UBM-BASED REAL-TIME SPEAKER SEGMENTATION FOR BROADCASTING NEWS

TingYao Wu[1] *Lie Lu* *Ke Chen* *Hong-Jiang Zhang*

Peking University, Beijing, China, 100871, tywu@cis.pku.edu.cn
Microsoft Research Asia, Beijing, China, 100080, llu@microsoft.com
Birmingham University, Birmingham, UK, K.Chen@cs.bham.ac.uk
Microsoft Research Asia, Beijing, China, 100080, hjzhang@microsoft.com

ABSTRACT

This paper addresses the problem of real-time speaker change detection in broadcast news, in which no prior knowledge on speakers is assumed. Our speaker segmentation is a "coarse to refine" process, which consists of two stages: pre-segmentation and refinement. In the pre-segmentation stage, a new approach based on Gaussian Mixture Model - Universal Background Model (GMM-UBM) is proposed to categorize feature vectors into three sets, i.e. reliable speaker-related set, doubtful speaker-related set and unreliable speaker-related set, in order to enhance the effect of the reliable speaker-related feature vectors. Then potential speaker change boundaries are detected based on a novel distance measure. In the refinement stage, incremental speaker adaptation (ISA), which is suitable for real-time requirement, is proposed to obtain considerably precise speaker models so that the potential speaker change boundaries can be confirmed and refined. Experimental results demonstrate our approach yields the satisfactory performance.

1. INTRODUCTION

Segmenting and indexing the speech stream based on speaker identities are very helpful in many scenarios, such as news broadcasting and net-meeting. Furthermore, real-time processing is also necessary in some applications with real-time requirement. So, in this paper, we propose a real-time speaker segmentation system for broadcast news processing. The objective is to detect speaker change boundaries in real-time speech stream and to segment the stream into homogeneous speaker clips regardless of changes of the background environment or channel conditions.

In real-time speaker segmentation, there is not any prior knowledge on speakers. Thus, no data can be used to train appropriate models for speakers *a priori*. On the other hand, the audio stream may be so long that it would cost too much storage if we store all received data. Thus, an efficient, low storage and high-accurate speaker segmentation approach is necessary.

In our former work [2][5], we have built a speaker segmentation system. It is a "coarse to refine" process, which consists of two stages: pre-segmentation and refinement. The stage of pre-segmentation is based on discriminative distance between two adjacent sub-segments. But it is unavoidable that there exist some non-speech frames, such as background noise or music, in a sub-segment of speech stream. It compromises the accuracy of speaker segmentation.

In order to tell background-related information from speaker-related information, Beigi [4] categorized the frames in one sub-segment into three distinct classes, i.e. silence-related or background-related frames, speech-related frames and speaker-related frames, using K-Means algorithm. However, K-Means algorithm can hardly tell which class is crucial to represent the speaker characteristics. Moreover, the iterative computation in K-Means is not avoidable, which may prohibit the applicability of this approach from a real-time task. Therefore, a new categorization approach based on universal background model is proposed. It categorizes the feature vectors into three categories: reliable speaker-related frames, doubtful speaker-related frames and unreliable speaker-related frames. These three categories can be properly discriminated based on the confidence of the frame to the speaker UBM model. In addition, this approach avoids the iterative computation, and it is more suitable for real-time processing.

In the refinement stage, pre-segment results are confirmed and refined by employing an accurate speaker model, which is adapted from new received data step by step. However, traditional speaker adaptation requires all the training data by expectation-maximization (EM) algorithm [3] and is not suitable for our real-time processing because of its computational complexity and storage requirement. Therefore, incremental speaker adaptation (ISA) is proposed to update current speaker model when the speaker data increase.

The rest of paper is organized as follows. The overview of our system is described in Section 2. Section 3 discusses our approach to feature categorization and pre-segmentation issues. Then ISA and refinement stage are described in Section 4. Section 5 shows our experimental results. The conclusion is drawn in Section 6.

2. SYSTEM DESCRIPTION

The flow chart of the proposed real-time speaker segmentation system is illustrated in Figure 1. It mainly consists of two modules: pre-segmentation and refinement.

