ABSTRACT
In this paper, we considered two representative speaker adaptation (SA) approaches – MAP and MLLR techniques – for our Korean isolated word recognition task. In addition, we proposed a new speaker adaptation algorithm to improve the performance of MAP technique. It is based on least squares method between MAP adapted mean vectors and the corresponding SI mean vectors. The results of our experiment using a CDHMM system indicated that the proposed SA technique yielded high performance improvement in the recognition rate especially when the number of adaptation data is very limited. Moreover, the computational load of proposed technique is much smaller than that of MLLR.

1. INTRODUCTION
In recent years, there is a growing interest in speaker adaptation (SA) techniques, which is an effective means of improving the performance of the speaker independent (SI) speech recognition systems. Among these adaptation schemes, maximum a posteriori (MAP) [1] and maximum likelihood linear regression (MLLR) [2] adaptation techniques are known to be particularly promising.

MAP adaptation tries to find the statistics of the SI models and that of the adaptation data. On the other hand, the MLLR adaptation transforms the SI models by linear transformation on the mean vector of the SI models.

Although MAP adaptation technique is theoretically optimal, its convergence is slow because it does not influence unobserved model parameters. MLLR provides a solution to this problem, which is to estimate a set of transforms of the HMM parameters and to adjust the degree of parameter tying, thus transforming all model parameters even if the number of adaptation data is small. Therefore, the recognition accuracy of MLLR is better than that of MAP when the adaptation data is limited.

2. REVIEW OF MAP AND MLLR ADAPTATION ALGORITHMS
This section contains a brief review of the MAP and MLLR adaptation techniques. In our experiments we use the conventional MAP and a simplified version of MLLR because of its computational efficiency.

2.1. MAP Adaptation
Assume that the mean is random with a prior distribution \( p(\mu) \) and the variance is known and fixed. It can be shown that the MAP estimate for the parameter is given by [3]

\[
\hat{\mu}_{\text{MAP}} = \frac{n\bar{y}}{\sigma^2 + n\sigma^2} + \frac{\sigma^2}{\sigma^2 + n\sigma^2} \bar{y}
\]

where \( n \) is the total number of training samples observed in the corresponding HMM state, and \( \bar{y} \) is the sample mean.

2.2. MLLR Adaptation
Conventional MLLR [2] is the algorithm that finds the
transformation matrix, which maximizes the likelihood of the adaptation data when the SI mean vectors are transformed. It needs a great deal of computations because it uses mean vectors and covariance matrices of Gaussians. On the other hand, simplified MLLR [4] finds the transformation matrix using least squares method between MLLR adapted mean vectors and corresponding feature vectors. It uses only mean vectors, and not the covariance matrices. Therefore the computational load of the simplified MLLR is much less than that of the conventional MLLR.

The simplified MLLR adaptation consists of the following two steps.

Step 1) Estimation of a transform matrix $W$, that is,

$$W = \left( \sum_{i=1}^{N} o_i \Sigma_i^T \right) \left( \sum_{i=1}^{N} \xi_i \Sigma_i^T \right)^{-1}$$

where $N$ is the number of observation vectors, and $\xi_i$ is the corresponding augmented SI mean vectors. This estimate is optimal in the sense that it minimizes

$$\sum_{i=1}^{N} \left\| o_i - \hat{\mu}_{MLLR} \right\|^2$$

where $\hat{\mu}_{MLLR}$ is the MLLR adapted mean vector corresponding to the $i$-th observation.

Step 2) By applying this matrix to the augmented SI mean vector, we can obtain SA mean vector as follows:

$$\hat{\mu}_{MLLR} = W \hat{\mu}$$

3. PREDICTIVE SPEAKER ADAPTATION BASED ON LEAST SQUARES METHOD

When few or no observations of a particular HMM state are available, MAP cannot or poorly estimates the mean vector of that state. As in the most predictive techniques [5]-[7], this can be overcome by finding relationships between the parameters of different models, and applying these relationships to the unseen HMM states.

In this paper, we proposed a new approach, which is to find relationships between MAP estimates and corresponding SI mean vectors. If all the dimensions in the feature vector are independent, we can deal with each vector element independently. This assumption is reasonable in most applications.

3.1. Preliminary Model Adaptation

The model parameters with enough adaptation data are first updated using equation (1). Let the resulting estimate for the mean of the $i$-th model be

$$\hat{\mu}_{MAP} = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in} \end{bmatrix}, \quad i = 1,2,\ldots,M$$

where $M$ is the number of MAP adapted HMM states and $n$ is the dimension of the feature vector.

3.2. Transformation Matrix Calculation

At this stage, to transform each dimension of the mean vector separately, we define new vectors which elements are composed of a certain dimension of the MAP adapted mean vectors in equation (5):

$$\mu_j = \begin{bmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{nj} \end{bmatrix}, \quad j = 1,2,\ldots,n$$

where $\mu_j$ is the $j$-th element of the MAP estimated mean vector of the $i$-th model in equation (5). The transformation matrix is obtained by minimizing each of the following objective functions.

