Probabilistic Latent Speaker Analysis for Large Vocabulary Speech Recognition

Dan Su, Xihong Wu and Huisheng Chi
Speech and Hearing Research Center
State Key Laboratory on Machine Perception, Peking University
Beijing 100871, China
sudan@cis.pku.edu.cn, wxh@cis.pku.edu.cn

Abstract
Trajectory folding problem is intrinsic for HMM-based speech recognition systems in which each state is modeled by a mixture of Gaussian components. In this paper, a probabilistic latent semantic analysis (PLSA)-based approach is proposed for use in speech recognition systems to alleviate this problem. The basic idea is that different speech trajectories are strongly correlated with speaker variation, and different speakers may have high scores on certain Gaussian components consistently. Thus, PLSA is adopted to perform co-occurrence analysis between Gaussian components and speakers and provide additional source of information to constrain searching path during decoding procedure. Experimental results show that 11.2% and 2.7% relative reduction on word error rate can be achieved on a homogeneous test set and the 2004 863 evaluation set, respectively.

Index Terms: speech recognition, PLSA, trajectory folding phenomenon, speaker variation

1. Introduction
Trajectory folding phenomenon [1] has been paid attention to for a long time since hidden Markov models (HMMs) [2] became the dominant methodology for automatic speech recognition. In traditional HMM framework, each state is modelled by a Gaussian mixture. The Gaussian mixture accounts for some variability of observation vectors within a state, whereas due to the independence assumption of acoustic observation vectors, the variability of speech trajectory can not be described adequately.

Figure 1 illustrates the so called Trajectory folding phenomenon. For a speech unit $u$, a two state ($s_1$, $s_2$) left-to-right HMM with two Gaussian pdfs per state has been trained with observations from two distinct speakers (speaker $a$, speaker $b$). It is reasonable to assume that, in each state, one Gaussian component models the speaker $a$’s and the other speaker $b$’s voice.

Figure 1: An example to illustrate the trajectory folding phenomenon.

For speaker $a$’s training data the model yields high probability for the path $s_1(a)$-$s_2(a)$, and for speaker $b$ for the path $s_1(b)$-$s_2(b)$. However, when presented with previously unseen input, the model may yield a high probability for path $s_1(a)$-$s_2(b)$, which has never been observed in the training set of $u$ and that might correspond to another sound $u'$. As a consequence, $u'$ could be misrecognized as $u$, which resulting in a performance degradation.

Many approaches have been proposed to overcome the adverse effect of the trajectory folding problem. In [2, 3], mixture of stochastic trajectory modelling was proposed. In [4, 5], multiple-HMMs modelling was employed. All experiments showed that multi-path modelling can improve speech recognition performance compared with single-path modelling method.

However, there are still some problems that are not completely resolved. First, both mixture stochastic trajectory modelling and multi-path modelling can only represent several limited speech trajectory paths, typically two or three, when the number of trajectory component increases, the performance will degrade due to data sparsity. Second, the length of trajectory modelling is also limited, in [2, 3], mixture stochastic trajectory model was built at phoneme level, in [4, 5], multi-path HMM model was experimented at syllable level, while the complicated coarticulation may even take effect at a longer-length level. When the longer-length level trajectory is modelled, the data sparsity problem arise again.

Figure 2 illustrates that the trajectory folding phenomenon may also exist at longer-length level. For example, a speech segment includes two units (unit1, unit2), the topology for each unit is the same as in Fig.1, the actual decoding result may contain such a trajectory that in unit1 it follows the speaker $a$’s path
while in unit2 it passes through the speaker b’s path. This can also bring performance degradation due to the trajectory inconsistency between the two adjacent units.

In this paper, a new way is presented to alleviate the trajectory folding phenomenon. And its basic idea is to introduce additional source of information characterizing the dependencies between Gaussian components. Many research works suggested that different trajectory correlates strongly with speaker variation, in [4], it was shown that trajectory clustering always detects the gender distinction as first factor.

Based on the above analysis, we first define the “dominant component” and assume that speech trajectory can be approximately represented by the sequence of dominant component. By adopting PLSA paradigm, the co-occurrence analysis between speakers and Gaussian components is performed, named probabilistic latent speaker analysis. With this method, probability of a dominant component can be measured given the dominant component history, thus, and in decoding procedure, this probability can be effectively integrated to constrain the searching path.

The rest part of this paper is organized as follows. In section 2, a detailed description of the proposed method will be given, the experiments and experimental results are introduced in session 3. And the conclusions and discussions are drawn in the last section.

2. The proposed approach

The proposed approach can be divided into training phase and recognition phase. In the training phase, state-level alignment is first done with each speaker’s utterances and then PLSA is performed. In the decoding phase, we use a linear interpolation method to integrate the probabilistic latent speaker score into searching procedure.

