Probabilistic Latent Speaker Training for Large Vocabulary Speech Recognition

Dan Su, Xihong Wu, and Huisheng Chi

Speech and Hearing Research Center
Key Laboratory of Machine Perception (Ministry of Education)
Peking University, Beijing 100871, China
sudan@cis.pku.edu.edu, wxh@cis.pku.edu.cn

Abstract
In this paper, we describe an improvement on probabilistic latent speaker analysis method and investigate the use of probabilistic latent speaker analysis for acoustic model training. By performing co-occurrence analysis between speaker and dominant components, speaker variation is dealt with based on different trajectories. Speech recognition experiment results show that our method, although with a general acoustic model and one-pass decoding, outperform the gender-dependent acoustic model with each gender is given for test set. Further experiment shows that the probabilistic latent speaker training method, although with no adaptation stage and no adaptation data, has outperformed the eigenMLLR adaptation method.

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

1. Introduction
The impact of speaker variability on the automatic speech recognition performance has been acknowledged for years. Speaker adaptation techniques have been fully investigated to deal with this problem and have been successfully applied to different adaptation scenarios. Maximum a posteriori (MAP) [1] and maximum-likelihood linear regression (MLLR) [2] are typical approaches when a certain amount of adaptation data is available. While both of them are based on using appropriate initial model. When speech data from a large number of speakers are pooled and trained, great confusion is introduced. Thus, clustered models such as gender dependent models are created, speaker adaptive training (SAT) [3] is developed and they are all in the effort to build a more compact and canonical model.

In most real scenarios where there is only a very small amount of adaptation data, rapid adaptation techniques such as eigenvoice [4] and eigenMLLR [5] are used. These techniques try to estimate the adapted model in the training speaker subspace expanded by a small number of eigenvectors. However, the eigenvoice method suffers from data sparsity in estimating speaker dependent models and the heavy storage requirements is another problem that prevents its application in large vocabulary speech recognition tasks. The eigenMLLR method alleviates these problems by representing each training speaker with a speaker specific transformation matrix rather than a speaker dependent model. Although transformation matrices are easier to estimate and store, they also limit the accuracy of subspace representation and adaption performance.

The probabilistic latent speaker analysis, as we proposed and described in [6], deals with speaker variation in a different way. It is originally motivated by trajectory folding problem.
tor and its labelled state, the dominant component is the component with the highest probability calculated by:

$$\arg \max_{k} w_k N(x; \mu_k, \Sigma_k)$$  \hspace{1cm} (1)$$

Here, we treat each speaker as a document and all the Gaussian densities as vocabulary. For example, we build a $M \times N$ cooccurrence matrix, if the training data set has $N$ speakers and the acoustic model consists of $i$ states and each state modelled with $j$ Gaussians and $M = i \times j$.

The probabilistic latent speaker analysis is based on an initial speaker-independent (SI) acoustic model. It can be divided into training phase and decoding phase.

2.1. Forced alignment and co-occurrence analysis

State-level forced alignment is performed on training set with an initial SI acoustic model and with each speaker respectively. During the alignment process, the number of times of each Gaussian component that appears as dominant component can be obtained and the co-occurrence matrix can be built, then, PLSA is used to perform decomposition on this matrix and construct a compact latent speaker space. Since 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.

Although in PLSA the sequential order is ignored, since here each Gaussian component belongs to certain state and the states have a left-to-right order, the sequential information is implicitly contained and thus it has certain ability to model speech trajectories.

