Speaker Change Detection based on Mean Shift

Ji-chen Yang*, Qian-hua He, Yan-xiong Li, Xue-yuan Zhang
South China University of Technology / School of Electronic and Information Engineering
Guangzhou, China
*Email: NisonYoung@yahoo.cn

***Abstract*—To settle out the problem that search of speaker change point (SCP) is blind and exhaustive, mean shift is proposed to seek SCP by estimating the kernel density of speech stream in this paper. It contains three steps: seeking peak points using mean shift firstly, using maximum likelihood ratio (MLR) to compute the MLR value of the peak points secondly, and seeking SCPs from MLR value using the maximum method thirdly. The relationship of MLR and BIC is given then. Compared with those methods of using metric or model, the process of seeking SCP is no longer blind because mean shift always points the direction of maximum increase in the density. The experiments show that the proposed algorithm can arrive a comparable result against to BIC and DISTBIC, while it can save detection time, for a 3-second speech segment , the time using the proposed algorithm is about 60% of DISTBIC and 45% of BIC . Further investigation and improvement about this method is discussed at the end of this paper.**

***Index Terms*—Speaker change detection, mean shift, kernel density estimation, peak point, maximum likelihood ratio**

I. INTRODUCTION

Audio is an important source of information for content-based multimedia indexing and retrieval and it can sometimes be even more important than visual part as it shows a stable behavior according to the content, From the content-based multimedia point of view, the audio information can be even more important than the visual part as it is more mostly unique and significantly stable within the entire duration of the content [1].

Speaker change detection (SCD) is the process of determining the time indices of the points of speaker change in a given conversational audio stream, without prior acoustic information on the speaker, which has received a great deal of interest in recent yeas.SCD has a range of applications in different areas including speaker tracking, indexing audio recording, and proving cues for scene/topic/program change [2]. In a word, SCD is the base and the most important step for speaker retrieval.

There are mainly three major categories of SCD algorithms: model-based, metric-based, and hybrid ones [3]. In model-based algorithms, which often initializes a set of models for different acoustic classes from training corpus to classify the input audio stream so as to locate the change [4]. There are a set of models have been used, for example, universal background model (UBM), universal general model (UGM), sample speaker model (SSM), anchor model and hidden Markov models (HMM) [3].In metric-based algorithms, detecting the local extrema of a proper distance between neighboring windows by using various distance, for example, generalized likelihood ratio, Kullback-Leibler divergence. In hybrid model algorithm, it is usually to use metric-based technique firstly and model-based secondly.

From the development history of mean shift, we can see that mean shift was mainly used in image processing application and there are not any researchers using it in audio processing yet. In order to settle out the problem that search of SCP is blind and exhaustive, mean shift is proposed to seek SCP by estimating the kernel density of speech stream in this paper, which is the first contribution of this paper. Because mean shift always points toward the direction of maximum increase in the density, it can seek SCP directly and fast.

The second contribution of the paper is that maximum method is proposed to seeking speaker change points from peak points.

The third contribution of the paper is discussing the relationship between Bayesian Information Criterion and maximum likelihood ratio and giving the conclusion under the maximum method.

The remainder of the paper is organized as following. In Section II, mean shift is introduced. In Section III, and in Section IV, we evaluate the proposed approach by comparing it to conventional methods. Section V concludes the paper finally.

II. RELEATED WORK

There are usually two steps for those SCD methods [4-6], potential speaker change is detected by metric or model at the first step, then potential speaker change point (SCP) is judged whether it is a genius or a false point at the second step. Bayesian Information Criterion (BIC) [4, 6] has been the most dominant approach for speaker change criterion in recent years; its popularity is mainly due to its superior ability to judge speaker change.

