An New Approach for Incremental Speaker Adaptation*

WANG Yu, ZHU Xiaoyan
State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science and Technology, Tsinghua University, Beijing
zxy-dcs@tsinghua.edu.cn

ABSTRACT

This paper presents an approach for fast, incremental speaker adaptation based on MAP algorithm with a simplified MLLR module, which is used to minimizes the mismatches caused by the different speaking environments and speaker connatural characteristics before MAP processing. The most important advantage of the new approach is that it can not only have a quick adaptation with a few short utterances but also be more accurate even in a noisy environment. Experimental results show that using the new approach can improve the word error rate by 20.3% in a quiet environment, and by 27.6% in a noisy environment.

1. INTRODUCTION

Speaker-dependent (SD) systems yield much higher recognition rates than speaker independent (SI) ones. However for many applications, it is not feasible to gather enough data from one speaker to train the system. To overcome the mismatch, speaker adaptation techniques are widely used to achieve recognition rates that come close to SD systems but only use a fraction of speaker dependent data. At the same time, speaker adaptation techniques can also be used to modify the system parameters to better match variations in microphone, transmission channel and environment noise.

In the past, Maximum *a posteriori* (MAP) method has proven to be quite effective for speaker adaptation of Hidden Markov Models (HMM) [1,2,3]. MAP method uses information from an initial model as *a priori* knowledge to complement the adaptation data. This *a priori* knowledge is statistically combined with *a posteriori* knowledge derived from adaptation data based on the amount of adaptation data. When the amount of adaptation data is small, the estimate is tightly constrained by *a priori* knowledge, and the estimate error is reduced. On the other hand, the availability of a large amount of adaptation data lessens the constraints of the *a priori* knowledge, thus preventing loss of the *a posteriori* knowledge. When the amount of adaptation data increases, the MAP estimate will approach the Maximum Likelihood (ML) estimate, and the adaptation performance will be much better. However, the MAP method still has the following problems.

First, when training data is not sufficient, the adaptation is rather coarse. And for many speech applications, such as a Command&Control system, the adaptation data is very limited. It is not desirable for the user to undergo a complex training procedure before he can use the system. This is in contrast to dictation systems.

Second, MAP estimate is based on the initial model, how to make the initial model accurate will directly affect the overall performance.

To solve these two problems, we introduce a simplified Maximum Likelihood Linear Regression (MLLR) module into the incremental adaptation scheme for getting the initial model. A global diagonal regression matrix is used for all acoustic models in the

* Supported by National Nature Science Foundation of China (69982005), and Projects of Development Plan of the State Key Fundamental Research (G199803050703). This paper is for International Symposium on Chinese Spoken Language Processing (ISCSLP2000), Beijing, China, Oct. 2000.

MLLR module, in order to minimize the mismatches caused by different speaking environments and speaker connatural characteristics. After the simplified MLLR module, the MAP processing module uses the whole incremental training data to modify the system parameters of every Gaussian densities again accurately. In our experiments, this approach not only shows good results even if adaptation data is a few short utterances but also has a good performance in noisy environment.

The rest of this paper is organized as follows: In Section 2 the new approach for incremental speaker adaptation is described in detail. After presentation of the experimental results and their discussion in Section 3 we finish with a short conclusion and future perspective in Section 4.

2. THE ADAPTATION APPROACH

Figure 1 illustrates our new approach for incremental speaker adaptation. The whole adaptation scheme is composed of the simplified MLLR module and an incremental MAP processing. In this section, we first briefly review the adaptation method of MAP, then show the simplified MLLR module and give out our incremental adaptation scheme.

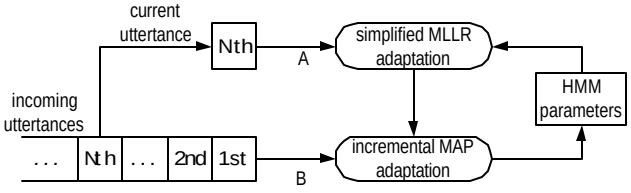

Figure 1: Illustration of the new approach for incremental speaker adaptation[1]

2.1 Review of the MAP adaptation

In MAP approach, we assume that a HMM is characterized by a parameter vector l that is a random vector and that the *a prior* knowledge about the random vector is available and characterized by *a prior* probability density function $P_p(l)$. With the observation data *O*, the MAP estimate is expressed as follow:

$$l_{MAP} = \arg\max_l P(l \mid O) = \arg\max_l P(O \mid l) P_p(l) \quad (1)$$

If we have no prior information, $P_p(l)$ is the uniform distribution and the MAP estimate becomes identical to the ML estimate.