2.2. Decoding with probabilistic latent speaker model

The probabilities produced by PLSA include $P(z_s)$, $P(g|z)$ and $P(s|z)$, in which $g$ represents a Gaussian component, $s$ represents a speaker and $z$ can be interpreted as a latent speaker cluster. Here, we call these probabilities the probabilistic latent speaker model (PLSM). This model can be integrated into Viterbi aligning step, and probabilistic latent speaker training process. By integrating the probabilistic latent speaker model, when the acoustic model parameters are estimated and probabilistic latent speaker analysis is done, the evaluation of the latent speaker cluster defaults to the distribution in the training data:

$$P(z_k|h_t) = P(z_k)$$  \hspace{1cm} (2)$$

Given current dominant component $g_t$, the distribution of latent speaker cluster can be updated with:

$$P(z_k|h_t) = \frac{1}{t+1} \sum_{q=1}^{t} P(g_t|z_k)P(z_q|h_{t-1}) + \frac{t}{t+1} P(z_k|h_{t-1})$$  \hspace{1cm} (3)$$

Here $h_t$ represents the dominant components history at time-frame $t$. The probability of the current dominant component given the history can be calculated as

$$P(g_t|h_1) = \sum_{k=1}^{K} P(g_t|z_k)P(z_k|h_1)$$  \hspace{1cm} (4)$$

During decoding process, this probability is integrated with the output probability using a linear interpolation method as in Equation (5) to constrain the searching path, making the searching path consistent with a latent speaker cluster, thereby the trajectory folding phenomenon could be alleviated.

$$\log p = (1 - \lambda) \log p(x|s_t) + \lambda \log p(g_t|h_1)$$  \hspace{1cm} (5)$$

3. Utterance-based probabilistic latent speaker analysis

In probabilistic latent speaker analysis as described in the last section, a co-occurrence matrix between speakers and dominant component is constructed. There, each speaker is regarded as a “document” and multiple utterances associated with a particular speaker are used to represent the “document”. A modification can be made here to directly regard each utterance as a “document” and build a co-occurrence matrix between utterances and dominant components, as shown in Figure 1.

Two advantages can be achieved by this modification, first, by treating each utterance as one document, the probabilistic latent speaker model can be generalized to describe not only inter-speaker but also intra-speaker variability. Second, speaker specific information in training set is not necessary, this is an important issue, since in many applications, training utterances are not labelled according to speaker.

4. Probabilistic latent speaker training

A potential deficiency of the probabilistic latent speaker analysis is that its performance also depends on the initial general acoustic model. Although we assume the sequence of dominant component correlates strongly with speaker variability, it is argued that Gaussian mixtures are not totally account for speaker variation. Thereby we propose to integrate the probabilistic latent speaker model to further re-estimating the acoustic model parameters, making it a more appropriate general acoustic model for probabilistic latent speaker analysis.

The proposed probabilistic latent speaker training (PLST) is based on the Viterbi training paradigm which contains two steps, first is to perform forced alignment on the whole training set and the second is to accumulate the statistics and re-estimate model parameters. In the forced alignment step, probabilistic latent speaker model is integrated as in Equation (5), same as its integration in the recognition stage. The reason that PLST is based on the Viterbi training paradigm rather than Baum Welch training paradigm is that best state sequence hypothesis could be obtained during Viterbi aligning process to integrate the probabilistic latent speaker model.

Figure 2 illustrates the probabilistic latent speaker analysis and probabilistic latent speaker training process. By integrating probabilistic latent speaker model into Viterbi aligning step, two advantages can be obtained. First, speaker specific informa-
tion is used during aligning process, that should result in more accurate segmentation and thus a better acoustic model can be re-estimated based on the statistics. Second, Gaussian mixture parameters are tuned to be more inclined to account for speaker variability. Thus the new model become a more appropriate model for probabilistic latent speaker analysis.

5. Experiments

This section presents experiments to evaluate the performance of the utterance-based probabilistic latent speaker analysis and the probabilistic latent speaker training method. We also use gender-dependent acoustic model and perform eigenMLLR adaptation as comparisons.

5.1. Experimental setup

In all our experiments, the test set includes 1114 utterances from Hub4 evaluation data set. Since the 1114 utterances are from different unknown speakers, if adaptation is performed, the scenario is that a quite small amount of adaptation data is available, that is, single utterances which are 2-5s in length, so only rapid adaptation technique can be used.

