right|$$

(3)

$L$ denotes acoustic log likelihood. If frame-level errors of $(b_t, s_t)$ state pair occur frequently, which means Gaussian mixture pdfs of $b_t$ and $s_t$ have similar acoustic features, their models may be mismatched by pronunciation variations. The smaller the $E$, the more similar $(b_t, s_t)$ pair.

Further more, At frame $t$, if $b_t \neq s_t$, when frame-level errors occur, according to input vector $x_t$, the acoustic likelihood of each mixture component in surface state $s_t$ is different. We also can say that with respect to baseform state $b_t$, the similarity of each mixture component in $s_t$ is different. Based on input vector, when variation occurs, only mixture component giving the highest acoustic likelihood has the highest contribution. So in order to catch alternate pronunciations in state level, we do not need to copy all mixture components or share mixture distributions of $s_t$ to $b_t$ for HMMs reconstruction. Instead, only the most likely mixture component of $s_t$ will be considered and shared among pre-trained HMMs. The most likely Gaussian component of $s_t$ can be obtained by

$$g_{\text{opt}} = \arg \max_{s_t} \{ p(x_t \mid g_{s_t}) \}$$

(4)

In equ.4, $g_{\text{opt}}$ is the optimal mixture component selected from surface state $s_t$ according to input vector $x_t$, and $g_{s_t}$ denotes total mixture components of $s_t$.

Table 1 gives the examples. The first column is input vector sequence number and the last column is the selecting optimal mixture component of $s_t$.

<table>
<thead>
<tr>
<th>$s_t$</th>
<th>$b_t$</th>
<th>$E$</th>
<th>$g_{\text{opt}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>ing[2]</td>
<td>1.914248</td>
<td>5</td>
</tr>
<tr>
<td>301</td>
<td>ing[3]</td>
<td>0.435846</td>
<td>1</td>
</tr>
<tr>
<td>302</td>
<td>ing[4]</td>
<td>1.181101</td>
<td>1</td>
</tr>
<tr>
<td>308</td>
<td>eng[4]</td>
<td>1.530333</td>
<td>3</td>
</tr>
<tr>
<td>325</td>
<td>xi[2]</td>
<td>1.482751</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1 variation information at frame level

In order to alleviate the dependency of $g_{\text{opt}}$ on the input acoustic data, in the following sections, we show that both frame-level errors and $g_{\text{opt}}$ occurrence numbers are considered. We also set threshold to cut those mappings with rare occurrence $g_{\text{opt}}$ and high frame-level error pairs, which maybe caused by accidental frame level errors or noise.

2.3 Calculate Pronunciation Variation Probability

Define $P(E, s_t \mid b_t)$ is conditional probability of frame-level variation pair $(b_t, s_t)$ occurrence.

The variation probability can be calculated as follows:
For each input vector $x_T = x_1, x_2, \ldots, x_T$. Matching $b_T = b_1, b_2, \ldots, b_T$ with $s_T = s_1, s_2, \ldots, s_T$, according to input frame number sequence.

(2) According to equ.3, calculate $P(E, s, \mid b)$ for all combinations of $b_1$ and $s_1$.

(3) Do (1) and (2) for all training data.

(4) If $P(E, s, \mid b) < \text{threshold}$ keep the variation pair $(b_1, s_1)$. Threshold is the permitted maximum likelihood score of state pair $(b_1, s_1)$, also it is expected to suppress the accidental frame level errors, which may be caused by noise.

(5) For the filtered variation pair $(b_1, s_1)$, use equ.4 to find the optimal mixture component $g_{\text{opt}}$ from $s_1$.

(6) Filter rare events of $(b_1, s_1)$ and $g_{\text{opt}}$ using empirical frequency (threshold = 0.1). The variation probability is estimated from equ.5. Here, $b_1$ and $s_1$ is state, and $g_{\text{opt}}$ is optimal mixture component of $s_1$.

$$P(s_1, g_{\text{opt}} \mid b_1) = \frac{\text{Occur} (s_1, g_{\text{opt}}, b_1)}{\text{Occur} (b_1)}$$

(7) End.

From above steps, not only variant in state level but also in mixture component level is considered. Moreover, since the variation probability is estimated from frame number sequence, there will not be insertion and deletion problems and no requirement for same state number to each HMMs.