For those methods, the search of SCP is blind and exhaustive. In [2], a speaker change is hypothesized at the midpoint of a window, which is a fixed-sized analysis window slid through the given audio at a predetermined rate. Potential SCP is detected by calculating the distance of two variable windows in a growing-window in [5]. This method is based on two hypothesis [9]: 1) the union

Manuscript received January 2, 2012
Ji-chen Yang is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong, China (E-mail: NisonYoung@ yahoo.cn).

doi:10.4304/jcp.8.3.638-644

of the feature vectors of the two windows forms a Gaussian cluster in the feature space, 2) the feature vectors of every window form a distinct Gaussian cluster. Two adjacent windows of the same size is shifted by a fixed step to detect potential SCP in [8]. A certain distance measure is used to evaluate the dissimilarity between the two windows. For this approach, generalized likelihood ratio and Kullback-Leibler divergence which derived from signal Gaussian model are popular because they have the advantage of low computational cost. The shortcoming of this method is to induce many potential speaker change points .In [9], speaker change point is detected by recursively partitioning a large analysis window into two sub-windows and recursively verifying the merging of two adjacent audio segments.

Mean shift was first proposed to estimate the gradient of a density by Fukunaga in 1975 [10]. Nonparametric density gradient estimation using a generalized kernel approach is investigated in [17]. The author pointed that mean shift is a hill-climbing algorithm on the density defined by a finite mixture or a kernel density estimate. By generalized the results for a Gaussian kernel function, a sampled mean shift estimate of the normalized gradient and extended it to a k-nearest-neighbor approach was also developed.

Cheng used it in mode seeking and clustering in 1995[11]. The author pointed that mean shift is a simple iterative procedure that shifts each data point to the average of data points in its neighborhood. It is shown that mean shift is a mode-seeking process on a surface constructed with a "shadow" kernel. For Gaussian kernels, mean shift is a gradient mapping .The two major applications of mean shift , which are cluster analysis and global optimization , were discussed .Because Mean shift has no fixed step or parametric to find mode, it has been successfully used in image processing, for example, image segmentation [12] . A general nonparametric technique is proposed for the analysis of a complex multimodal feature space and to delineate arbitrarily shaped clusters in [12], the authors proved for discrete data the convergence of a recursive mean shift procedure to the nearest stationary point of underlying density function and its utility in detecting the modes of the density. In 2007, Carreira-Perpinan proposed Mean shift is an EM algorithm [13]. The paper shown that, mean-shift is an expectation-maximization (EM) algorithm when the kernel is Gaussian and mean-shift is a generalized EM algorithm when the kernel is non-Gaussian. Which implies that mean-shift converges from almost any starting point and that, in general, its convergence is of linear order. Mean shift has widely used in object tracking [14-18]. Object tracking using mean shift plays an important role in computer vision applications because of its robustness implementation and computational efficiency. A new approach toward target representation and localization, the central component in visual tracking of non-rigid objects is proposed in [14], in which feature histogram-based target representations are regularized by spatial masking with an isotropic kernel .A metric derived from the Bhattacharyya coefficient as similarity measure is employed and mean shift procedure to perform the optimization is used. The method successfully coped with camera motion, partial occlusions, clutter, and target scale variations. In [15], a new mean-shift algorithm to tackle some tracking difficulties, such as background clutter and partial occlusion is proposed. First, all mean-shift-like tracking algorithms are compared and indicate that the main difference among them is weight calculation. Then, a new fusion strategy is proposed to unify all weight calculation methods into a framework. A fully automatic multiple-object tracker based on mean shift algorithm is presented in [16], Foreground is extracted using a mixture of Gaussian followed by shadow and noise removal to initialize the object trackers and also used as a kernel mask to make the system more efficient by decreasing the search area and the number of iterations to converge for the new location of the object. Trackers are automatically refreshed to solve the potential problems that may occur because of the changes in objects' size, shape, to handle occlusion-split between the tracked objects and to detect newly emerging objects as well as objects that leave the scene. Using a shadow removal method increases the tracking accuracy. In [17], a corrected background-weighted histogram (CBWH) formula is proposed by transforming only the target model but not the target candidate model in mean-shift tracking. The CBWH scheme can effectively reduce background's interference in target localization. The experimental results show that CBWH can lead to faster convergence and more accurate localization than the usual target presentation in mean-shift tracking. For the purpose of algorithmic speedup, an agglomerative mean shift (MS) clustering method along with its performance analysis is proposed in [18]. That method, namely Agglo-MS, is built upon an iterative query set compression mechanism which is motivated by the quadratic bounding optimization nature of MS algorithm. The whole framework can be efficiently implemented in linear running time complexity. In [19], in order to solve the problem of neglecting the fact that the background images consist of different objects whose conditions may change frequently in the current methods of performing background subtraction for detecting moving objects in videos. A novel hierarchical background model is proposed based on segmented background images by using mean shift and a hierarchical model.

