

# UBM-BASED INCREMENTAL SPEAKER ADAPTATION

*TingYao Wu*<sup>1</sup>

Center for Information Science,  
Peking University  
Beijing, China, 100871  
tywu@cis.pku.edu.cn

*Lie Lu*

Microsoft Research Asia  
Beijing, China, 100080  
llu@microsoft.com

*Ke Chen*

Birmingham University  
Birmingham, UK  
K.Chen@cs.bham.ac.uk

*Hong-Jiang Zhang*

Microsoft Research Asia  
Beijing, China, 100080  
hjzhang@microsoft.com

## ABSTRACT

This paper addresses a novel algorithm of incremental speaker adaptation (ISA) based on universal background model (UBM) for saving storage and real-time processing. This algorithm can be seen as an extension of traditional speaker adaptation. It consists of two steps, adaptation and combination. It not only considers the speaker's characteristics in limited training data, but also prohibits over-fitting of the updated model. The incremental adaptation algorithm needs little storage and meets the requirement of real-time processing. In order to evaluate the efficiency and effectivity of the proposed approach, a real-time speaker segmentation system for broadcasting news is built. Experiment results demonstrate that our approach yields real time operation and achieves satisfactory performance.

## 1. INTRODUCTION

Speaker adaptation has been a crucial problem in speech recognition and speaker recognition recently. Speaker adaptation is often used to obtain a proper speaker-dependent model for the specific speaker, by updating a speaker-independent model with a speech set from the corresponding speaker.

Several approaches have been presented for speaker adaptations. The widely used algorithms are Vector Quantization (VQ) [6], Maximum Likelihood Linear Regression (MLLR) [2] and Maximum a Posteriori (MAP). Reynolds [3] also proposed a speaker adaptation approach based on universal background model (UBM), which is now widely used in speaker recognition. In this approach, a speaker-independent UBM is trained by plenty of speech data. A speaker model is derived by updating the well-trained parameters in UBM via adaptation using the speech data of the corresponding speaker. Experimental results showed that the adaptation algorithm can obtain relatively accurate speaker model. However, in this approach, all the speech data belonging to one speaker is needed to adapt the speaker model. Thus, it is not suitable for real-time processing.

However, real-time processing is necessary in some applications, such as, real-time speaker segmentation system [4][5]. In such a real-time processing system, the received speech data has to be processed immediately and then discarded; it is not affordable to store all received data and finally process them, which cost much storage and waiting time. Thus, to achieve a more flexible scheme, the speaker model should be adapted or updated

incrementally by an efficient, low storage and high-accurate speaker adaptation approach.

Traditionally, VQ can be used to adapt the speaker model (codebook in this case) incrementally according to update the statistical parameters of codebook. However, as a simplified example of Gaussians Mixture Model (GMM), VQ is less accurate than GMM since it only considers the influence of inquiry feature vector to the closest codebook. Moreover, in real time speaker segmentation, it is also difficult to achieve the original codebook for unknown speaker.

Therefore, to extend the traditional speaker adaptation and keep high accuracy, an incremental speaker adaptation (ISA) base on UBM is proposed. It obtains a speaker-dependent gradually by adapting the UBM model when the speaker data increase incrementally. The proposed algorithm consists of two steps: adaptation and combination. It can restrict the over-fitting of adapted speaker model, only needs little storage requirement, and is suitable for real-time processing, although it sacrifices little accuracy.

The rest of paper is organized as follows. The incremental speaker adaptation algorithm is introduced in Section 2. Section 3 introduces our real-time speaker segmentation system briefly, as an instance of applying incremental speaker adaptation algorithm. The experimental results are shown in Section 4 and the conclusion is given in Section 5.

## 2. INCREMENTAL SPEAKER ADAPTATION

In this section, we will first introduce the traditional speaker adaptation based on UBM. Then the algorithm of incremental speaker adaptation is presented and illustrated.

### 2.1. Traditional speaker adaptation based on UBM

As Reynolds proposed [3], a speaker model can be adapted by updating the statistical parameters of speaker-independent GMM-UBM model. Similar to the Expectation-Maximum (EM), the algorithm of speaker adaptation based on UBM also consists of two-steps.