2.1. State-level alignment

First, the term “dominant component” needs to be explained, given a frame of feature vector and its labelled state, the index k of the dominant component can be obtained as:

$$\text{arg max}_k w_k N(x; \mu_k, \Sigma_k)$$  

(1)

In another word, the dominant component is the component with the highest probability given the feature vector.

In PLSA method, a co-occurrence matrix of words and documents needs to be first obtained. Here, in our method, we propose to build a co-occurrence matrix between Gaussian components and speakers, as shown in figure 3. In an acoustic model with i states and each state modeled with j Gaussians, the total number of Gaussian components is $M = i \times j$. Here, the M rows correspond with all M Gaussian components in the acoustic model and the N columns correspond with all N speakers in the training set.

To construct this matrix, firstly, state level alignment is performed on training set with each speaker respectively. During the alignment process, the indices of the dominant Gaussian components for each state are recorded. Then the number of times of each Gaussian component that appears as dominant component in its mixtures for each specific speaker can be obtained. Each element in the matrix represents the frequency of a Gaussian component becomes dominant component within a specific speaker’s utterances. Figure 3 illustrates the matrix used in our method.

2.2. PLSA analysis

After the co-occurrence matrix is constructed, PLSA is used to perform decomposition on this matrix and construct a compact latent speaker and component space. The observed speech trajectory is approximately represented by the sequence of dominant components. Then PLSA technique is used here to explore the relation between different trajectories and speakers.

A reader might argue, that PLSA only performs a co-occurrence analysis, it lacks the ability to model trajectory because it has neglected the temporal sequential order, while since here each Gaussian component comes from a determinate state and the states have a left-to-right order, the temporal sequential information is implicitly included and thus it gets a certain ability to model speech trajectory.

2.3. Decoding with probability latent speaker score

PLSA analysis provides the probabilities as

- $P(z)$
- $P(w|z)$
- $P(d|z)$

in which z represents the aspect, w represents a word and d represent a document, in our method, we rewrite them as

- $P(z)$
- $P(g|z)$
- $P(s|z)$

in which g represents a Gaussian component, s represents a speaker and z can be interpreted as a latent speaker cluster.

Given the model produced by PLSA, a measure of the distance between each two speakers in the probabilistic space could provide us with the information of whether the model has captured some information of speaker variation. The probability can be calculated using the follow equation

$$P(s_1, s_2) = \sum_z P(s_1|z)P(z)P(s_2|z)$$  

(2)

Just in the way PLSA is integrated with traditional language model [6], here, the probability of a dominant component given the dominant component history needs to be calculated.

Given the dominant component history and current dominant component, the distribution of latent speaker cluster can
be adapted using the history. For the first dominant component to be evaluated, the distribution of the latent speaker cluster defaults to the distribution in the training data

\[ P(z_k|h_i) = P(z_k) \] (3)

For the other dominant components, the distribution is adapted with

\[ P(z_k|h_i) = \frac{1}{i+1} \sum_{q=1}^{K} P(g_i|z_k) P(z_q|h_i) \] (4)

Then the probability of the current dominant component given the history can be calculated as

\[ P(g_i|h_i) = \sum_{k=1}^{K} P(g_i|z_k) P(z_k|h_i) \] (5)

During decoding procedure, this probability provides additional information to constrain the searching path consistent with a latent speaker cluster, thereby the trajectory folding phenomenon can be alleviated. To integrate this probability, we use a linear interpolation method, as in equation (5) in log domain.

\[ \log p = \lambda \log p(x|s) + (1 - \lambda) \log p(g|h) \] (6)

The interpolation weight can be optimized on the development set.

3. Experiments and results

3.1. Speech material

In our experiments, the speech material is taken from 863 continuous Mandarin speech corpus. This corpus totally consists of 166 speakers (83 male speakers and 83 female speakers). Each speaker has more than 500 utterances (about more than 40 minutes), which are all read speech recorded in clean studio environment. To evaluate our approach, we separated the speech material into the training, development and test sub-sets. Details of the composition of the three sets are given in Table 1.

Table 1: Main statistics of the speech data used in our experiment.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Training</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers</td>
<td>166</td>
<td>166</td>
<td>166</td>
</tr>
<tr>
<td>Utterances</td>
<td>85490</td>
<td>332</td>
<td>498</td>
</tr>
<tr>
<td>Hours</td>
<td>113.1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Considering that the environmental background and speakers in test data are quite matching with the training condition, a second test is designed to perform on the evaluation data of the 2004 863 Evaluation on Mandarin speech recognition (we will call this evaluation data as the 2004 863 evaluation set in the rest part of this paper). Speech utterances in this data set were recorded in various type of environmental noises. Speakers in this data set speak even with different accents. Details of the development set and test set are given in Table 2.

Feature extraction was carried out at a frame rate of 10ms using a 25ms Hamming window. A pre-emphasis factor of 0.97 was employed. 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first and second order time derivatives were calculated. Channel normalization was applied using cepstral mean normalization over complete recordings.

