MAXIMUM A POSTERIORI LINEAR REGRESSION FOR
SPEAKER ADAPTATION WITH THE PRIOR OF MEAN

Chih-Heng Lin, Wern-Jun Wang

Chunghwa Telecommunication Laboratories

12, lane 551, min-tsu rd. sec 5, yang-mei, Taoyuan, Taiwan 326
e-mail: chih@ms.chttl.com.tw Tel: +886 3 424-5861 fax: +886 3 424-4167

ABSTRACT
An efficient method for speaker adaptation (SA) is proposed in this paper. Let the relationship between the mean parameters of adapted model and the mean parameters of the speaker independent (SI) model be represented by sets of linear transformations like that of maximum likelihood linear regression (MLLR) approach, we try to estimate the transformations by maximum a posteriori (MAP) criterion. The prior mean distribution is considered in the estimation.

The experiments on Mandarin speech recognition show the proposed approach is superior to the MLLR approach when only little speech is available for speaker adaptation.

1. INTRODUCTION
A reasonable design of a practical speech recognition system is to start with a less accurate SI system, and then adapt the system to a new user using limited amount of training data from him. On real applications, the amount of speech for adaptation is generally little. A key issue for speaker-adaptation techniques is therefore to extract the appropriate information about the new speaker from as little data as possible.

Presented in [1], the MLLR approach is a useful technique for speaker adaptation. The means of the speaker-adapted models are derived by using a set of linear transformations, which are estimated in a maximum likelihood (ML) criterion from the new speaker’s speech. Experiments has reported that such transformations can capture the new speaker’s characteristics, and are effective even when some models do not have adaptation data. Nevertheless, this approach requires a fair amount of speech data before it starts to be effective.

To conquer the problem of insufficient training data, the MAP estimation is superior to the ML estimation. The MAP estimation employs the prior information besides the new training data, while the ML estimation employs the data only. As an improvement of the MLLR approach, the maximum a posteriori linear regression (MAPLR) approach estimates the transformation matrixes using the MAP criterion. The key issue of this approach is how to select the appropriate prior and estimate the corresponding transformation matrixes. Many previous researches [2][3][4] focused on the prior of the transformation matrix itself. Instead we consider the prior of the mean parameters for the proposed approach. This is because the prior of the mean is informative [5][6] and the transformation matrix can be easily calculated during the MAPLR estimation.

This paper is organized as follows. In Section 2, the MAPLR approach that we proposed is introduced. In Section 3, our implementation of the speaker adaptation experiments is described. In Section 4, the results of the experiments are discussed. Finally, in Section 5, some concluding remarks are given.

2. THE MAPLR APPROACH WITH THE PRIOR OF MEAN
Like the MLLR approach [1], let the adapted mean vector $m_{ad}$ of a Gaussian distribution be a transformation of the corresponding mean vector $m$. The mean vector is of dimension $n$ and the Gaussian distribution belongs to a continuous density hidden Markov model. The transformation is expressed by

$$m_{ad} = Au$$

(1)

where $A$ is the transformation matrix of dimension $n$ by $(n+1)$ and $u$ is the vector $m$ extended by a constant $1$ for the $(n+1)$th dimension.

To estimate the transformation matrix $A$ by MAP approach, the previous researches [2][3][4] start from...
\[ \max_{A} P(A | O) = \max_{A} P(O | A) g(A) \]  
\[ \text{where } O \text{ represents the observations of the distribution and } g \text{ represents the prior distribution.} \]

However the prior probability \( g(A) \) of the transformation matrix is not easily determined, so we propose an alternative approach.

Starting from estimating the adapted mean \( m_{ad} \) by MAP criterion, we have

\[ \max_{m_{ad}} P(m_{ad} | O) = \max_{m_{ad}} P(O | m_{ad}) g(m_{ad}). \]  

Substituting the transformation of (1) for \( m_{ad} \) of (3), we have

\[ \max_{A} P(Au | O) = \max_{A} P(O | Au) g(Au). \]  

Because \( u \) is known, to find the optimal solution \( m_{ad} \) is equivalent to find the solution \( \text{MAP}_A \).

\[ A_{\text{MAP}} = \arg \max_{A} P(O | Au) g(Au) \]  