We can use the EM algorithm to estimate the parameters of HMM in Equation (1). Finally we get the solution as follows:

$$\hat{m}_k = \frac{t_k m_k + n_k m_k}{t_k + n_k} \quad (2)$$

where m_k and $\hat{m}_k$ are the k-th Gaussian means of the initial model and the adapted model respectively. m_k is the sample mean of the k-th Gaussian distribution and also the ML estimate. n_k is the total number of adaptation samples observed for the corresponding Guassian mixture component. t_k is a balancing factor between the priori mean and the ML estimate. To describe these variables more concretely, m_k represents an *a priori* guess or knowledge for $\hat{m}_k$ and t_k measures the certainty of this guess or knowledge. Therefore, the specification of this *a priori* parameter t_k is a key factor in MAP estimation. It can be estimated from initial model parameters or the training data [1,2]. In this paper t_k is estimated using the reciprocal of the Mahalanobis distance between the sample mean m_k of adaptation data and the priori mean m_k of the initial model [3].

2.2 Simplified MLLR Module

The MAP method has a good performance when the amount of adaptation data increases, but if the adaptation data is only a few short utterances it can not work well. Correspondingly, the MLLR method has a faster adaptation rate than the MAP method. On the other hand, MAP adaptation also needs an accurate initial model. So we introduce a simplified MLLR module into the incremental adaptation scheme.

In the standard MLLR approach [4,5], the mean vectors m of the Gaussian densities are updated using a $n \times (n+1)$ transformation matrix W calculated from the adaptation data by applying

[1] In this paper, as HMM parameters, only the means are considered, though other parameters, such as variances and mixture weights, can also be adapted.

$$\hat{m}_s = W_s x_s \tag{3}$$

where x_s is the extended mean vector:

$$x_s = [w, m_{s_1}, ..., m_{s_n}]^T = [w : m_s]^T \tag{4}$$

w is the offset term for regression, n is the dimension of the feature vector.

In the standard approach, re-estimation of the transformation matrix is computationally expensive which makes the algorithm time-consuming and unpractical. To simplify the calculations, the estimated transformation matrix W is restricted to a diagonal matrix in our system. Moreover, because the amount of adaptation data is limited, we also use only one globe regression class in MLLR adaptation module. So that, from the adaptation data of all HMM models, we can get a globe transformation matrix to adapt all Gaussians.

By introducing this simplified MLLP module, the mismatches caused by the different speaking environments or the speaker connatural characteristics in all models can be minimized remarkably, and a more accurate initial model can be given to the MAP processing. Our experiment shows this method can work well, especially in a noisy environment.

2.3 Incremental adaptation scheme

In incremental adaptation, we can basically distinguish between two ways of using the adaptation data, which are illustrated in Figure 2. One way *A* is to use only the current utterance to re-estimate the most recently adapted HMM parameters. The drawback of this method is that if an utterance is rather short, the estimation is unreliable and may give bad results. The other way *B* is to adapt the HMM parameters using the whole adaptation sample collected so far. Depending on the adaptation method, however, this approach may cause problems if the computation must be done for the whole sample every time.

In our experiments, both two ways are used respectively. For the simplified MLLR module, we only use the current utterance to adapt the HMM parameters. Because the MLLR has only one globe regression class for all acoustic models, the adaptation data of all models is enough to estimate the globe transformation matrix. Moreover, the module is mainly used for the environment adaptation, and the environment noise is also most changeful, so the way *A* is reasonable for this module. On the other hand, to accurately adapt the mismatches in elaborate speaker characteristics of every model, the way *B* is used in the MAP progressing. For reducing the computation of the whole samples, we reuse some values from the previously observed samples[2]. Our whole incremental adaptation scheme is illustrated in Figure 3.

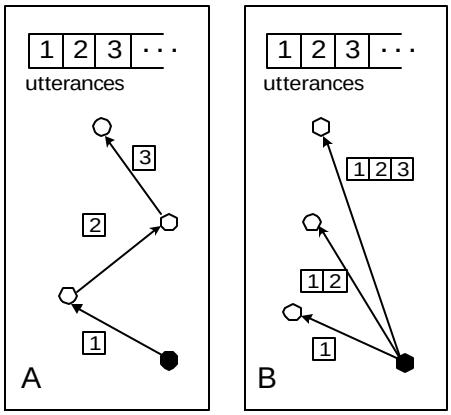

Figure 2: Illustration of two incremental adaptation ways: *A* and *B*

3. EXPERIMENTAL RESULTS

Training and testing data were from a Chinese database, CIDS speech database[3]. A set of SI models was trained using 40 speakers from CIDS (2800 utterances). We used 3-state HMMs with 5 Gaussian per state. The speaker adaptation experiments were conducted on two test sets: Set A and Set B. Set A consisted of 10 speakers from CIDS recorded in a quiet environment. For testing our new approach under various noise conditions, we also recorded 10 speakers' speech (Set B) in a noisy office by different microphones. All results shown in this paper are averaged over all 10 speakers in Set A or Set B.

---

[2] A detailed description can be found in [5]

[3] CIDS is recorded and made by State Key Laboratory of Intelligent Technology and Systems.