In front-end processing, the input speech stream is first segmented into 3s sub-segments with 2.5s overlapping. That is, the resolution of the speaker segmentation algorithm is 0.5s, and the 3-second sub-segment is used as the basic unit for initializing current speaker model and comparing the dissimilarity. The sub-segment is further divided into non-overlapping 25ms-long frames, where 16-order Mel-scaled Cepstrum Coefficients (MFCC) are extracted [2].

1. This work was completed when this author was a visiting student in Media Computing Group, Microsoft Research Asia.

In pre-segmentation module, feature vectors in a sub-segment are categorized by UBM first into three categories, namely, reliable speaker-related frames, doubtful speaker-related frames and unreliable speaker-related frames. Only the reliable speaker-related frames are utilized to compute the dissimilarity between each two adjacent sub-segments. Then potential speaker change boundaries are detected according to this dissimilarity sequence. If no potential speaker change is found, current existing speaker model will be adapted incrementally using the just received speech data.

Once a potential change point is found, the dissimilarity 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 a false alarm, 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. Here, a novel ISA approach is employed incremental speaker updating.

3. FEATURE CATEGORIZATION AND PRE-SEGMENATION

In this section, we propose a new approach to categorize the feature vectors based on UBM-GMM. We also present a new distance measure for computing dissimilarity between two adjacent sub-segments.

3.1. Feature categorization based on UBM-GMM

In order to discriminate the difference between speakers, channels and environments, Beigi [4] implemented K-Means algorithm to categorize the frames in each sub-segment into three clusters. These three clusters represent background-related frames, speech-related frames and speaker-related frames. However, it is not proved that speech can be split into these three clusters. Moreover, it is also difficult to know exactly which class is corresponding to speaker-related class by this simple unsupervised cluster algorithm. If we can know the exact type of each cluster, it is possible to grasp the more reliable speaker characteristics. Thus, feature categorization based on GMM-UBM is proposed to solve these problems.

In our system, speaker-independent UBM is trained off-line by using plenty of speech data in broadcast news through EM algorithm. Such UBM model represents the global speaker characteristics. For a feature vector in the speech stream, we can get its speaker-related confidence according to its likelihood probability to UBM model.

Denote GMM-UBM as $G(\omega_s, \boldsymbol{m}_s, \sum_s)$ $(0 < s \le S-1)$, where S is the number of Gaussians in GMM, ω_s, $\boldsymbol{m}_s$ and $\sum_s$ are the weight, mean and diagonal covariance of each Gaussian component; and denote the *N*-dimensional feature vectors of the *i*-th sub-segment in a speech stream as $\boldsymbol{X}_i = (\boldsymbol{x}_0^i, \boldsymbol{x}_1^i, ..., \boldsymbol{x}_{k_i-1}^i)$, where k_i is the number of frames in *i*-th sub-segment. Thus, the confidence of $\boldsymbol{x}_l^i$ $(0 \le l < k_i)$ can be represented by the likelihood probability function, which is defined by:

$$p_l^i(\boldsymbol{x}_l^i \mid G) = \sum_{s=1}^{S} \omega_s p_s(\boldsymbol{x}_l^i) \quad (1)$$

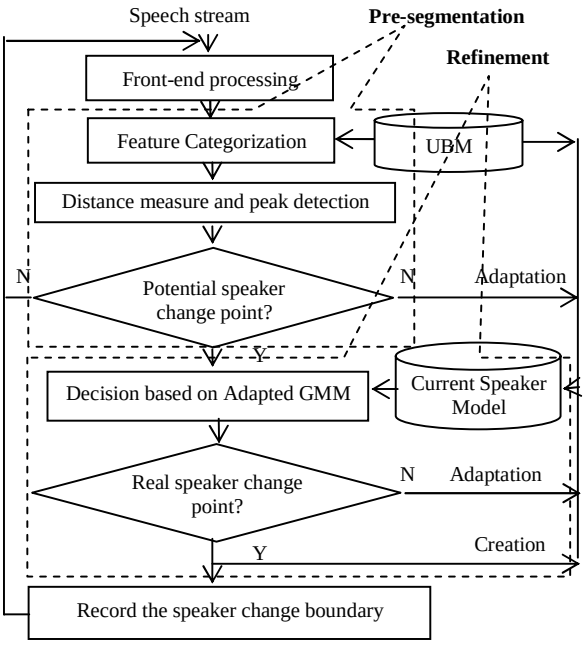

Figure 1. A brief flow diagram for speaker change detection

where

$$p_s(\boldsymbol{x}_l^i) = \frac{1}{(2\pi)^{\frac{N}{2}} \mid \sum_s \mid^{\frac{1}{2}}} \exp\left\{ -\frac{(\boldsymbol{x}_l^i - \boldsymbol{m}_s)^T \sum_s^{-1} (\boldsymbol{x}_l^i - \boldsymbol{m}_s)}{2} \right\}. \quad (2)$$