To estimate the \( A_{\text{MAP}} \) through (5), the EM algorithm[6][7] can be applied. The algorithm is suitable for the estimation of parameters in the incomplete-data cases such as mixture density and hidden Markov model. The EM algorithm has two steps. First the expectation step determines the auxiliary (expectation) function according to the transformation matrix currently estimated, second the maximization step estimates the new transformation matrix by maximizing the auxiliary function. This is achieved by means of differentiating the corresponding auxiliary function to zero by the transformation matrix to be estimated. The two steps are iterated until a stable solution is reached. It is guaranteed that the posterior likelihood is increased during the iterations. Denote \( E \) as the expectation operator, the auxiliary function \( R \) is expressed by

\[ R = Q(A, \tilde{A}) + \log(g(Au)) \]  

with

\[ Q(A, \tilde{A}) = E(\log(P(O | \tilde{A}u)) | O, A) \]  

where \( A \) and \( \tilde{A} \) are the current transformation matrix and the transformation matrix to be estimated respectively.

For MAP estimation, the conjugate prior of the parameter to be estimated is generally considered[5][6]. Because we want to estimate the mean of a Gaussian distribution, the prior distribution \( g(m_{ad}) \) of mean-adapted is assumed to be a single Gaussian distribution. Substituted \( m_{ad} \) by (1), the prior distribution is

\[ g(Au) = \frac{1}{(2\pi)^{\frac{n}{2}} |\alpha \gamma|^{\frac{n}{2}}} e^{-\frac{1}{2} \left( (u - \mu) \alpha^{-1} (u - \mu)^T \right)} \]  

where \( \mu \) is the mean vector and \( \gamma \) is the covariance matrix of the prior distribution. By scaling the covariance matrix using \( \alpha \), the contribution of the prior distribution in (5) is controlled. The ML estimation (i.e. the MLLR case) happens when \( \alpha \) approaches infinity.

Having the prior as (8), the transformation matrix \( A_{\text{MAP}} \) which maximizes the posterior probability of (5) can be estimated. The derivation is similar to that in MLLR approach.

When the transformation matrix \( A \) is common for a tied group of \( J \) Gaussians, (5) becomes

\[ A_{\text{MAP}} = \arg \max_{A} \left[ \Pi_j P(O_j | Au_j) g(Au_j) \right] \]  

and the matrix \( A_{\text{MAP}} \) can be estimated in the same way.

3. EXPERIMENT SETUP

The recognition of continuous Mandarin speech is experimented to test the effect of speaker adaptation. To compare the effect, test speech of a new speaker is recognized by the SI model and the models adapted by the MLLR and the proposed approaches.

The microphone speech from 45 speakers is used in the experiment. Each speaker has about 400 utterances. An utterance consists of 1 to 7 syllables. The database is recorded using 16k Hz sampling rate. The features employed include 12 mel-scaled cepstral coefficients, 12 related delta cepstral coefficients and 1 delta energy coefficient. The window size for the cepstral coefficients is 20ms and is shifted every 10 ms. The window size for the delta coefficients is 7 frames long. The speech of 40 speakers is used for training the SI model and the speech of the rest 5 speakers is used for checking the adaptation efficiency. Each of the 5 test speakers has 200 utterances for adaptation and the rest 200 utterances for testing.

Each mandarin syllable is made up with a beginning initial and an ending final. As described
in [8], the modeling of Mandarin syllable is based on the assumption that the initial is right context dependent on the beginning phone of the following final and the final is context independent. In the experiment, an initial is represented by a hidden Markov model (HMM) of 3 states and a final is represented by a HMM of 5 states. The HMM of a syllable is made up of HMMs of corresponding initial and final, while the HMM of an utterance is made up of HMMs of corresponding syllables and inter-syllable silence. The silence is modeled by 1 state HMM and can be skipped within the state sequence. There are totally 440 states in the HMM and each state has 4 Gaussian distributions.

Based on the ML estimation, the SI HMM is trained first. The model is trained by means of the segmental k-means algorithm [9]. For the purpose of adaptation, the SI distributions provide not only mean bases for transformation, i.e. $u$ of (1), but also the parameters of the prior distributions of mean.

The segmental k-means algorithm is employed to combine with the speaker adaptation algorithm to achieve speaker adaptation effect. There are two steps in the segmental k-means algorithm. At the first step, the speech is labeled optimally based on the adapted model. At the second step, a new transformation matrix is then re-estimated by the EM algorithm based on the labeled speech. The two steps of the segmental k-means algorithm are iterated until the model converges. At the second step, the MLLR transformation matrix is estimated by approach described in [1] and the MAPLR transformation matrix is estimated by the proposed approach described in Section 2.

