SPEAKER, ENVIRONMENT AND CHANNEL CHANGE DETECTION AND CLUSTERING VIA THE BAYESIAN INFORMATION CRITERION

Scott Shaobing Chen & P.S. Gopalakrishnan
IBM T.J. Watson Research Center
email: schen@watson.ibm.com

ABSTRACT

In this paper, we are interested in detecting changes in speaker identity, environmental condition and channel condition; we call this the problem of *acoustic change detection*. The input audio stream can be modeled as a Gaussian process in the cepstral space. We present a maximum likelihood approach to detect turns of a Gaussian process; the decision of a turn is based on the *Bayesian Information Criterion* (BIC), a model selection criterion well-known in the statistics literature. The BIC criterion can also be applied as a termination criterion in hierarchical methods for clustering of audio segments: two nodes can be merged only if the merging increases the BIC value. Our experiments on the Hub4 1996 and 1997 evaluation data show that our segmentation algorithm can successfully detect acoustic changes; our clustering algorithm can produce clusters with high purity, leading to improvements in accuracy through unsupervised adaptation as much as the ideal clustering by the true speaker identities.

1. INTRODUCTION

Automatic segmentation of an audio stream and automatic clustering of audio segments according to speaker identities, environmental conditions and channel conditions have received quite a bit of attention recently [4, 8, 6, 10]. For example, in the task of automatic transcription of broadcast news [3], the data contains clean speech, telephone speech, music segments, speech corrupted by music or noise, etc. There are no explicit cues for the changes in speaker identity, environment condition and channel condition. Also the same speaker may appear multiple times in the data. In order to transcribe the speech content in audio streams of this nature,

- we would like to *segment* the audio stream into homogeneous regions according to speaker identity, environmental condition and channel condition so that regions of different nature can be handled differently: for example, regions of pure music and noise can be rejected; also, one might design a separate recognition system for telephone speech.
- we would like to *cluster* speech segments into homogeneous clusters according to speaker identity, environment and channel; unsupervised adaptation can then be performed on each cluster. [8, 10] showed that a good clustering procedure can greatly improve the performance of unsupervised adaptation such as MLLR.

Various segmentation algorithms have been proposed in the literation [2, 4, 6, 8, 10, 14], which can be categorized as follows:

- Decoder-guided segmentation. The input audio stream can be first decoded; then the desired segments can be produced by cutting the input at the silence locations generated from the decoder [14, 8]. Other informations from the decoder, such as the gender information, could also be utilized in the segmentation [8].
- Model-based segmentation. [2] proposed to build different models, e.g. Gaussian mixture models, for a fixed set of acoustic classes, such as telephone speech, pure music, etc, from a training corpus; the incoming audio stream can be classified by maximum likelihood selection over a sliding window; segmentation can be made at the locations where there is a change in the acoustic class.
- Metric-based segmentation. [4, 6, 10] proposed to segment the audio stream at maxima of the distances between neighboring windows placed at every sample; distances such as the KL distance, the generalized likelihood ratio distance have been investigated.

In our opinion, these methods are not very successful in detection the acoustic changes present in the data. The decoder-guided segmentation only places boundaries at silence locations, which in general has no direct connection with the acoustic changes in the data. Both the model-based segmentation and the metric-based segmentation rely on thresholding of measurements which lack stability and robustness. Besides, the model-based segmentation does not generalize to unseen acoustic conditions.

Clustering of audio segments is often performed via hierarchical clustering [10, 8]. First, a distance matrix is computed; the common practice is to model each audio segment as one Gaussian in the cepstral space and to use the KL distance or the generalized likelihood ratio as the distance measure [6]. Then bottom-up hierarchical clustering can be performed to generate a clustering tree. It is often difficult to determine the number of clusters. One can heuristically pre-determine the number of clusters or the minimum size of each cluster; accordingly, one can go down the tree to obtain desired clustering [14]. Another heuristic solution is to threshold the distance measures during the hierarchical process; the thresholding level is tuned on a training set [10]. Jin et al. [7] shed some light on automatically choosing a clustering solution.