The first step is identical to Expectation step of EM algorithm, where it estimates the expectation of statistical parameters of GMM. Denoting the training feature vectors of speaker be  $X = (x_0, x_1, \dots, x_{K-1})$ , where  $K$  is the volume of training set, and the UBM model be  $S$ -Gaussian  $G_U(\varnothing, m, \Sigma)$ , we compute:

---

1. This work was completed when the first author was a visiting student in Media Computing group, Microsoft Research Asia.

$$P(s | x_t) = \frac{\omega_s p_s(x_t)}{\sum_{j=1}^S \omega_j p_j(x_t)}, \quad (1)$$

where

$$p(x_t) = \frac{1}{(2\pi)^{\frac{s}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{(x_t - m)^T \Sigma^{-1} (x_t - m)}{2}\right\}. \quad (2)$$

Then the statistic expectations of weights, means and variance are updated:

$$\gamma_s = \sum_{k=0}^{K-1} P(s | x_t), \quad (3)$$

$$E(x_t) = \frac{\sum_{k=0}^{K-1} P(s | x_t) x_t}{\gamma_s}, \quad (4)$$

$$E(x_t^2) = \frac{\sum_{k=0}^{K-1} P(s | x_t) x_t^2}{\gamma_s}. \quad (5)$$

In the second step, the parameters in UBM are updated. This step is not identical to Maximum step of EM algorithm:

$$\hat{\omega}_s = [\alpha_s^\omega \gamma_s / (K-1) + (1 - \alpha_s^\omega) \omega_s] \beta, \quad (6)$$

$$\hat{m}_s = \alpha_s^m E(x_t) + (1 - \alpha_s^m) m_s, \quad (7)$$

$$\hat{\Sigma}_s = \alpha_s^\Sigma E_s(x_t^2) + (1 - \alpha_s^\Sigma) (\Sigma_s + m_s^2) - \hat{m}_s^2. \quad (8)$$

where  $\beta$  is the scale factor, and  $\alpha_s^\rho$ ,  $\rho \in (\omega, m, \Sigma)$  is data-dependent coefficient. It allows a mixture-dependent adaptation of parameters.

The above two steps iterate until the convergence condition is satisfied.

The published results indicate that the adaptation based on UBM is better than the standard approach of maximum likelihood training of a model [7]. Thus, it provides more accurate speaker models in speaker adaptation. However, it requires receiving all the speaker data at one time before adaptation. This requirement is usually unsatisfied in real time processing. It is necessary to achieve speaker adaptation incrementally for real-time requirements. In this paper, incremental speaker adaptation based on UBM is proposed to solve this issue.

## 2.2. Incremental speaker adaptation model based on UBM

Because of the constraints of real time operation, we can not save all received data since the length of speech stream is unknown. It is also not affordable to store all received data and finally process them, which cost much storage and waiting time. Hence, we have to adapt the speaker model instantly using current received data then discard them when new data comes. At each adaptation, the received data is very limited, for example, about 0.5s in our real time speaker segmentation system. If we update the statistics parameters incrementally using the approach of traditional speaker adaptation directly, the model parameters would be over-fit to this small data and biased seriously after iterating several times. The contributions of each segment belonging to one speaker at different time should be the same. If we employ traditional speaker adaptation method, the new coming data will contribute much more than former received data after several iterations. Thus, the model would incline towards the characteristics of new data and be over-fit after updating, but not balance all segments. In order to prevent the over-fitting of

adapted speaker model, a new method is needed for incremental speaker adaptation.

The proposed incremental speaker adaptation also consists of two steps, adaptation and combination, as the Figure 1 illustrates.