III. INTRODUCTION OF MEAN SHIFT

Given a set $\{x_i\}_{i=1...n}$ of n points in the d dimensional space R^d, the estimation of kernel density f (x) with kernel $K(x)$ and window bandwidth h is:

$$\hat{f}_K(x_t) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{h}\right) w(x_i) \qquad (1)$$

where $\int K(x)dx = 1$, $w(x_i)$ is weight at x_i and $\sum_{i=1}^{n} w(x_i) = 1$.

Then we can get the gradient of $\hat{f}_K(x)$

$$\hat{\nabla} f_K(x) = \frac{2}{nh^{d+2}} \sum_{i=1}^{n} (x - x_i) k' \left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i) \qquad (2)$$

If we let $g(x) = -k'(x)$ and $G(x) = g\left(\|x\|^2\right)$, formula (2) can be written as:

$$\hat{\nabla} f_K(x) = \frac{2}{nh^{d+2}} \sum_{i=1}^{n} (x_i - x) G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)$$

$$= \frac{2}{h^2} \left[\frac{\sum_{i=1}^{n} G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)}{nh^d} \right] \times$$

$$\left[\frac{\sum_{i=1}^{n} (x_i - x) G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)}{\sum_{i=1}^{n} G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)} \right]$$

$$= \frac{2}{h^2} \left[\frac{\sum_{i=1}^{n} G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x)}{nh^d} \right] \times$$

$$\left[\frac{\sum_{i=1}^{n} x_i G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)}{\sum_{i=1}^{n} G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)} - x \right] \qquad (3)$$

The formula (3) is composed of two parts. The first part is the estimation of kernel density f (x) with kernel $G(x)$, the second part is mean shift vector, if we define it as $M_h(x)$.

$$M_h(x) = \frac{\sum_{i=1}^{n} x_i G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)}{\sum_{i=1}^{n} G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)} - x \qquad (4)$$

So we can get

$$\nabla \hat{f}_K(x) = \frac{2}{h^2} \hat{f}_G(x) M_h(x) \qquad (5)$$

From (5), we can further get

$$M_h(x) = \frac{1}{2} h^2 \frac{\nabla \hat{f}_K(x)}{\hat{f}_G(x)} \qquad (6)$$

We can know that mean shift vector $M_h(x)$ computed with kernel $G(x)$ is proportional to the density gradient estimate obtained with kernel $K(x)$. Thus, mean shift vector always points toward the direction of maximum increase in the density.

If we let the first part as $m_h(x)$ in formula (4)

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)}{\sum_{i=1}^{n} G\left(\left\| \frac{x_i - x}{h} \right\|^2 \right) w(x_i)} \qquad (7)$$

$m_h(x)$ can be used as a recursive formula to find the mode. In speech stream, the density of SCP is more than its neighbor field, so we can use mean shift to detect the peak points.

IV. Speaker Change Detection based on Mean Shift

In this section, the algorithm of speaker change detection based on mean shift is proposed.

It contains three stages:

1. Seeking peak point using mean shift.
2. Using maximum likelihood ratio (MLR) to compute the MLR value of the peak points.
3. Seeking SCPs from MLR value using the maximum method.

The first stage is the more important and the base of the latter two stages. Next we will introduce them step by step.

