# SPEAKER ADAPTATION USING MAXIMUM LIKELIHOOD MODEL INTERPOLATION

*Zuoying Wang    Feng Liu*

Email: gaoyu@hotmail.com
Department of Electronic Engineering,
Tsinghua Univ., Beijing 100084, P.R.China

## ABSTRACT

A speaker adaptation scheme named maximum likelihood model interpolation (MLMI) is proposed. The basic idea of MLMI is to compute the speaker adapted (SA) model of a test speaker by a linear convex combination of a set of speaker dependent (SD) models. Given a set of training speakers, we first calculate the corresponding SD models for each training speaker as well as the speaker-independent (SI) models. Then, the mean vector of the SA model is computed as the weighted sum of the set of the SD mean vectors, while the covariance matrix is the same as that of the SI model. An algorithm to estimate the weight parameters is given which maximizes the likelihood of the SA model given the adaptation data. Experiments show that 3 adaptation sentences can give a signaificant performance improvement. As the number of SD models increases, further improvement can be obtained.

## 1. INTRODUCTION

In recent years, there is a growing interest in speaker adaptation (SA) techniques, which has been shown to be an effective means of improving the performance of the large vocabulary continuous speaker independent (SI) speech recognition systems[1]. While many adaptation schemes have been proposed, MAP estimation[2] and MLLR[3] seem to be particularly promising.

Most speaker adaptation schemes attempt to find a map from the speaker independent acoustic model to the speaker adapted model in such a way as to more closely match the characteristics of the test speaker. For example, the MAP adaptation tries to construct the SA models by finding the statistics of the SI model and that of the adaptation data. Since there are so many parameters need to be estimated, the MAP requires a large amount of adaptation data. On the other hand, the MLLR transforms the SI model by applying a linear transformation on the mean vector of the SI model. The schemes used in[1] [4] extend the idea of MLLR. Instead of transforming the SI model, it makes only use of a subset of the speaker-dependent (SD) models which are acoustically close to the test speaker. The SA models are computed by the average of the set of the transformed SD models.

In this paper, a new SA scheme named maximum likelihood model interpolation (MLMI) is proposed which is based on the assumption that in the feature space, the SA model to a new speaker can be approximated by a linear convex combination of the SD models in the training set, and that these SD models from each training speaker have different effect on the SA model, so the weight parameters corresponding each of the SD models are different. If the characteristics of the test speaker may be closely represented by the linear combination of the set of SD models, MLMI is able to achieve the performance of a SD system. In addition, since there is a small number of parameters to be estimated, MLMI needs only a little amount of adaptation data. Experimental results show that with 3 adaptation sentences, MLMI gives a significant performance improvement in the error rate. In addition, as the number of training speakers in the training set increases, further improvement can be obtained.

Because of the special structure and various characteristic features of Mandarin Chinese[5], the speech recognition system is very different from that of the western language. For sake of ease of interpretation, the speech recognition system especially the acoustic model is first introduced in section 2, and the basic idea of MLMI and the algorithm are described in section 3 and 4, respectively. Section 5 gives the experimental results and the conclusions are drawn in section 6.

## 2. DURATION DISTRIBUTION BASED HMM

The recognition system is composed of two parts, the acoustic part and the language part. The acoustic part converses the input speech data into syllable strings, and the language part converses the syllable strings into Chinese characters. Here we only introduce the acoustic model. In the following development, we assume that each state comprises of a single Gaussian with the full covariance matrix.

The acoustic part is based on a modified HMM called the duration distribution based hidden Markov model (DDBHMM)[6]. The DDBHMM is an inhomogeneous HMM which is based on the fact that the state duration distribution is relatively stationary. In DDBHMM, the duration distribution probability of the state is used instead of the state transition probability which is used in the classical HMM. To be more specific, given T frames of observation feature vector of speech $X = (x_1, x_2, \ldots, x_T)$ and the word string $W = (w_1, w_2, \ldots, w_K)$, the optimum word string $W^*$ is defined as:

$$W^* = \arg\max_W P(X|W)$$

$$= \arg\max_W \{ \max_{S_2,\ldots,S_N} \prod_{i=1}^{N} P_i(\tau_i) \prod_{t=S_i+1}^{S_{i+1}} b_i(x_t) \} \qquad (1)$$

Here, P(X|W) is the probability of the observation sequence X given the word string W, and $\tau_i$ is the number of frames of the observation vectors belonging to i-th state, or the state duration of state i:

$\tau_i = S_{i+1} - S_i$ , i=1,2,...,N

where $S_i$ is the state segment point and N is the number of states. $P_i(\tau)$ is the state duration distribution function of the i-th state. $b_i(x_t)$ is the probability density of observation vector $x_t$ in state i and

$$b_i(x_t) = \frac{1}{(2\pi)^{D/2} |R_i|} \exp\{-(x_t - u_i)^T R_i^{-1} (x_t - u_i) / 2\} \qquad (2)$$

where $u_i$ and $R_i$ is the i-th mean vector and covariance matrix of the model and D is the dimension of the observation vector.