### 3. HMM Model Reconstruction

#### 3.1 New Output Distribution Generation

The idea of reconstructing HMM model is to share Gaussian mixture components not mixture pdfs of alternate states to Gaussian mixture density of canonical states. The reconstructed acoustic model can include both canonical and alternate realizations in one set of HMM states, and the mixture weights are governed by pronunciation variation probabilities.

Our method focus on continuous density HMMs. First, we assume $P(x \mid b)$ is the output distribution of state $b$ in the original pre-trained HMMs and $P(s \mid b)$ is variation probability.

$$P(x \mid b) = \sum w_{j,b} f_j (\nu, \mu_j, \Sigma_j)$$

where $w_{j,b}$ is mixture weight of the $j$th mixture component. In the next follows, we use $f_j (\cdot)$ to denote $f_j (\nu, \mu_j, \Sigma_j)$.

We define $P'(x \mid b)$ is new output distribution of state $b$ in reconstructed HMM acoustic models, that is:

$$P'(x \mid b) = \lambda P(x \mid b) + (1 - \lambda)P(x \mid s) \cdot P(s \mid b)$$

Since the shared mixture components are from different alternate states $s_j$, so we have:

$$P'(x \mid b) = \lambda P(x \mid b) + (1 - \lambda) \sum w_{j,b} P(x \mid s_j) \cdot P(s_j \mid b)$$

When we consider the optimal mixture components instead of distributions of surface state for model reconstruction, equ.9 is written as:

$$P'(x \mid b) = \lambda \sum w_{j,b} P(x \mid s_j) + (1 - \lambda) \sum w_{i,b} \cdot P(s_i \mid b)$$

where $w_{j,b}$ and $w_{i,b}$ are new mixture weights of state $b$ in reconstructed HMM model, and $\lambda$ is normalized rate, which guarantees the sum of new mixture weights equals to 1. From equ.10, we can see that after HMM model reconstruction, the output distribution of each state includes both canonical and alternate realizations. In addition, the weights of shared mixture components are governed by pronunciation variation probability.

#### 3.2 Acoustic Model Re-estimation

The parameters of reconstructed HMMs will be re-estimated. We will use traditional Baum-Welch algorithm, all the configurations for re-training are exactly same as those used in pre-trained procedures. So that the pure effect given by model reconstruction from pronunciation modeling can be evaluated. Mixture weights of reconstructed HMMs are initialized by equ.11 and re-estimated as free variable in the following iterations. Two methods of acoustic model re-estimation are used: only re-train the mixture weights and transition probabilities, re-train all parameters.

### 4. Speech Recognition Experiments

We use Hub4NE 1997 Mandarin Broadcast News Database to evaluate the effectiveness of our approach. The total number of CD-Initial and CI-Final is 139, and 415 for Chinese standard syllables. Three-states, left-to-right HMM topology is applied. The acoustic features are 13 MFCC, 13ΔMFCC and 13ΔΔMFCC. The data for acoustic model training and HMM reconstruction consists of 10 hours of speech, the testing data is 724 utterances. Most of training and testing data are spontaneous speech and conversational speech.

We trained 4, 8, 12 and 16 Gaussian mixtures HMM models. 4 and 8 mixture HMM models are used for acoustic model reconstruction, and 12, 16 Gaussian mixture HMMs are used for recognition performance evaluation. In the following tables, "−" means only re-estimated weights and transition probabilities of reconstructed model, and "−w−" means re-estimated all parameters.

In table 2, we compared the recognition performance of different variation probability estimation methods. The previous DP alignment and proposed method are used to train state level variation probability respectively. In order to make a fair comparison, during HMM model reconstruction, we
copied mixture pdfs from alternate states to their canonical state. Starting from 4 Gaussian mixture per state HMMs and selecting top 3 variation probabilities, according to eq.9 we reconstructed new HMMs with 12 mixtures per state. Table 2 represent that the improvement gained from proposed method is almost double with respect to that from previous DP alignment. It proved that variation probability estimated from acoustic scores at frame level is more accurate.