Assuming the prior mean distribution of each mixture to be a single Gaussian distribution like (12), we choose the mean vector $\mu$ to be $m_{SI}$ and the covariance matrix $\gamma$ to be $\Sigma_{SI}$ following the works of [5][6]. Both the mean vector $m_{SI}$ and the covariance matrix $\Sigma_{SI}$ belong to the corresponding SI Gaussian distribution. Because the covariance matrices of both the adapted distribution and the SI distribution are assumed to be diagonal, the estimation of the transformation matrix can be done easily.

To make the most efficient use of the data available for adaptation from the new speaker, the sharing of one common transformation matrix for all the HMM Gaussian distributions is considered. For each distribution of the adapted model, the mean is derived by transforming the mean of the corresponding SI distribution using the regression matrix estimated, whereas the covariance matrix is kept the same as that of the SI distribution.

The one stage algorithm is used for recognition. The syllable-string hypothesis having the best matching score is decided as the recognition result. The recognition rate is the syllable accuracy rate after subtracting the substitution error, the deletion error and the insertion error form the total number of syllables.

4. RESULTS

The average recognition performance of the five test speakers is listed in the Table 1. The upper part of the Table 1 shows results based on the SI model and is for reference. The lower part of the Table 1 shows results based on the adapted models of the MLLR approach and the proposed MAPLR approach and is for comparing the speaker-adaptation capability.

By varying the number of utterances for adaptation, the performance of the two SA algorithms is evaluated. The numbers of utterances for adaptation are listed in the first column and the corresponding speech durations are listed in the second column. As the amount of speech for adaptation becomes more, the recognition rates of both SA methods become higher. Comparing the performances of the two approaches, the proposed approach performs better than the MLLR approach when the amount of speech is 2 and 5, and the two approaches have similar performance as expected when the amount of speech is increased to 200.

It is noticed when the utterance numbers is 2 or 5, the MLLR model performs even worse than the SI model, but the models adapted by the proposed approach using other values of $\alpha$ perform better than the SI model. As the priori probabilities are varied by $\alpha$ with values smaller than infinity (the MLLR case), the effect of speaker adaptation increases at the beginning then decreases. This indicates an appropriate $\alpha$ is important to enhance the effect of speaker adaptation.

According to the results listed at the Table 1, the proposed approach keeps the speaker adaptation capability even though the speech for adaptation is little. This shows its usefulness on real application.

5. CONCLUSION

Based on the prior distribution of mean, we
have proposed a new MAPLR method for speaker adaptation. The mean of the adapted distribution is derived by transformation from corresponding mean of SI distribution. The estimation of the transformation is based on maximizing the a posteriori probability of mean. To evaluate the posteriori probability, the prior distribution of mean is assumed to be a simple Gaussian with parameters derived from corresponding SI distribution. Experiments on Mandarin speech recognition show the model adapted by the proposed method has better performance than the SI model when there is little speech for adaptation.

6. ACKNOWLEDGMENT

The authors would like to thank Dr. K.Y. Chang, Dr. J.K. Chen, Dr. E.F. Huang and the colleagues of speech recognition group at Chunghwa Telecommunication Laboratories for their invaluable advice and timely help.

7. REFERENCES


<table>
<thead>
<tr>
<th>Utterance no</th>
<th>Duration(sec)</th>
<th>MLLR ($\alpha = \infty$)</th>
<th>$\alpha = 5$.</th>
<th>$\alpha = 1$.</th>
<th>$\alpha = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.5</td>
<td>39.3</td>
<td>54.4</td>
<td>53.9</td>
<td>52.4</td>
</tr>
<tr>
<td>5</td>
<td>6.2</td>
<td>48.9</td>
<td>55.1</td>
<td>54.3</td>
<td>52.8</td>
</tr>
<tr>
<td>10</td>
<td>10.97</td>
<td>55.4</td>
<td>57.2</td>
<td>56</td>
<td>53.4</td>
</tr>
<tr>
<td>200</td>
<td>181.84</td>
<td>60.7</td>
<td>60.8</td>
<td>61</td>
<td>59.6</td>
</tr>
</tbody>
</table>

Table 1. The recognition rates of both SI and SA models