It can be assumed that the frames whose likelihood probabilities are relatively large have high confidences to represent the speaker characteristics, and the frames whose likelihood probabilities are relatively small are the relatively unreliable speaker-related vectors. Under this assumption, we categorize the feature vectors in one sub-segment into three categories, i.e. reliable speaker-related frames, doubtful speaker-related frames and unreliable speaker-related frames, according to their confidences. Due to no prior knowledge, we assume that frames in a sub-segment are uniformly distributed in our implementation [2].

Feature categorization based on UBM is not the same as traditional audio classification [6]. The aim of audio classification is usually to classify the audio segment into speech, music, silence, etc. Generally, it classifies the whole audio clip to the dominant audio type, often by a *hard decision*. Furthermore, in audio classification, although some segment is classified as speech, it may still contain some speech and noise-like frames simultaneously. In contrast, our feature categorization based on GMM-UBM here is to give a confidence to each frame and make a *soft decision*.

3.2. Distance Measure

Since we focus on the change of speaker identity in broadcast news but do not care about whether the background conditions change or not, we only consider the most reliable speaker-related frames categorized by feature categorization based on UBM and ignore the doubtful and unreliable frames which are more af-

fected by background or channel conditions. Thus, the dissimilarity D between two neighboring sub-segments is defined as the distance between their reliable speaker-related sets [2]:

$$D = d_{rs} . \tag{3}$$

Here, d_{rs} is the distance between two reliable speaker-related frame sets. It is defined as [1][5]:

$$d_{rs} = \frac{1}{2} tr[(\boldsymbol{C}_f - \boldsymbol{C}_l)(\boldsymbol{C}_f^{-1} - \boldsymbol{C}_l^{-1})], \tag{4}$$

where $\boldsymbol{C}_f$ and $\boldsymbol{C}_r$ represent the covariance of reliable speaker-related vectors in former sub-segment and latter sub-segment respectively. Eq. (3) means only the cluster of reliable speaker-related frames is considered to adapt the speaker model and measure the dissimilarity, and the effect of background and channel conditions is reduced to highlight the speaker characteristics.

3.3 Potential Speaker Change Detection

In speaker pre-segmentation, we compare the dissimilarity between every two adjacent sub-segments and detect the local peaks in the dissimilarity sequence, which are hypothesized to be the potential change points, if they satisfy the following conditions, as Fig.2 illustrates:

$$\begin{aligned} &D(i,i+1) - D^{i}_{\min,left} > \theta_i , \\ &D(i,i+1) - D(i+1,i+2) > 0 , \\ &D(i,i+1) > Th_i \end{aligned} \tag{5}$$

Here, $D(i, j)$ is the distance between *i-th* sub-segment and *j-th* sub-segment, $D^{i}_{\min,left}$ is the left minima around the *i-th* window. θ_i and Th_i are dynamic thresholds, which are computed as:

$$Th_i = \alpha_1 \cdot \frac{1}{M} \sum_{m=1}^{M} D(i-m-1, i-m) , \tag{6}$$

$$\theta_i = \alpha_2 \cdot \frac{1}{M} \sum_{m=1}^{M} D^{i-m}_{\min,left} , \tag{7}$$

where M is the number of previous distances used for predicting threshold. α_1 and α_2 are amplifiers pre-defined to 1.05 and 1.2 respectively after optimization in training set.

4. SPEAKER ADAPTATION IN REFINEMENT

Because there are no sufficient data to estimate the speaker model accurately from only one short speech sub-segment, such an estimated speaker model would be biased. Thus there are still some false alarms in the pre-segmentation. To solve this problem, we use as many data as possible to update the existing speaker model, and then a more accurate refinement method is also proposed to refine the pre-segment results.