Comparing equ(1) with the conventional HMM, we may see that the state duration probability density of HMM is inherently exponential, while DDBHMM may take arbitrary form, which is more appropriate in application.

## 3. BASIC IDEA OF MLMI

The basic idea of model interpolation is very simple. Given a set of M training speakers and the corresponding speaker dependent models, the speaker adapted model is computed by the linear convex combination of the set of SD models and can be expressed as follows:

$$u_j^{(SA)} = \sum_{m=1}^{M} \alpha_m \cdot u_{mj}^{(SD)} \qquad (3)$$

and

$$U_j^{(SA)} = R_i \qquad (4)$$

with the constraint set:

$$\Omega = \{\alpha_m | \sum_{m=1}^{M} \alpha_m = 1, 0 \le \alpha_m \le 1, m = 1, \ldots M\} \qquad (5)$$

where

$\{\alpha_m | m = 1,2,\ldots, M\}$ is a set of weight parameters corresponding to the set of the training speaker;

M is the number of training speakers in the training set;

$u_j^{(SA)}$ and $U_j^{(SA)}$ are respectively the SA mean vector and covariance matrix belonging to state j. j=1,2,...,N, and N is the number of states in the SA models;

$u_{mj}^{(SD)}$ is the j-th mean vector of the SD model belonging to the m-th training speaker, m=1,2,...,M,

$R_i$ is the j-th covariance matrix of the SI model.

As shown in equation(3-5), the mean vector of the SA model is the weighted sum of the SD means from the set of training speakers, while the covariance matrix is the same as that of the SI models. That is to say, the SA mean vector of any new speaker can be computed by the interpolation of the SD means in the SD model set. Hence the name model interpolation.

Intuitively, when there is no adaptation data available, all the weights should have the same value because all the SD models have the same effect, and then the new mean vector is calculated as the average of all the SD mean vectors, and this becomes a SI model. On the other hand, when the adaptation data from a new test speaker is available, the weights to each SD model should not be the same, since each SD model has different effect on the SA model. Some of the SD models which are closer to the test speaker should have larger weights, while those which are very different from the test speaker should have smaller weights. In the extreme case, when the test speaker is acoustically similar to one of the training speaker m, we can expect that the weight parameter corresponding to speaker m is close to 1, and other parameters are close to 0, and this becomes a SD model.

## 4. THE ALGORITHM OF MAXIMUM LIKELIHOOD MODEL INTERPOLATION

The MLMI tries to estimate the weight vector  to maximize the likelihood of the speaker adapted models given the available adaptation data. Here an object function is defined as follows:

$$J(a) = \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=0}^{T_i - 1} \left\| x_{it} - \sum_{m=1}^{M} \alpha_m u_{mi}^{(SD)} \right\|^2 \qquad (6)$$

where

$a = (\alpha_1, \alpha_2, \ldots, \alpha_M)^T$ is the weight vector;

$x_{it}$ is the t-th frame aligned to the i-th SI mean;

$T_i$ is the number of frames aligned to the i-th SI mean;
n is the number of states in the adaptation data;
$\| \cdot \|^2$ is defined as the norm of the vector:

$$\|x_i\|^2 =< x_i , x_i >= x_i^T \cdot R_i^{-1} \cdot x_i \qquad (7)$$

where $R_i^{-1}$ is the inverse of the i-th SI covariance matrix.

The optimum solution of the weight vector which minimize the objective function J(a) is

$$a^* = \arg\min_{a \in \Omega} J(a) \qquad (8)$$

with the constraint set of equ (5).

From equations above we may see that the optimum solutions $a^*$ also maximize the likelihood of the SA models given the adaptation data, hence the name maximum likelihood model interpolation .

The equ (8) may be rewritten as :

$$J(a) = a^T C a - 2b^T a + q \qquad (9)$$

where C is a M*M matrix, with the component

$$c_{lk} = \sum_{i=1}^{n} < u_{li}^{(SD)} , u_{ki}^{(SD)} > , \; l,k=1,...,M;$$

b is a M-dimension vector

$$b_l = \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=0}^{T_i-1} < x_{it} , u_{li}^{(SD)} >$$

$$= \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=0}^{T_i-1} x_{it}^T R_i^{-1} u_{li}^{(SD)}$$

and

$$q = \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=0}^{T_i-1} < x_{it} , x_{it} >$$

$$= \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=0}^{T_i-1} x_{it}^T R_i^{-1} x_{it}$$

where the symbol < > is defined in equ(7).

Given the constraint condition of equ(5), the optimum solution of equ(9) can be calculated with any constraint optimization algorithm, for example, the gradient projection algorithm, and the procedure is described as follows:

1) Make up a set of SD models for each of the training speaker, comprising a single Gaussian per context-dependent state, and compute the initial SI models $\{u_i^{(SD)}, R_i | i = 1,2,\ldots,N\}$;

2)Initialize the weight vector $\alpha$ with a proper value, for example, 1/M;

3)Decode the adaptation data for the test speaker using SI models, and do a Viterbi-alignment of the adaptation data against the decoded script to assign all the frames of the adaptation data to the corresponding states (supervised or unsupervised);

4) Calculate J(a) according to equ(9) and use the constraint optimization algorithm to find the optimum $\alpha$ that minimizes equ(9);

5) Calculate the SA model according to equ(3).