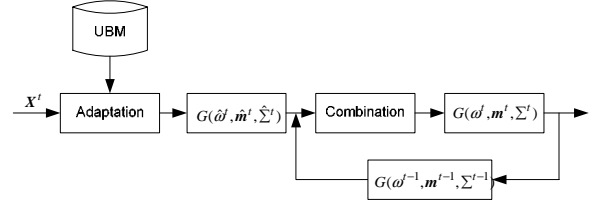


Figure 1. The structure of incremental speaker adaptation

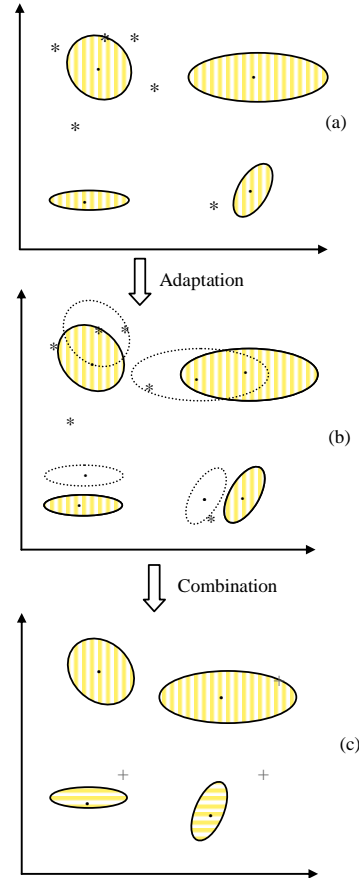


Figure 2. Pictorial illustration of the two steps in ISA algorithm. '\*' and '+' represent the received feature vector in time  $t$  and  $t+1$  respectively. (a) Original status. Ellipses represent the distribution of speaker model at  $t-1$ . Feature vectors '\*' are received in time  $t$ . (b) Adaptation step. Dashed ellipses represent the distribution of  $G(\hat{\omega}^t, \hat{m}^t, \hat{\Sigma}^t)$ , which adapted from UBM using vectors '\*'. This step is identical to adaptation of traditional speaker model based on UBM. (c) Combination step. Ellipses represent the distribution of speaker model at  $t$ . Feature vectors '+' in time  $t+1$  are received.

The first step is identical to traditional adaptation approach, which has been described in Section 2.1. Suppose at time  $t$ , the existing adapted speaker model is  $G(\omega_s^{t-1}, \mathbf{m}_s^{t-1}, \Sigma_s^{t-1})$  and the training set currently received is  $\mathbf{X}^t = (\mathbf{x}_0^t, \mathbf{x}_1^t, \dots, \mathbf{x}_{k_t-1}^t)$ , where  $k_t$  is the number of frames at time  $t$ . We achieve the estimated statistical parameters by adapting speaker independent UBM, and denote those parameters as  $G(\hat{\omega}_s^t, \hat{\mathbf{m}}_s^t, \hat{\Sigma}_s^t)$ . Here we do not adapt the covariance in order to prohibit the distribution of GMM too sharp and the over-fitting of the adapted speaker model. This step is also called adaptation.

The second step is to combine the currently adapted model  $G(\hat{\omega}_s^t, \hat{\mathbf{m}}_s^t, \hat{\Sigma}_s^t)$  with existing model  $G(\omega_s^{t-1}, \mathbf{m}_s^{t-1}, \Sigma_s^{t-1})$ . Let the total number of frames at time  $t-1$  be  $F_{t-1} = \sum_{j=1}^{t-1} k_j$ , the parameters of our new updated GMM are then combined by:

$$\omega_s^t = \frac{\omega_s^{t-1} F_{t-1} + \hat{\omega}_s^t k_t}{F_t}, \quad (9)$$

$$\mathbf{m}_s^t = \frac{\omega_s^{t-1} F_{t-1} \mathbf{m}_s^{t-1} + \hat{\omega}_s^t k_t \hat{\mathbf{m}}_s^t}{\omega_s^{t-1} F_{t-1} + \hat{\omega}_s^t k_t}. \quad (10)$$

This combination step allows for efficient computation of weights and means of GMM components by using above recursions.

A more visualized illustration of the two-step procedure of incremental speaker adaptation based on UBM is shown in Figure. 2. A transitional speaker model is achieved by updating UBM from current data in adaptation step. Thus, this transitional model is fit for new received data. In the second step of combination, the new adapted model combines the transitional model and existing model. Therefore, it integrates the characteristics represented by current data and former data, and prohibits the over-fitting caused by the small set training data in the procedure of adaptation.

### 3. SPEAKER SEGMENTATION

In order to further evaluate the performance of proposed ISA, a real time speaker segmentation system is built. Real time speaker segmentation is to detect speaker change points in a speech stream in real time and then segment the stream into homogeneous speaker clips. There is no any prior knowledge on speakers. Hence no data can be used to train appropriate models for speakers *a priori*. Speaker model should be established incrementally when the more and more speaker data is received from the speech stream.