$$F_j = \left\| \mu_j - K_j w_j \right\|^2, \quad j = 1,2,\ldots,n$$

where

$$K_j = \begin{bmatrix} 1 & \xi_{1j} \\ 1 & \xi_{2j} \\ \vdots & \vdots \\ 1 & \xi_{nj} \end{bmatrix}$$

and $\xi_j$ is the $j$-th element of the $i$-th SI mean vector and $w_j$ is a 2x1 vector which consists of the $j$-th row of the transformation matrix $W$ of the following form

$$W = \begin{bmatrix} w_{11} & w_{12} & 0 & 0 \\ w_{21} & 0 & w_{23} & 0 \\ M & M & M \\ w_{n1} & 0 & 0 & w_{n,n+1} \end{bmatrix}$$

This leads to the following formula [8].

$$w_j = \begin{bmatrix} w_{j1} \\ w_{j2} \end{bmatrix} = (K_j^T K_j)^{-1} K_j^T \mu_j, \quad j = 1,2,\ldots,n$$
3.3. Prediction of the Mean Vectors

The last stage of this algorithm is the prediction of the mean vector by using the transformation matrix obtained from the previous stage. This is done as follows,

\[ \tilde{\mu} = W \tilde{\xi} \]  

(11)

where \( \tilde{\xi} \) is the augmented SI mean vector, that is,

\[ \tilde{\xi} = \begin{bmatrix} 1 \\ \xi_1 \\ M \\ \xi_n \end{bmatrix} \]  

(12)

3.4. The Comparison of the Computational Loads

Once the MAP estimates are obtained at first stage, the additional computation of the proposed scheme is very small. Since \( K^T K \times 2 \) matrix, the amount of computation in calculating the inverse matrix is negligible.

Therefore, the amount of computation of the proposed algorithm is slightly larger than that of MAP, but is much smaller than that of MLLR. Moreover, as the adaptation data increase, computational difference between the proposed algorithm and MLLR increases. For example, in case of 10 adaptation data, computational complexity of the proposed algorithm is three times as much as that of MAP, and one 26th as little as that of MLLR. In case of 30 data, it is only twice of that of MAP and one 29th of that of MLLR.

4. EXPERIMENT AND RESULTS

4.1. Experimental Setup

Phonetically optimized word (POW) database and phonetically balanced word (PBW) database from ETRI were used for our experiments [9]. The HTK [10] was used to train and test all models. The speech was coded into 20ms frames, with a frame advance of 10ms. Each frame was represented by 24-component vector including 12 MFCCs and their delta parameters.

POW database and PBW database were used to construct SI HMMs and to evaluate the previously mentioned SA techniques, respectively. The POW database consists of 10 repetitions of 3,848 words from 80 speakers (40 males and 40 females). The models were state clustered triphones, containing a total of 1,597 tied states. Among these words, we have selected 19,026 words to train a set of SI models. This set of SI models gives recognition accuracy of 96.6% on our test database, which will be used for baseline. The PBW database consists of 2 repetitions of 452 words from 70 speakers (38 males and 32 females), and speaker adaptation and testing were carried out using this database. For speaker adaptation, we used 1 repetition of 35 words from each of 9 male speakers. Also, we used other 100 words from the same speakers for testing.

In the following experiments, we investigated the performances of various SA schemes. All of them were performed in a static supervised manner using labeled adaptation data.

4.2. Results and Discussion

Experiments were performed using the adaptation techniques described in Section 2. Figure 1 illustrates the recognition results of MAP, MLLR and the proposed algorithm with varying number of adaptation data. In this figure, we found that the MAP performance is much lower than the baseline performance. One possible reason for this is that, when the number of adaptation data is small, the state that has small adaptation data may be poorly updated by MAP algorithm. However it has not been confirmed yet and requires further investigations. For additional information, in Figure 2, we plotted the recognition accuracy of MAP with enough adaptation data. When the number of adaptation data becomes large, high recognition accuracy is achieved, but one can see that the speed of convergence is slow as expected.

![Figure 1: Performance comparison of the various SA schemes.](image-url)
adaptation, the proposed method, and the SI HMMs, respectively.

Figure 1 also shows that MLLR outperforms MAP when the number of adaptation data is larger than 5. Asymptotic convergence is achieved when the number of adaptation data is about 10, and the accuracy is very high. However, MLLR performance is not guaranteed when the number of adaptation data is very small.

![Figure 2: Recognition results of MAP with sufficient amount of adaptation data.](image)

From Figure 1, we can see that the performance of the proposed algorithm is much higher than that of the conventional MAP algorithm. While the MLLR performance is not guaranteed when the adaptation data is limited, the proposed algorithm records 55% error rate reduction compared to the baseline system even though only one utterance is available. As the number of adaptation data becomes larger, the recognition accuracy of the proposed algorithm is slightly lower than that of MLLR. However, from the viewpoint of computational load, the proposed algorithm has greater strength to implement the on-line adaptation than MLLR. In addition, as the number of adaptation data increases, the MAP performance becomes higher, and hence, the additional performance improvement of the proposed algorithm will be obtained.

5. CONCLUSION

In this paper, we compared various speaker adaptation schemes in the Korean isolated word recognition experiments. Significant increase in recognition accuracy was obtained by those adaptation schemes compared to the baseline system. We have also proposed a new speaker adaptation algorithm that is based on least squares method. Experimental results confirm its superiority to MAP algorithm. Even though only one utterance is available, the error rate reduction of 55% can be obtained compared to the baseline system.

When the number of adaptation data is relatively large, the performance of the proposed algorithm is slightly lower than that of MLLR. However, since the proposed method has much less computational load than MLLR, it has a great advantage in real-time implementation of on-line speaker adaptation.

6. REFERENCES

arch 1997.