A. Seeking Peak Point

Because the kernel density of SCP is more than its neighbor field, that is to say, the position of SCP is a peak point, so we can use formula (7) to seek the peak point.

In conversation methods, in order to seek peak point, some distance formulas are often used to compute the distance of two variable windows in a growing-window [5] or two adjacent windows of the same size shifted by a fixed step [8] first, and then seeking the peak points by some rules from the points which are gotten from the former step. We can see that the process of seeking peak point is blind and exhaustive.

Because mean shift always points toward the direction of maximum increase in the density and it will reach the maximum point finally, so the process of seeking peak point using mean shift is no longer blind and exhaustive. That mean shift doesn't have fixed step or parameters is another advantage.

In order to seek the peak point using formula (7), kernel function $G(x)$ is set as $e^{-\frac{x^2}{2}}$, window bandwidth h is set as 4s and all the weights are set as the same.

The basic idea of seeking peak point is as follows: randomly selection a frame x in window ; using formula (7) to get $m_h(x)$, if the result meets convergence condition, that means seeking peak point, else using

formula (7) as a recursive formula continue; seeking other peak points according the foregoing steps gradually.

The process of seeking peak point is as following:

(1) Initialize a=3, b=10.
(2) Randomly selection a frame as x in interval [a, b]
(3) Using formula (7) to calculate $m_h(x)$.

Let the frame corresponding to $m_h(x)$ be x_R

(4) If $\|m_h(x)-x\|$ >TH (TH is a threshold),

Let x = $m_h(x)$, go to (3)

Else x_R is a peak point;

Let a= x_R +1+a, b= x_R +1+b , go to (2)

Figure 1 displays the process of seeking peak point using mean shift.

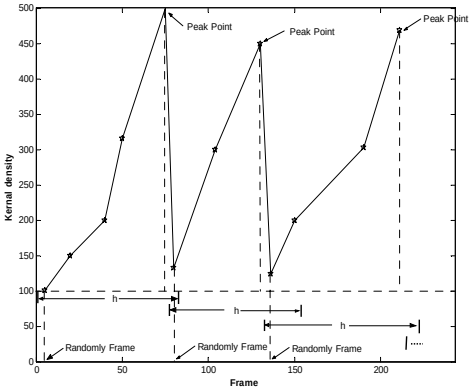

Figure 1. The process of seeking peak point using mean shift

B. Calculation MLR Value of Peak Points

From last stage, we can know that peak point is only a maximum value in a window, not all peak points are SCPs , in order to seek real SCPs, in this stage, we use MLR (formula 8) to get MLR value of every peak point.

$$d_{MLR} = N\log|C| - N_1\log|C_1| - N_2\log|C_2| \quad (8)$$

where N_1 and N_2 denote the frame number of left and right 3s speech segments of peak point respectively and N denotes the number of the merged speech segments; C_1, C_2 and C stand for covariance of them respectively.

Supposing the peak points are p_i (i=1, 2, 3, …n), using formula (8), we can get the MLR value M (p_i) (i=1, 2, 3, …n) of p_i (i=1, 2, 3, …n).

C. Seeking SCPs From MLR Value

We all know that M (p_i) (i=1, 2, 3, …n) stands for the relationship between left and right speech segment of peak point p_i (i=1, 2, 3, …n) . The more M (p_i), the more possible SCP is .Therefore, in this stage; we use maximum method to seek SCP, which runs as follows:

(1) Calculation the mean value u of all M (p_i), discarding the points those M (p_i) is lower than u.

(2) If the global maximum M (pi) of the remainder M (p_i) satisfies the condition: max>1.2u. Then the maximum M (p_i) corresponding point is regard as a genuine SCP, else there is no SCP in the remainder M (pi).

(3) Discarding the point whose position satisfy the condition: $|P_{max} - P| < 2s$, where P_{max} stands for the position of the maximum M (p_i) corresponding and P stands for the position of M (p_i).

(4) go to (2).