Acoustic models in our experiments are context dependent phoneme based acoustic models in which each unit is modelled by 3-state left-to-right HMM. The feature extraction is Mel-spectrum based, with corresponding first and second order time derivatives resulting in 39 dimensional features. Channel normalization is applied using cepstral mean normalization over complete recordings.

5.2. Evaluation of utterance-based probabilistic latent speaker analysis

In the experiment to evaluate the utterance-based probabilistic latent speaker analysis method, the training set is 863-1 continuous Mandarin speech corpus. This corpus totally consists of more than 113 hours speech from 166 speakers (83 male and 83 female). After decision tree-based state tying, the SI baseline acoustic model consists of 3000 tied states and each state is modelled with 32 Gaussian mixtures.

Based on the baseline acoustic model, state-level alignment is first performed on training data. In speaker-based probabilistic latent speaker analysis we built a co-occurrence matrix which consists of 3000*32 rows and 166 columns. In the utterance-based method, the matrix consists of 3000*32 rows and 96701 columns, since the training set has totally 96701 utterances. In recognition stage, the PLSM is integrated into decoding process.

For comparison we also construct gender-dependent acoustic models, they are adapted from the baseline acoustic model using MAP adaptation with either complete male or female speech. In recognition stage, speaker gender is directly given for each test utterance instead of using a gender identification step.

Table 1: WER for speaker-based and utterance-based probabilistic latent speaker analysis compared with baseline and gender-dependent model.

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI baseline</td>
<td>42.75</td>
</tr>
<tr>
<td>Gender-dependent</td>
<td>41.71</td>
</tr>
<tr>
<td>SI model decoding with PLSM</td>
<td></td>
</tr>
<tr>
<td>Interp-weight (λ)</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>42.18</td>
</tr>
<tr>
<td>0.15</td>
<td>41.92</td>
</tr>
<tr>
<td>0.20</td>
<td>41.76</td>
</tr>
<tr>
<td>0.25</td>
<td>41.74</td>
</tr>
<tr>
<td>0.30</td>
<td>42.23</td>
</tr>
</tbody>
</table>

Table 2: WER of utterance-based method for different choices of the number of latent speakers.

<table>
<thead>
<tr>
<th># of latent speaker</th>
<th>2</th>
<th>5</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>41.77</td>
<td>41.66</td>
<td>41.47</td>
<td>41.72</td>
</tr>
</tbody>
</table>

Experimental results are given in Table 1, including baseline and gender-dependent performance, speaker-based and utterance-based probabilistic latent speaker analysis performance. Here, we present the results for different interpolation weight λ as in Equation (5). As it is shown, the utterance-based method has shown its superiority over the speaker-based one, though the reduction on WER is not very significant. In the best case, the SI baseline model, when decoding with utterance-based PLSM, has outperformed the gender-dependent models although in the former one no gender specific information was given. We also perform experiments to choose an appropriate number of latent speakers (or K in Equation (4)), experiment results are given in Table 2.

Based on the above results, in our remaining experiments, utterance-base method is used, the number of latent speaker is fixed to be 20 and the interpolation weight λ is fixed to be 0.25.

5.3. Evaluation of PLST

In order to faithfully evaluate the probabilistic latent speaker training method we trained two different baseline acoustic model. The first baseline model is trained with the whole 863-1 training set as described in subsection 5.2. Here, we performed one iteration PLST on the baseline model and then PLSM was

Figure 2: A diagram to illustrate the probabilistic latent speaker analysis and the proposed probabilistic latent speaker training method.
regenerated based on the new model. For comparison, we also performed eigenMLLR adaptation. For eigenMLLR adaptation, full transformation matrices were estimated for each training speaker using all their utterances. 10 primary eigenvectors were preserved after eigen-decomposition. Unsupervised adaptation was performed on the test set and the adaptation data were single utterances.