The flow diagram of the real-time speaker segmentation system is illustrated in Fig. 3. It consists of two modules: pre-segmentation and refinement. Potential speaker change is found in pre-segmentation by comparing the dissimilarity between two adjacent sub-segments. These potential speaker change points would be verified in refinement using speaker model adapted incrementally.

The input speech stream is first segmented into 3s sub-segments with 2.5s overlapping. The first 3-second sub-segment is used as the basic unit for initializing current speaker model. The dissimilarity is estimated between each two neighboring

sub-segments. In pre-segmentation module, the peaks in the dissimilarity sequence are hypothesized to be potential speaker change points. If no speaker change boundary is detected, it means the current sub-segment is from the same speaker as the previous sub-segment. Thus, the current existing speaker model can be updated using this available new data.

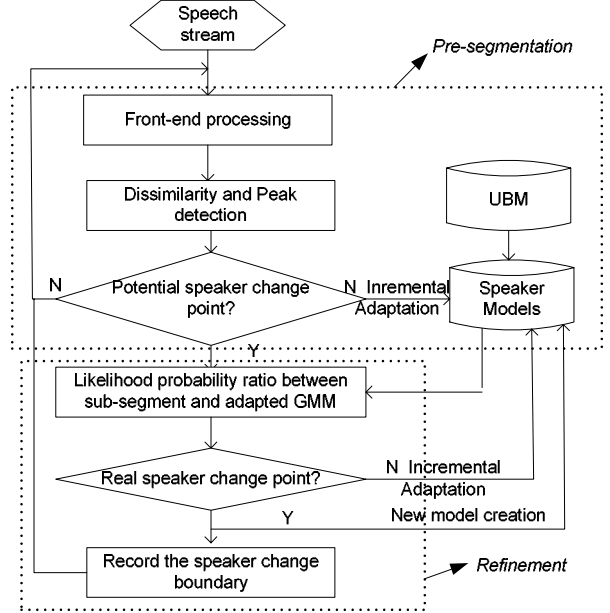


Figure 3. A brief flow diagram of real-time speaker segmentation system

The current speaker model is adapted by currently received speech data using the method of incremental speaker adaptation based on UBM until a potential speaker change point is found. Log-likelihood ratio between the existing speaker model and current speech sub-segment will be estimated to verify whether this potential change is a real change boundary in refinement module. If this potential speaker change is not a real speaker boundary, current existing speaker model will be adapted incrementally again using current sub-segment data. Otherwise, a new speaker model will be created from UBM to substitute for existing speaker model using the current sub-segment.

Since the paper mainly focuses on incremental speaker adaptation, and speaker segmentation system is used only as one of its application, for more detail on this system, please refer to [4][5].

### 4. EXPERIMENTAL RESULTS

In this section, the proposed ISA algorithm is evaluated on our news broadcasting database and in speaker segmentation system.

#### 4.1. Database

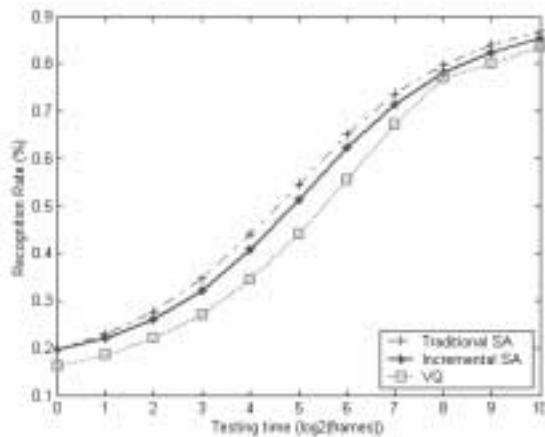
The evaluation of the proposed incremental speaker adaptation is performed on Hub-4 1997 English Broadcast News Speech Database. It consists of about 97 hours news broadcasting, which are from different radios, such as CNN, ABC, CRI and C-SPAN;

About 10 hours speech data is selected randomly for training speaker independent GMM-UBM, and the remaining speech data is for evaluation. Each testing file is either about 30 minutes or 60 minutes, and there are about 30 speakers and about 100-200 speaker changes in each file.