D. Relationship of MLR and BIC

The formula of BIC is as following.

$$d_{BIC} = N\log|C| - N_1\log|C_1| - N_2\log|C_2| - \frac{(d^2+3d)\lambda}{4}\log N \quad (9)$$

where λ is penalty factor, d is dimension of feature.

Though BIC has been the most dominant approach for speaker change criterion in recent years, the trouble of using BIC is to tune penalty factor λ, whose choice is task-dependent and non-robust to different acoustic and environment conditions [21]. That is to say, in order to let the d_{BIC} of SCP great than zero and d_{BIC} of false SCP less than zero, we have to tune penalty factor λ repeatedly. It is a disadvantage for BIC.

From formula (8) and (9) , we can get

$$d_{BIC} = d_{MLR} - \frac{(d^2+3d)\lambda}{4}\log N \quad (10)$$

If we let the length of speech segment equal the same every time and λ be a constant, the value of $\frac{(d^2+3d)\lambda}{4}\log N$ is also a constant. That is to say, the differences of d_{MLR} and d_{BIC} is a constant, too. So we can see that the effort of d_{MLR} and d_{BIC} is the same here.

Because the maximum method is seeking SCP by selecting maximum value of M (p_i) (i=1, 2, 3, …n) every time. Compared with BIC, it is no need to tune penalty factor. But their effort is the same under the maximum method.

V. Experiments and Evaluations

A. Data and Preprocessing

The data used in the experiments were from VOA (voice of America) broadcast news. All the data were sampled at 16 KHz, 16bits and saved as mono channel wav formats. There were 2.5 hours length; we divided it into two parts in our experiments: 1 hour was used to train and the other was used to test our algorithm.

In the front-end process, a speech activity detector is used to discard silence/noise, and then speech is firstly divided into frames of 32ms with 50% overlap and 12

mel-frequency cepstral coefficients (MFCCs) are extracted for each frame.

B. Selection N to Form Length-frame

From formula (7) we can seek that $m_h(x)$ is decided by the data not only from neighbor field of frame x but also frame x. Because the information in a frame is little, so we can use n frames to form a length-frame with more information.

In order to find the best value of n, firstly, we use n frame to form length-frame, so that new combined MFCCs can be gotten according to length-frame, then experiments are done to study the relationship between peak-speaker-ratio (its definition is as formula (11)) and n. The result is as Figure 2.

$$\text{peak-speaker-ratio} = \frac{n_{PS}}{n_{TS}} \tag{11}$$

where n_{PS} represents the number of speaker change points in peak points, n_{TS} is the total number of speaker change points.

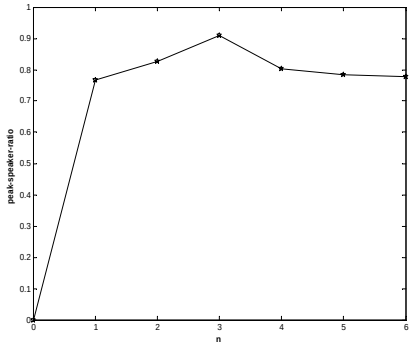

Figure 2. The relationship between peak-speaker-ratio and n

From figure 2, we can see that, when n increases from 1 to 3, the peak-speaker-ratio is increasing and it reaches the peak when n equals 3. While n increases continue, the peak-speaker-ratio is decreasing; the peak-speaker-ratio is nearly the same when n equals 6 and 1. So we can know that n equals 3 is the best value and we let n be 3 to form length-frame.

C. Figures of Merit

The performances of the proposed approach was evaluated by precision, recall and F_1. Their definition is as following:

$$\text{precision} = \frac{n_{rs}}{n_s}$$

$$\text{recall} = \frac{n_{rs}}{n_r}$$

$$F_1 = 2\frac{n_{rs}}{n_s + n_r} \tag{12}$$

where n_{rs} denotes the number of correctly detected speaker change points, n_s represents the number of all detected speaker change points, n_r is the total number of speaker change points.