In this paper, we are interested in detecting changes in speaker identity, environmental condition and channel condition; we call this the problem of *acoustic change detection*. The input audio stream can be modeled as a Gaussian process in the cepstral space. We present a maximum likelihood approach to detect turns of a Gaussian process; the decision of a turn is based on the *Bayesian Information Criterion* (BIC), a model selection criterion in the statistics literature. The BIC criterion can also be applied as a termination criterion in the hierarchical methods for speaker clustering: two nodes can be merged only if the merging increases the BIC value. Our experiments on the Hub4 1996 and 1997 evaluation data show that our segmentation algorithm can successfully detect acoustic changes; our clustering algorithm can produce clusters with high purity and enhance unsupervised adaptation as much as the ideal clustering by the true speaker identities.

This paper is organized as follows: section 2 describes model selection criterions in the statistics literature; section 3 and section 4 explains our maximum likelihood approach for acoustic change detection and our clustering algorithm based on BIC; we present our experiments on the Hub4 1996 and 1997 evaluation data; we compare our algorithms with other recent works in the literature.

2. MODEL SELECTION CRITERIA

The problem of model identification is to choose one among a set of candidate models to describe a given data set. We often have candidates of a series of models with different number of parameters. It is evident that when the number of parameters in the model is increased, the likelihood of the training data is also increased; however, when the number of parameters is too large, this might cause the problem of *overtraining*. Several criteria for model selection have been introduced in the statistics literature, ranging from non-parametric methods such as cross-validation, to parametric methods such as the Bayesian Information Criterion (BIC) [11].

BIC is a likelihood criterion penalized by the model complexity: the number of parameters in the model. In detail, let $\mathcal{X} = \{x_i : i = 1, \cdots, N\}$ be the data set we are modeling; let $\mathcal{M} = \{M_i : i = 1, \cdots, K\}$ be the candidates of desired parametric models. Assuming we maximize the likelihood function separately for each model M, obtaining, say $L(\mathcal{X}, M)$. Denote $\#(M)$ as the number of parameters in the model M. The BIC criterion is defined as:

$$BIC(M) = \log L(\mathcal{X}, M) - \lambda \frac{1}{2} \#(M) \times \log(N) \qquad (1)$$

where the penalty weight $\lambda = 1$. The BIC procedure is to choose the model for which the BIC criterion is maximized. This procedure can be derived as a large-sample version of Bayes procedures for the case of independent, identically distributed observations and linear models [11].

The BIC criterion is well-known in the statistics literature; it has been widely used for model identification in statistical modeling, time series [13], linear regression [5], etc. It is commonly known in the engineering literature as the *minimum description length* (MDL). It has been used in the speech recognition literature, e.g. for speaker adaptation [12]. BIC is closely related to other penalized likelihood criterions such as AIC [1] and RIC [5]. One can vary the penalty weight λ in (1), although only $\lambda = 1$ corresponds to the definition of BIC.

3. CHANGE DETECTION VIA BIC

In this section, we describe a maximum likelihood approach for acoustic change detection based on the BIC criterion. Denote $\mathbf{x} = \{x_i \in \mathcal{R}^d, i = 1, ..., N\}$ as the sequence of cepstral vectors exacted from the entire audio stream; assume $\mathbf{x}$ is drawn from an independent multivariate Gaussian process:

$$x_i \sim N(\mu_i, \Sigma_i)$$

where μ_i is the mean vector and Σ_i is the full covariance matrix.

3.1. Detecting One Changing Point

We first examine a simplified problem: assume that there is at most one changing point in the Gaussian process.