## 4.2. Experiment Results

Twenty broadcasting news files, totally about 10 hours length, are randomly selected to train speaker independent GMM-UBM. Speech data are extracted according to the corresponding transcription files. Furthermore, the silence segments in speech data are discarded using simple energy threshold so that only speech data (loud enough) are considered. These data are blocked into 25ms-frame without overlapping. 16-order MFCC are extracted from each frame; and GMM-UBM is trained by the classical EM algorithm.

In order to evaluate the performance of proposed ISA method, 50 speakers in testing set are selected to perform a speaker identification system. Each speaker model is trained or updated by 30-second speech data. The residual data are used as testing set. Our proposed approach is compared with traditional speaker adaptation based on UBM, and incremental speaker adaptation based on VQ, with different testing length. Figure 4 shows their recognition accuracies in speaker identification system with different testing length. It shows our proposed ISA is comparable with traditional non-incremental speaker adaptation and better than incremental speaker adaptation based on VQ. For example, when the length of testing utterance is 1.6s, the performance of our ISA approach is about 7% higher than VQ, but just a little (about 2%) less than traditional speaker adaptation based on UBM. The loss is paying for the cost of reduction of the storage requirement and real-time processing.



**Figure 4.** Recognition rates of traditional non-incremental speaker adaptation (Traditional SA), VQ and incremental speaker adaptation (Incremental SA) in different testing length. The length of each frame is 25ms

The performance of real-time speaker segmentation system is also tested with using proposed ISA. False alarm rate (FAR) and missed detection rate (MDR) are used to evaluate the performance. In pre-segmentation, we can allow more FAR but keep less MDR, since we could correct many false potential changes in the refinement stage. The results of FAR and MDR

are shown in Table 1. It can be observed that in the refinement stage, we decrease many false alarms while sacrificing few missing detection, after utilizing the speaker model adapted incrementally based on our proposed ISA.

	Pre-segmentation	Refinement
FAR	33.8%	19.23%
MDR	10.83%	13.65%

**Table 1.** False Alarm Rate (FAR) and Missed Detection Rate (MDR) in speaker segmentation module

## 4.3. Computation costs and storage requirement

We also test the computation load of the real time speaker segmentation with proposed ISA. With P4 1.8GHz PC /Windows XP operation system, the segmentation is finished synchronously with broadcasting news using 20% CPU time.

## 5. CONCLUSION

In this paper, a novel algorithm of incremental speaker adaptation based on UBM is proposed. The algorithm helps to establish an accurate speaker model incrementally when the speech data is received gradually. It is very suitable for real-time processing and needs a little storage. The performance of our proposed ISA is better than that of VQ, and comparable with that of traditional non-incremental speaker adaptation in a speaker identification system, while our method needs little storage and does not add more computational cost. Incremental speaker adaptation is also integrated in a real time speaker segmentation system. The results are also very encouraging.

## 6. REFERENCES

- [1] J. P. Campbell, JR. "Speaker Recognition: A Tutorial", *Proc. of the IEEE*, v1.85, No. 9 (1997), pp.1437-1462.
- [2] S.-J.Doh, and R. M. Stern, "Inter-class MLLR for speaker adaptation", *Proceeding of International Conference of Acoustics, Speech, and Signal Processing* (2000), Vol. 3, pp. 1543-1546.
- [3] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing* 10 (2000), pp.19-41.
- [4] T. Y. Wu, L. Lu, K. Chen, H. J. Zhang, "UBM-Based real-time speaker segmentation for broadcasting news", submitted to *International Conference on Acoustic, Speech and Signal Processing* (2003)
- [5] L. Lu, H.J. Zhang, "Speaker Change Detection and tracking in Real-time News Broadcasting Analysis," *Proc of ACM Multimedia* (2002), pp. 602- 610.
- [6] F. K. Soong, A.E. Rosenberg, L. R. Rabiner, and B. H. Zhuang, "A vector quantization approach to speaker identification", in *Proceeding of International Conference of Acoustics, Speech, and Signal Processing*, Tampa (1985), FL, pp.387-390.
- [7] D.A. Reynolds, Comparison of background normalization methods for text-independent speaker verification, In *Proceedings of the European Conference on Speech Communication and Technology* (1997), pp. 963-966