D. Comparison of Distance Formulas

After all the peak points were gotten, in order to discard false SCPs and get real SCPs, we calculate maximum likelihood ratio (MLR) value of peak points, we also compute Kullback-Leibler divergence (KLD) [22] and arithmetic harmonic sphericity (AHS)[23] of peak points to compare. The speaker change detection result is as listed in Table I.

$$d_{KLD} = \frac{1}{2}tr[(C_1 - C_2)(C_1^{-1} - C_2^{-2})] \tag{13}$$

$$d_{AHS} = \log[tr(C_1C_2^{-1}) \times tr(C_2C_1^{-1})] - 2\log(D) \tag{14}$$

where C_1 and C_2 stand for covariance of the speech segments at the side of peak point respectively, C_1^{-1} and C_2^{-1} are their inverse matrix. $tr[\bullet]$ represents the trace of the matrix and D stands for the dimension of the feature vector.

TABLE I.
THE SPEAKER CHANGE DETECTION RESULT USING DIFFERENT DISTANCE FORMULAS

	precision/%	recall/%	F_1/%
KLD	42.48	53.04	47.18
AHS	44.03	65.19	52.56
MLR	74.86	75.69	75.27

From Table I, it is observed that using MLR is much better than KLD and AHS. From formula (8), (13) and (14), we can see that the reason may be as follows: MLR not only considers the left and right data of peak point but also the merged data, while KLD and AHS only consider the left and the right data .

E. Comparison with Some Conventional Algorithms

In order to evaluate our algorithm, we use conversation speaker change detection algorithm, e.g. BIC [5] and DISTBIC [8] to do experiment. The comparison result is as Table II.

TABLE II.
COMPARISON WITH SOME CONVERSATION ALGORITHM

	precision/%	recall/%	F_1/%
BIC	61.37	93.92	74.23
DISTBIC	86.75	71.27	76.79
Ours	74.86	75.69	75.27

From Table II, it can be seen that the proposed algorithm can arrive a comparable result against to BIC and DISTBIC. In BIC and DISTBIC, the process of seeking potential SCPs is blind and exhaustive by calculating the distance of adjacent windows, while ours can escape the trouble.

F. Comparison of Detection Time

Because the proposed algorithm can seek SCP directly and fast, so it can save detection time. Figure 3 displays the detection time using different algorithm for different length speech segments.

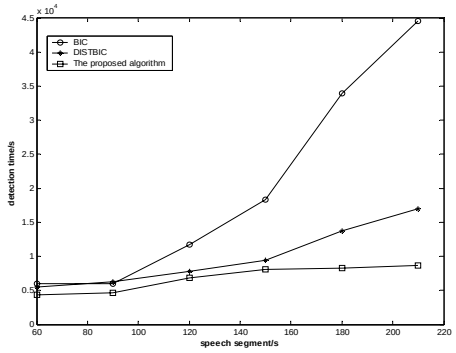

Figure.3. Detection time using different algorithm for different length speech segments

From figure 3, it is observed that the proposed algorithm can save detection time. When the length of speech segment becomes longer, the detection time is increasing, but ours increases slowly, DISTBIC increases faster than ours while BIC increases fastest. For a 3-second speech segment, the time using ours is about 60% of DISTBIC and 45% of BIC.

VI. CONCLUSION AND DISCUSSION

In this paper, the algorithm of speaker change detection based on mean shift was proposed, which contain three stages: firstly, peak point was gotten by estimating the kernel density using mean shift ; secondly , MLR value of peak point was gotten by using maximum likelihood ratio ; thirdly, seeking real speaker change point from MLR value. Compared with BIC and DISTBIC, the proposed algorithm can arrive at a comparable result against them and save detection time.