<table>
<thead>
<tr>
<th></th>
<th>%Syll Acc. And Improvement Relative to Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>12Gau Mix. (Baseline)</td>
<td>66.18 (0.00)</td>
</tr>
<tr>
<td>Proposed*</td>
<td>67.63 (+1.45)</td>
</tr>
<tr>
<td>Proposed**</td>
<td>68.16 (+1.96)</td>
</tr>
<tr>
<td>Previous DP Align*</td>
<td>66.84 (+0.66)</td>
</tr>
<tr>
<td>Previous DP Align**</td>
<td>67.20 (+1.02)</td>
</tr>
</tbody>
</table>

Table 2 Results of different probability estimation methods

We will compare the recognition performance of reconstructed HMMs in the following tables. Starting from 8 mixture HMMs system, from eq.10, the original 3336 mixture components inflate to 4670, and on average there are 11.2 Gaussian mixtures per state. The reconstructed HMMs will be re-estimated as shown in section 3.2. To make a fair comparison, we compared the reconstructed HMMs with 12 and 16 Gaussian mixtures systems, the decoder and lexicon are exactly same.

Table 3 and table 4 give the phone and syllable recognition performance.

<table>
<thead>
<tr>
<th></th>
<th>%Ph. Acc. And Improvement Relative to Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>12Gau Mix. (Baseline)</td>
<td>66.10 (0.00)</td>
</tr>
<tr>
<td>Reconstructed*</td>
<td>67.00 (+0.84)</td>
</tr>
<tr>
<td>Reconstructed**</td>
<td>67.10 (+0.94)</td>
</tr>
<tr>
<td>16Gau Mix.</td>
<td>67.11</td>
</tr>
</tbody>
</table>

Table 3 Performance of phone recognition results

<table>
<thead>
<tr>
<th></th>
<th>%Syll Acc. And Improvement Relative to Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>12Gau Mix. (Baseline)</td>
<td>66.18 (0.00)</td>
</tr>
<tr>
<td>Reconstructed*</td>
<td>67.83 (+1.65)</td>
</tr>
<tr>
<td>Reconstructed**</td>
<td>68.21 (+2.03)</td>
</tr>
<tr>
<td>16Gau Mix.</td>
<td>67.75</td>
</tr>
</tbody>
</table>

Table 4 Performance of syllable recognition results

In table 5, during decoding for baseline systems (12 or 16 mixtures), we use multi- pronunciation lexicon [3]. This decoding lexicon has 2.4 pronunciations per syllable on average, and each pronunciation attached related probability. However, for reconstructed HMM models, we still use one-to-one mapping lexicon.

<table>
<thead>
<tr>
<th></th>
<th>%Syll Acc. And Improvement Relative to Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>12Gau Mix. (baseline)</td>
<td>67.00 (0.00)</td>
</tr>
<tr>
<td>Reconstructed*</td>
<td>67.83 (+0.83)</td>
</tr>
<tr>
<td>Reconstructed**</td>
<td>68.21 (+1.21)</td>
</tr>
<tr>
<td>16Gau Mix.</td>
<td>68.12</td>
</tr>
</tbody>
</table>

Table 5 Performance of using multi-pronunciation lexicon

Table 3 and table 4 showed that after reconstructing HMM, both phone accuracy and syllable accuracy are all improved, especially for syllable accuracy, which improved 2.03% absolutely with respect to the baseline. Also, the syllable accuracy is even higher than using 16 Gaussian mixtures although the reconstructed HMM only has 11.2 mixtures per state. Results in table 5 represent an important advantage of proposed method. The dictionary need not be expanded to include alternate pronunciations, since we have included mixture components of alternate realizations after HMM reconstruction, so that even use one-to-one standard lexicon, the syllable accuracy is higher than 12 mixtures baseline system and still can compare to 16 Gaussian mixtures with multi-pronunciation lexicon.

5. Conclusion

We present new approaches to estimate pronunciation variation probability from acoustic likelihood score at frame level and reconstruct HMM model by sharing dominant mixture components. The experimental results show that proposed method for variation probability estimation is more accurate than previous DP alignment and efficient in improving recognition performance. According to estimated variation probability, we can include alternate realizations by sharing mixture components instead of mixture pdfs from surface states to canonical states. Further more, we do not need to modify lexicon and decoder and we do not need a huge amount of data for variation probability training. Our future work includes discovering more efficient criterion for threshold setting and using a hand transcribed data for bootstrapping.

6. Acknowledgements

We would like to thank Dr. Bill Byrne, Johns Hopkins University for his valuable and instructive suggestions in this paper. This work is supported by research grants from Hong Kong Research Grant Council.

7. References