Speech recognition WER are displayed in Table 3 including the performance for baseline and PLST, both with and without probabilistic latent speaker model. The last row is eigenMLLR adaptation result.

Table 3: WER for probabilistic latent speaker training compared with EigenMLLR adaptation.

<table>
<thead>
<tr>
<th>Model</th>
<th>without PLSM</th>
<th>with PLSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>42.73</td>
<td>41.47</td>
</tr>
<tr>
<td>PLST</td>
<td>41.55</td>
<td>39.66</td>
</tr>
<tr>
<td>EigenMLLR</td>
<td>40.89</td>
<td></td>
</tr>
</tbody>
</table>

The second baseline model was trained with a much larger training set including 863-1, 863-2 and Intel Mandarin speech corpus. The whole training set consists of more than 600 hours of speech data and 1510 speakers. After decision tree-based state tying, the baseline model contains 8000 tied states and each HMM state is modelled with 16 Gaussian mixtures. One iteration PLST was performed on the baseline model and speech recognition results were given in Table 4.

Table 4: WER for probabilistic latent speaker training on the larger training data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Without PLSM</th>
<th>With PLSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>35.13</td>
<td>34.22</td>
</tr>
<tr>
<td>PLST</td>
<td>34.42</td>
<td>32.59</td>
</tr>
</tbody>
</table>

There are several observations that can be made from Table 3 and Table 4. First, the PLST resulted in reduction in WER relative to the baseline model, even without PLSM in the recognition stage. When we ran a further iteration of standard Viterbi training on a Baum Welch trained acoustic model, the recognition performance may not necessarily have improvement since standard Viterbi training is not as reliable as Baum Welch training. While in our experiments, the PLST consistently yielded better results than baseline models. This could be attributed to the speaker specific information introduced by the PLST during the PLST process. Second, after PLST was performed, the reduction in WER with PLSM relative to without PLSM was more significant than the reduction on the baseline model. This result has confirmed that PLST could generate more appropriate model for probabilistic latent speaker analysis. Third, as can be seen in Table 3, the PLST method, recognition with PLSM has outperformed the eigenMLLR adaptation method, although the former used no adaptation data and only included a one-pass decoding procedure.

6. Discussion

The effect of PLST need to be further investigated, it could run iteratively, and this process should converge conceptually and result in a model in which the trajectory patterns of a particular speaker could be stably represented by some certain Gaussian components. However, since the integration of PLSM is in an interpolation way, the interpolation weight is infeasible to be optimized in each iteration. In our experiment, the interpolation weight was fixed and only one iteration was run. The computational requirements in space and time for recognition with PLSM is quite moderate. The PLSM in our experiments has a smaller size than the Gaussian mean parameters. Time required for recognition with PLSM is about 1.5 times of recognition time for a standard recognizer. The additional computation time is necessary for computing probability items in Equation (3) and (4).

7. Conclusions

Further studies on probabilistic latent speaker analysis have been presented in this paper. First, utterance-based probabilistic latent speaker analysis is developed, in which more general variation such as intra-speaker variability should be described. Second, we proposed the use of probabilistic latent speaker model for acoustic model training, resulting in a more accurate and appropriate model for probabilistic analysis. Experiments shown that the utterance-based approach yield better result than gender-dependent models. And by probabilistic latent speaker training approach, WER reduction of about 7% relative to baseline model could be obtained. Furthermore, although the PLST is a one-pass decoding procedure, it has outperform the eigenMLLR adaptation in our experiment.

8. Acknowledgements

The work was supported in part by the National Natural Science Foundation of China (60435010; 60535030; 60605016), the National High Technology Research and Development Program of China (2006AA01Z196; 2006AA010103), the National Key Basic Research Program of China (2004CB318105; 2004CB318005), and the New-Century Training Program Foundation for the Talents by the Ministry of Education of China.

9. References