While the results of speaker change detection reported in this paper are promising, the proposed algorithm still leaves considerable room for further investigation and improvement. For example, in the recursive process of seeking the peak point, window bandwidth h is fixed as 4s in our algorithm, there will be more peak point for speech stream without speaker changing long time, the best solution is to use variable window bandwidth; because the position of $m_h(x)$ is not precisely, so the weight is difficult to get, we have to let all the weight be the same in our experiment. In our future work, in order to get better speaker change detection result using mean shift, variable window bandwidth and proper weight is the key point to study, which are good directions in the future work.

ACKNOWLEDGMENT

This work was supported National Natural Science Foundation of China (NSFC) (Item No. 61101160, 60972132) and Natural Science Foundation of Guangdong province, China (No. 9351064101000003, No. 10451064101004651), and the Fundamental Research Funds for the Central Universities, South China University of Technology, China (No. 2011ZM0029).

REFERENCES

[1] S.Kiranyaz, M.Gabbouj, "Generic content-based audio indexing and retrieval framework", *IEE Proc.-Vis Image Siganl*, vol.153, no.3, pp.285-297, 2006.

[2] Amit S. Malegaonkar, Aladdin M.Ariyaeeinia, and Perasiriyan Sivakumaran , "Efficient speaker change detection using adapted Gaussian mixture models", *IEEE Trans. on Audio, speech and language processing*, vol .15, no.6, pp.1859-1869, 2007.

[3] Margarita Kotti, Emmanoui Benetos, Constantine Kortropoulos, "Computainally efficient and robust BIC-based speaker segmentation", *IEEE Trans. on Audio, speech and language processing*, vol.16, no.5, pp.920-933, 2008.

[4] L.Lu, H.Jiang and H.Zhang, "A robust audio classification and segmentation method", *ACM Multimedia*, pp.203-211, 2001.

[5] Scott shaobin chen, P.S.Gopalakrishnan, "Speaker , environment and channel change detection and clustering via the Bayesian information criterion" , *DARPA Broadcast News Transcription & Understanding Workshop*, 1998, pp.127-132.

[6] Sunil kumar kopparapu, Ahmed Imran, G Sita, "A two pass algorithm for speaker change detection", pp.755-758, *IEEE TENCON 2010*.

[7] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE J. Sel. Topics Signal Process*. vol. 4, no. 6, pp. 1059–1070, 2010.

[8] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segment for audio data indexing", Journal *of Speech Communication*, vol.32, pp. 111-126, 2000.

[9] Shih-Sian Cheng, Hsin-Min Wang, Hsin-Chia Fu, "BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization", *IEEE Trans. on Audio, speech and language processing*, vol .18, no.1, pp.141-157, 2012.

[10] K. Fukunaga and L.D. Hostetler, "The estimation of the gradient of a density function, with application in pattern recognition", *IEEE Trans. Information Theory*, vol.21, pp.32-40, 1975.

[11] Yizong Cheng, "Mean shift, Mode seeking and clustering", *IEEE Trans. on Pattern analysis and machine intelligence*, vol.17, no.8, pp.790-799, 1995.

[12] D. Cormaniciu, P. Meer, "Mean Shift: A robust approach toward feature space analysis", *IEEE Trans. On Pattern analysis and machine intelligence*, vol.24, no.5, pp. 603-619, 2002.

[13] M. Carreira-Perpinan, "Gaussian mean-shift is a EM algorithm", *IEEE Trans. On Pattern analysis and machine intelligence*, vol. 29, no.5, pp.767-776, 2007.

[14] D. Cormaniciu, V. Ramesh and P. Meer, "Kernel-based object tracking", *IEEE Trans on Pattern analysis and machine intelligence*, vol.25, no.5, pp.564-575, 2003.

[15] M. Carreira-Perpinan, "Gaussian mean-shift is a EM algorithm", *IEEE Trans. On Pattern analysis and machine intelligence*, vol. 29, no.5, pp.767-776, 2007.

[16] Lingfeng Wang, Chunhong Pan, Shiming Xiang, "Mean-shift tracking algorithm with weight fusion strategy", *18th IEEE International Conference on Image Processing*, pp.473-476, 2011.