3.2. Training phase

The baseline model is a context dependent phoneme based acoustic model in which each unit is modelled by 3-state left-to-right HMM. After decision tree-based state tying, the baseline acoustic model totally consists of 3000 tied states and each HMM state is modelled by 32 Gaussian mixtures.

State-level alignment was performed on training set with each speaker respectively. During alignment process, the number of times for each of the 3000*32 Gaussian components appears as dominant component was accumulated within each speaker’s utterances. Thus, the co-occurrence matrix could be constructed, consisted of 3000*32 rows and 166 columns.

PLSA was then performed on this matrix. In our experiment, we set z=20. In an attempt to see whether PLSA has captured some information of speaker variation through co-occurrence analysis between speakers and Gaussian components, we calculated the joint probabilities of each two speakers according to equation (2), and plot these probabilities onto a 2 dimension plane, as in figure 4.

With the speaker indices from 1 to 166, the first 1 to 83 are male speakers and the last 84 to 166 are female speakers. Figure 4 demonstrates that the probabilities differ obviously between male and female speakers, this result indicates that the resulted model of PLSA has encapsulated some speaker variation information.

3.3. Recognition phase

We have revised the Sphinx 3.2 decoder to integrate the probabilistic latent speaker model into decoding procedure. When a frame of feature vector was evaluated by a Gaussian mix-
ture of a state, the index of the dominant component was preserved. The probability of this newly estimate dominant component given the history was linearly interpolated with the output probability as in equation (6).

In large vocabulary speech recognition systems, the Viterbi decoding procedure could produce an exhaustive amount of state-level path histories, thus, the requirement of disk space may increase greatly. Due to efficiency consideration, the dominant component history only reserved within a lexical range. It is efficiently implemented since the Sphinx decoder is based on a lexical tree framework.

### 3.4. Experimental results

For the first experiment, the interpolation weight $\lambda$ in equation (6) was first optimized on the development set, then, it was fixed during recognition procedure on the test set. Experimental result is given in Table 3.

<table>
<thead>
<tr>
<th>Recognizer</th>
<th>Word Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>9.69</td>
</tr>
<tr>
<td>With probabilistic latent speaker model</td>
<td>8.60</td>
</tr>
</tbody>
</table>

As can be seen in Table 3, the probabilistic latent speaker model has brought a significant improvement on recognition performance, a 11.2% relative reduction on word error rate can be observed.

Then the second experiment was performed on the 2004 863 evaluation set. The development set was used to determine the optimum interpolation weight.

<table>
<thead>
<tr>
<th>Recognizer</th>
<th>Word Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>53.00</td>
</tr>
<tr>
<td>With probabilistic latent speaker model</td>
<td>51.57</td>
</tr>
</tbody>
</table>

As can be seen in Table 4, due to the difficulty of this second task, the recognition performance was much lower than the first one. A 2.7% relative WER reduction was achieved with probabilistic latent speaker model.

### 4. Conclusions and discussions

In this article, motivated by trajectory folding phenomenon and inspired from probabilistic latent semantic analysis, we propose a new approach to alleviate this problem. Here, we consider speaker variation as a critical factor affecting speech trajectories, and approximately represent speech trajectories with dominant component sequences. By performing co-occurrence analysis between Gaussian components and speakers using PLSA. The resulted model which captures the relation between speech trajectories and speakers, named probabilistic latent speaker model, are integrated into the first-pass decoding to constrain the searching path during decoding procedure. Both experiments on 863 continuous Mandarin Corpus and the 2004 863 evaluation set show that the proposed approach has brought obvious improvement on recognition performance, 11.2% and 2.7% relative WER reduction respectively.

The recognition results suggest that the co-occurrence analysis performed by PLSA do capture some relation between Gaussian components and speakers. Since in our experiment, the dominant component sequence is used to approximately represent speech trajectory, the experimental results have justified our previous assumption that speaker variation is one of the most critical factors affecting speech trajectories. A favorable aspect for our experiment is that in our training set, each speaker has sufficient training data.

Although the 2004 863 evaluation set is quite unmatched with the training set, about 3% relative reduction on WER was also achieved. This improvement is not so significant as on the first test set. It can be explained by that PLSA usually suffers from overfitting problem. This has limited the effect of the proposed approach when test speakers are quite discrepant with speakers in training data. Latent Dirichlet Allocation (LDA) is a new method which improves upon PLSA by placing a Dirichlet prior on topic distributions to reduce overfitting problem. Therefore, in future work, PLSA may be replaced by LDA to get better performance.

The probabilistic latent speaker model could be effectively integrated into decoding procedure, so far, it has been achieved at lexical level, the effect of integration at much longer level need to be experimented in future. Furthermore, since the proposed method was performed in first-pass decoding procedure, it will provide better transcription for the subsequent speaker adaptation methods.

### 5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (60305004; 60435010; 60535030; 60605016), the National High Technology Research and Development Program of China (2006AA01Z196, 2006AA010103), and the Trans-Century Training Programme Foundation for the Talents by the State Education Commission.

### 6. References