Once the potential speaker change is detected in pre-segmentation stage, we will verify whether this potential change is a true speaker change or not through comparing the likelihood probability of the receiving sub-segment to existing speaker model with the dynamic threshold similarly as defined in Eq. (6). Therefore, it is important to obtain an accurate speaker model through speaker adaptation.

In order to get as many data as possible for estimating a speaker model more accurately, we utilize the results of the potential speaker change detection. If no potential 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 is updated using this available new data.

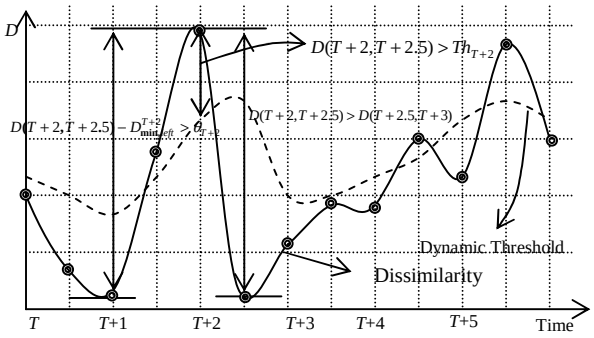

Figure 2. Illustration of finding potential speaker changes, which occur at (T+2) and (T+5.5)

Unlike the traditional speaker adaptation, because of the real-time constraint, we cannot store all former data that may need a large of space. Moreover, even if we can store all former data while taking no account of storage requirement, the adaptation using all former data will be time-consuming. Thus, incremental speaker adaptation (ISA) is proposed to deal with this problem.

Suppose at time t, the existing speaker adapted GMM is denoted as $G(\omega_s^{t-1}, \boldsymbol{m}_s^{t-1}, \Sigma_s^{t-1})$, and the training set received is denoted as $\boldsymbol{X}^t = (\boldsymbol{x}_0^t, \boldsymbol{x}_1^t, ..., \boldsymbol{x}_{k_t-1}^t)$. Thus, we can achieve a new set of parameters, which is denoted as $G(\hat{\omega}^t, \hat{\boldsymbol{m}}^t, \hat{\Sigma}^t)$, using training set $\boldsymbol{X}^t$ through EM algorithm adapted from GMM-UBM. If we denote the total number of frames at time t-1 as F_{t-1}, where $F_{t-1} = \sum_{j=1}^{t-1} k_j$, the weight and mean parameters of our new updated GMM are then combined by:

$$\omega_s^t = \frac{\omega_s^{t-1} F_{t-1} + \hat{\omega}_s^t k_t}{F_t} , \tag{8}$$

$$\boldsymbol{m}_s^t = \frac{\omega_s^{t-1} F_{t-1} \boldsymbol{m}_s^{t-1} + \hat{\omega}_s^t k_t \hat{\boldsymbol{m}}_s^t}{\omega_s^{t-1} F_{t-1} + \hat{\omega}_s^t k_t} . \tag{9}$$

Here we do not adapt the diagonal covariance in order to prohibit the components of GMM too sharp.

Obviously, adaptation in terms of Eq. (8) and (9) is less precise than traditional adaptation based on EM algorithm [3]. However, this approach reduces greatly the storage and its performance is comparable to the traditional speaker adaptation, which can be seen in our experiments.

5. EXPERIMENTAL RESULTS

In this section, we describe the database used to evaluate the proposed speaker segmentation algorithms. Then feature categorization based on GMM and K-Means are compared, while non-feature categorization is considered as a baseline. We also report the results of performance of ISA in speaker identification, as well as the pre-segmentation and refinement in our system.

5.1. Database

The evaluation of the proposed speaker change detection is performed on Hub-4 1997 English Broadcast News Speech Database. The database 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. In this database, each file is either about 30 minutes or 60 minutes, and there are about 30 speakers and about 60-80 speaker in these two kinds file, respectively.

5.2 Experimental Results

Twenty broadcasting news files, altogether about 10 hours, 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.

To investigate the performance of feature categorization based on UBM, we compare the false alarm rate (FAR) and missed detection rate (MDR) among non-feature categorization (no FC), feature categorization based on UBM-GMM (FC based on UBM) and K-Means (FC based on K-Means) in pre-segmentation respectively. The ROC curves of these three approaches are plotted in Figure 3. It shows that feature categorization based on UBM-GMM boosts the performance of speaker pre-segmentation considerably.