[17] C. Beyan A. Temizel, "Adaptive mean-shift for automated multi object tracking", *IET Computer Vision*, vol.6, iss.1, pp.1-12, 2012.

[18] J. Ning, L. Zhang, D. Zhang, C. Wu1, "Robust mean-shift tracking with corrected background-weighted histogram", *IET Computer Vision*, vol.6, iss.1, pp.62-69, 2012.

[19] Xiao-Tong Yuan, Bao-Gang Hu, Ran He, "Agglomerative Mean-Shift Clustering", *IEEE Trans. Knowledge and data engineering*, vol. 24, no. 2, pp.209-219, 2012.
[20] Shengyong Chen, Jianhua Zhang, Youfu Li, Jianwei Zhang, "A Hierarchical Model Incorporating Segmented Regions and Pixel Descriptors for Video Background Subtraction", *IEEE Trans. on industrial informatics*, vol.8, no.1, pp.118-127, 2012.
[21] Margarita K., Vassiliki M., Constanting K., "Speaker segmentation and clustering", *Signal Processing*, vol.88, pp.1091-1124, 2008.
[22] Lie Lu, Hong-jiang, Zhang , and Hao Jiang, . Content analysis for audio classification and segmentation [J], *IEEE Transactions on Speech and Audio Processing*, vol.10, no.7, pp.504-516, 2002.
[23] Bimbot, F. Mathan L., Text-free speaker recognition using an arithmetic harmonic sphericity measure. In: proceedings of Europspeech, 1993, pp.169-172.

Ji-chen Yang was born in Jieshou, Anhui province, China in June, 1980. He received the B.Eng. degree in electronic and information engineering from Guangdong University of Petrochemical Technology (GDUPT), Maoming, Guangdong province, China, in 2004, he received the M. Eng. Degree in system engineering from Guangdong University of Technology (GDUT), Guangzhou, Guangdong, China, in 2007.he received the Ph.D. in Communication and Information System from South China University of Technology (SCUT), Guangzhou, Guangdong, China, in 2010.

Since Oct. 2011, he has been a post doc. researcher in South China University of Technology. His currents interest is speaker indexing and retrieval.

Qian-hua He was born in Shaodong, Hunan province, China, in Feb. 1965.He received the B. S. Degree in physics from Hunan Normal University in 1987, the M. S. Degree in medical instrument engineering from Xi'an Jiaotong University in 1990, and the Ph. D degree in communication engineering from South China University of Technology, in 1993.

Since 1993, he has been at the Institute of Radio and Auto-control of South China University of Technology. His research interests include speech recognition, speaker recognition and its security, optimal algorithm design, such as genetic algorithm and neural networks, embedded system design.

From 1994 to 2001, he worked with the department of computer science, City University of Hong Kong for about 3 years in 4 periods. From 2007.11 to 2008.10, he worked with University of Washington in Seattle as a visiting scholar.

Yan-xiong Li was born in Jiahe, Hunan province, China, in Aug. 1980. He received the B.S. Degree and the M.S. Degree both in Electronic Engineering from Hunan Normal University (HNU) in 2003 and 2006, respectively, and the Ph.D. Degree in Communication and Information System from South China University of Technology in 2009.

From 2 September 2008 to 1 September 2009, he worked as a research associate with the Department of Computer Science at the City University of Hong Kong. From 23 March 2010 to present, he has been working as a lecturer with the School of Electronic and Information Engineering at South China University of Technology. His current research interests include speech/audio signal processing, pattern recognition.

He has two research grants from the National Natural Science Foundation of China and from the Natural Science Foundation of Guangdong Province, China.

Xue-yuan Zhang was born in Shijiazhuang, Hebei province, China, in Nov.5 1987. He received the B.S. Degree in Information Engineering from South China University of Technology in 2009. From 2009 to present, he is pursuing his PhD degree in Information and Communication Engineering at South China University of Technology. His current research interests include speech/audio signal processing, pattern recognition.