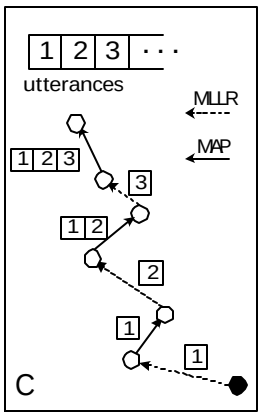

Figure 3: Illustration of our incremental adaptation scheme

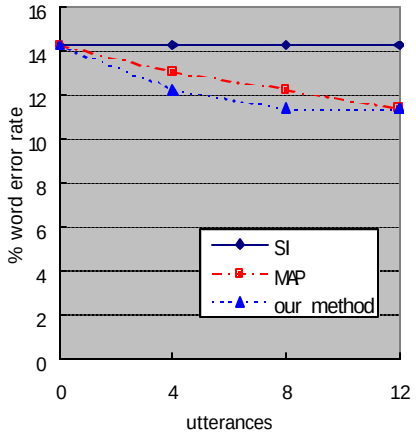

Figure 4: %WERs of Set A using the SI system, MAP adaptation and our incremental adaptation after about 4, 8, 12 utterances.

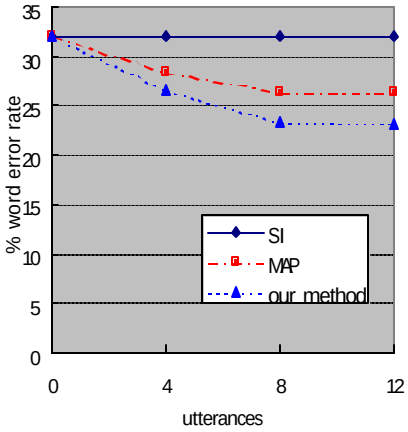

Figure 5: %WERs of Set B using the SI system, MAP adaptation and our incremental adaptation after about 4, 8, 12 utterances.

Table 1 shows the results for the MAP method and our new approach discussed above in % word error rate (WER). It is remarkable that our incremental adaptation scheme can improve the recognition result more effectively than the MAP adaptation, even in a noisy environment, such as Set B. After only 8 utterances our new approach can achieve a WER reduction of 20.3% in Set A and 27.6% in Set B.

Table 1: WERs using the SI system, MAP adaptation and our incremental adaptation after 8 utterances.

WER(%)	Set A	Set B
SI	14.3	32.2
MAP Adaptation	12.3	26.4
Our Incremental Adaptation	11.4	23.3

Figure 4 and Figure 5 illustrate the all experiment results of different adaptation methods after about 4, 8, 12 utterances, for Set A and Set B. They show the system perfo rmance is improved by the increased amount of training data. However, after 8 utterances the improvement is not significant, so for our new method 8~10 utterances are enough. From the figures, it is clearly demonstrated that our incremental adaptation scheme has better performance than the MAP adaptation even when the amount of training data is limited. The comparison of the two figures also shows that our incremental adaptation scheme is working more effectively in the noisy environment.

4. CONCLUSION

A new approach for rapid, incremental speaker adaptation has been presented in this paper. By introducing a simplified MLLR module to minimize the environment mismatches, the incremental MAP approach shows good results even if adaptation is conducted after only a few short utterances. In our new scheme, two incremental adaptation ways are used respectively for MLLR and MAP. Experimental results show that using the new approach can improve the WER by 20.3% in a quiet environment and by 27.6% in a noisy environment. These results demonstrate that this new approach is well suited for the robust speech recognition even when the adaptation data is limited. Future works involve experiments on much large vocabulary robust speech recognition and some application systems.

REFERENCES

[1] C-H. Lee, C-H. Lin, B-H. Juang, A Study on Speaker Adaptation of the Parameters of continuous density Hidden Markov Models, *IEEE Trans on Speech Signal Proc.*, VOL.39, NO.4, pp.806~814, Apr. 1991.

[2] Seyed Mohammad Ahadi-Sarkani, Bayesian and Predictive Techniques for Speaker Adaptation, *PhD Thesis*, Cambridge Univ., 1996.

[3] Masahiro Tonomura, Tetsuo Kosaka, Shoichi Matsunaga, Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum a Posteriori Estimation, *Computer Speech and Language*, pp.117~132, 1996.

[4] C.J. Leggetter, P.C. Woodland, Speaker Adaptation of HMMs using Linear Regression, *Technical Report*, Cambridge Univ., 1994.

[5] C.J. Leggetter, Improved Acoustic Modeling for HMMs using Linear Transform, *PhD Thesis*, Cambridge Univ., 1995.

[6] Michal Schuβler, Florian Gallwitz, Stefan Harbeck, A Fast Algorithm for Unsupervised Incremental Speaker Adaptation, *Proceedings of ICASSP97*, Munich, 1997.