In order to evaluate the performance of proposed ISA method, fifty speakers who have plenty of speech data in training set are selected to form a speaker identification system. Each speaker model is trained by 30-second speech data. The remaining data are used as testing set. Figure 4 shows accuracy of speaker identification of two aforementioned approaches in different testing length. It illustrates that the performance of our ISA approach is just a little less than that of classical EM algorithm (about 3%). The loss is paying for the cost of reduction of the storage requirement and real-time processing.

In pre-segmentation, we can allow more FAR but prohibit less MDR, since we could correct many false potential changes in the refinement stage. Thus, in the speaker pre-segmentation module, FAR=33.8% and MDR=10.83% are obtained, while in the refinement stage, FAR=19.23% and MDR=13.65% are achieved. It can be also observed that in the refinement stage, we can decrease many false alarms while sacrificing few missing detection.

6. CONCLUSION

In this paper, we propose a two-stage approach for a real-time speaker segmentation system. A new feature categorization method is proposed to efficiently emphasize on the reliable speaker-related frames. Incremental speaker adaptation is also presented to suit real-time processing requirement. Experimental results demonstrate our approach improves the system performance in comparison with the existing system. Further work will be focused on the recognition of unknown speakers.

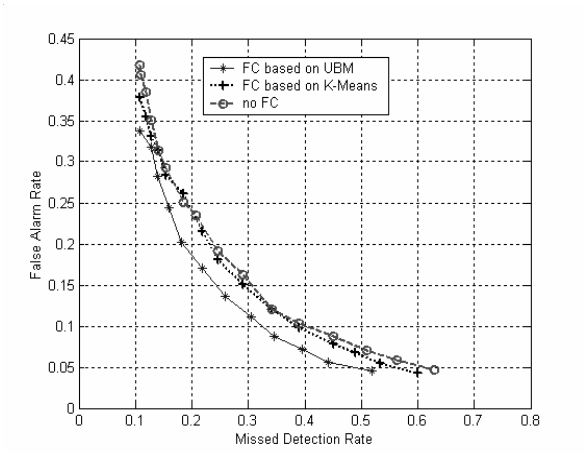

Figure 3. ROC curves of FC based on UBM, FC based on K-Means and no FC

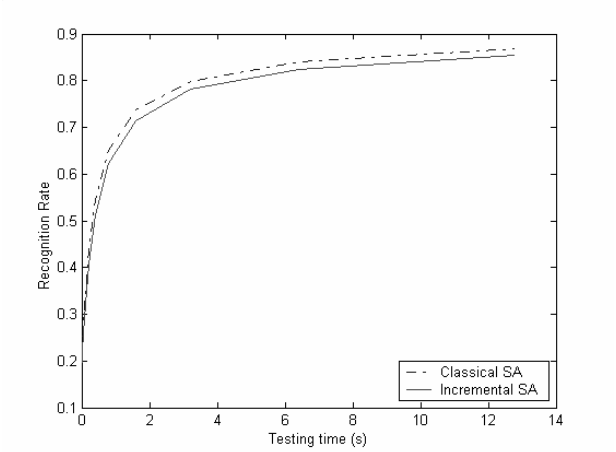

Figure 4. Recognition rates of traditional SA and ISA in different testing length

ACKNOWLEDGEMENT

This work was in part supported by an MSR grant on non-verbal speech analysis and an NSFC grant (60075017) to KC.

7. REFERENCES

[1] J. P. Campbell, JR. "Speaker Recognition: A Tutorial", *Proc. of the IEEE*, vl.85, No. 9 (1997), pp.1437-1462.

[2] T. Y. Wu, L. Lu, etc, "Universal Background Models for Real-time Speaker Change Detection", Proc of *9th International Conference on Multimedia Models (MMM2003)*, pp.135-149

[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] H. S. Beigi, Maes, S., "Speaker, Channel and Environment Change Detection", In: *World Congress of Automation* (1998).

[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] L. Lu, H. Jiang and H.-J. Zhang, "A Robust Audio Classification and Segmentation Method", *Proc. of ACM Multimedia* 2001, pp.203-211