From the procedure above, we may expect that MLMI is able to achieve the performance of a SD system if we select the set of SD models carefully so that the characteristics of any test speaker may be closely represented by the linear combination of the set of SD models. In addition, since there is only M parameters needed in equ(6), a very little amount of adaptation data is enough to estimate the weight vector. This means low computation cost as well as fast adaptation.

## 5. EXPERIMENT RESULTS

In the following experiments, we investigate the performance of the MLMI. To illustrate the effectiveness of the adaptation algorithm, only the acoustic performances are given. The baseline system has been described above, and details are available in [7].

In 1998, the recognition system won the first place with the character error rate 4% in the National Assessment on the Large Vocabulary Continuous Speech Recognition, which is sponsored by the National 863 High-Tech Project. All the tests are based on this recognition system.

The test data comprises of 100 sentences each from 3 speakers, and there are totally 300 sentences. All the training data and test data are provided by the National 863 High-Tech Project for large vocabulary continuous speech recognition. Supervised batch adaptation is used in all the following tests.

| # of sentences | S1 | S2 | S3 | Avg. |
|---|---|---|---|---|
| 0 baseline | 28.11 | 25.93 | 25.38 | 26.47 |
| 1 | 22.37 | 30.1 | 24.74 | 25.73 |
| 2 | 22.19 | 26.31 | 22.82 | 23.77 |
| 3 | 23.14 | 25.11 | 22.05 | 23.43 |
| 4 | 23.31 | 24.98 | 21.67 | 23.32 |
| 5 | 23.31 | 25.50 | 22.35 | 23.72 |
| 10 | 22.71 | 25.06 | 21.98 | 23.25 |
| 20 | 21.68 | 24.03 | 22.65 | 22.78 |
| 30 | 21.42 | 24.29 | 22.69 | 22.80 |
| 40 | 22.02 | 24.99 | 22.74 | 23.25 |

**Table 1** the error rate with different number of adaptation sentences

In the first test, we assess how the amount of the adaptation data affects the performance of the scheme. Table 1 gives the error rate of MLMI with different number of adaptation sentences. There are totally 80 SD models from 80 training speakers used in the test. 3 test speaker's performances (s1~s3) as well as the average error rate are listed. As can be seen in Table 1, the MLMI is able to reduce the error rate by 10% with only 1~3 sentences. But when there are more sentences in adaptation, the performance improvement is insignificant. This is because there are only 80 parameters to be estimated. So as the number of adaptation sentences increases, the performance saturates quickly. This limitation can be overcome by using more SD models in the scheme which is shown in the next test.

In the second test, we investigate the performance of MLMI with different number of SD models in the training set. 10 sentences are used for adaptation since when the number of SD models increases, more parameters need to be estimated, and thus more adaptation data are needed. Table 2 shows that as the number of training speaker increases, the MLMI gives better performance. In the extreme case, if there are enough SD models in the set so that any SA models for the new speaker can be represented exactly by that set of SD models, speaker dependent performance can be expected.

| #of SD models | avg. Error rate |
|---|---|
| 40 | 25.24 |
| 80 | 23.33 |
| 120 | 22.84 |
| 160 | 22.20 |

**Table 2.** Performance with different number of adaptation sentences

## 6. CONCLUSIONS

In this paper, a new adaptation scheme called MLMI is proposed which is based on model interpolation. MLMI makes up the SA model of the test speaker by the linear combination of a set of SD models. Experimental results show that the scheme gives significant performance improvement with only 3 sentences as adaptation data. As the number of SD models increases, better performances can be obtained. Also the computation cost is relatively low. It still remains unresolved how we could select a set of SD models, which consist of as few training speakers as possible while still can approximate almost all the test speakers.

## REFERENCES

[1] M.Padmanabhan, L.R.Bahl, and M.A.Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulary Speech Recognition Systems", Proc. ICASSP-96

[2] C.J.Legetter and P.C.Woodland, "Maximum likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM's", Computer Speech and Language, vol.9, no.2, pp171-186

[3] J.L.Gauvain and C.H.Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Observations of Markov Chains", IEEE Trans. SAP, vol.2, no.2, pp291-298, Apr.1994

[4] Yuqing Gao, M.Padmanabhan, and M.Picheny, "Speaker Adaptation Based on Pre-Clustering Training Speakers", Proc. ICASSP-98

[5] Linshan Lee, "Voice Dictation of Mandarin Chinese", IEEE SP Mag., July 1997, pp63-101

[6] Zuoying Wang, "Inhomogeneous HMM for Speech Recognition and THED Recognition and Understanding System", Telecommunication Science, Vol.9, No.4, July 1993, pp31-36(in Chinese)

[7] www.thsp.ee.tsinghua.edu.cn