We are interested in the hypothesis testing of a change occurring at time i:

$$H_0 : x_1 \cdots x_N \sim N(\mu, \Sigma)$$

versus

$$H_1 : x_1 \cdots x_i \sim N(\mu_1, \Sigma_1); x_{i+1} \cdots x_N \sim N(\mu_2, \Sigma_2).$$

The maximum likelihood ratio statistics is

$$R(i) = N log|\Sigma| - N_1 log|\Sigma_1| - N_2 log|\Sigma_2| \qquad (2)$$

where Σ, Σ_1 and Σ_2 are the sample covariance matrices from all the data, from $\{x_1, \cdots, x_i\}$ and from $\{x_{i+1}, \cdots, x_N\}$, respectively. Thus the maximum likelihood estimate of the changing point is

$$\hat{t} = \arg\max_i R(i).$$

On the other hand, we can view the hypothesis testing as a problem of model selection. We are comparing two models: one models the data as two Gaussians; the other models the data as just one Gaussian. The difference between the BIC values of these two models can be expressed as

$$BIC(i) = R(i) - \lambda P \qquad (3)$$

where the likelihood ratio $R(i)$ is defined in (2), the penalty

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N$$

and the penalty weight $\lambda = 1$; d is the dimension of the space. Thus if (3) is positive, the model of two Gaussians is favored. Thus we decide there is a change if

$$\{\max_i BIC(i)\} > 0. \qquad (4)$$

It is clear that the m.l.e. of the changing point also can be expressed as

$$\hat{t} = \arg\max_i BIC(i). \qquad (5)$$

Comparing with the metric-based segmentation described in the introduction, our BIC procedure has the following advantages:

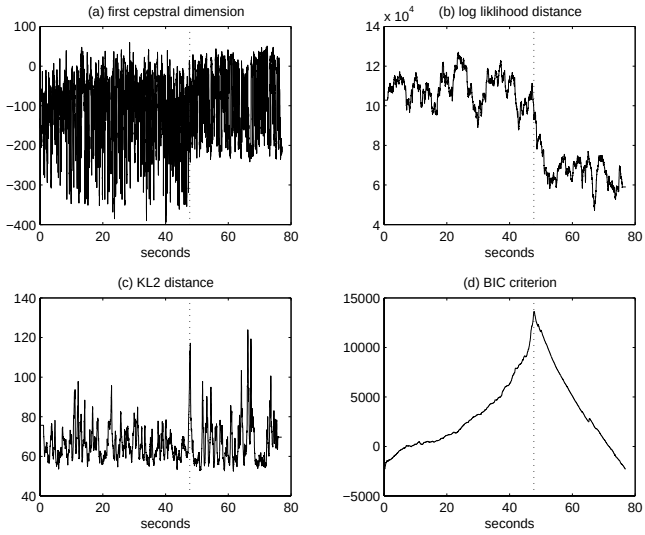

Figure 1. Detecting one changing point

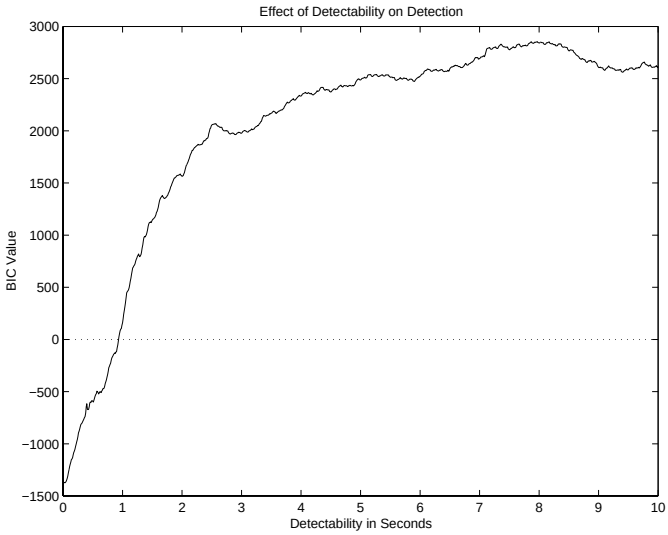

Figure 2. The detectability of a change

- *Robustness.* [10, 4] proposed to measure the variation at location i as the distance between a window to the left and a window to the right; typically the window size is short, e.g. two seconds; the distance can be chosen to be the log likelihood ratio distance [6] or the KL distance. In our opinion, such measurements are often noisy and not robust, because it involves only the limited samples in two short windows. In contrast, the BIC criterion is rather robust, since it computes the variation at time i utilizing all the samples. Figure 1 shows an example which indicates the robustness of our procedure. We experimented on a speech signal of 77 seconds which contains two speakers. Panel (a) plots the first dimension of the cepstral vectors; the dotted line indicates the location of the change. One can clearly notice the changing behavior around the changing point. We computed both the log likelihood ratio distance (i.e. the Gish distance) and the KL2 distance [10] between two adjacent sliding windows of 100 frames. Panel (b) shows the log likelihood distance: it attains local maximum at the location of the change; however, it has several maxima which do not correspond to any changing points; it also seems rather noisy. Similarly Panel (c) shows the KL2 distances: there is a sharp spike at the location of the change; however, there are several other spikes which do not correspond to any changing points. Panel (d) displays the BIC criterion; it clearly predicts the changing point.
- *Thresholding-free.* Our BIC procedure is able to automatically performs model selection, whereas [10] is based on thresholding. As shown in Figure 1 (b) and (c), it is difficult to set a thresholding level to pick the changing points. Figure 1(d) indicates there is a change since the BIC value at the detected changing point is positive.
- *Optimality.* Our procedure is derived from the theory of maximum likelihood and model selection. It can be shown that our estimate (5) converges to the true changing point as the sample size increases.

The performance of our procedure relies heavily on the amount of data available for each of the two Gaussian models separated by the true changing point. We define the *detectability* of a changing point at t as

$$D(t) = \min(t, N - t). \tag{6}$$

In general the BIC procedure is less accurate as the detectability decreases. This can be demonstrated in the following experiment. We placed multiple windows of the same size around a speaker changing point in an audio stream, with each window corresponding to different detectability. Within each window, the BIC procedure was performed to detect if there was a change. Figure 2 plots the BIC value against the detectability of the sampling. The BIC value starts as negative, suggesting that there is only one speaker. As the detectability increases, the BIC value also increases sharply; it is well above zero for detectability greater than 2 seconds, strongly supporting the change point hypothesis.

3.2. Detecting Multiple Changing Points

We propose the following algorithm to sequentially detect the changing points in the Gaussian process $\mathbf{x}$:

(1) initialize the interval $[a, b] : a = 1; b = 2$.
(2) detect if there is one changing point in $[a, b]$ via BIC.
(3) if (no change in $[a, b]$)
    let $b = b + 1$;
  else
    let $\hat{t}$ be the changing point detected;
    set $a = \hat{t} + 1$; $b = a + 1$;
  end
(4) go to (2).

By expanding the window $[a, b]$, the final decision of a change point is made based on as much data points as possible. In our view, this can be more robust than decisions based on distance between two adjacent sliding windows of fixed sizes [10], though our approach is more costly.

The BIC criterion can be viewed as thresholding the log likelihood distance, with the thresholding level automatically chosen as $\lambda \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N$ where N is the size of the decision window and d is the the dimension of the feature space.

Again we emphasize that the accuracy of our procedure depends on the detectabilities of the true changing points. Let $\mathcal{T} = \{t_i\}$ be the true changing points; the detectability can be defined as

$$D(t_i) = \min(t_i - t_{i-1} + 1, t_{i+1} - t_i + 1).$$

When the detectability is low, the current changing point is often missed; moreover, this error contaminates the statistics for the next Gaussian model, thus affects the detection of the next changing point.

Our algorithm has a quadratic complexity; however, one can reduce the complexity dramatically by performing a crude search without much sacrifice of the resolution.

3.3. Change detection on the Hub4 1997 evaluation data

We applied our algorithm on the Hub4 1997 evaluation data, which consists of 3 hour broadcasting news programs; detection was performed using 24-dimensional Mel-cepstral vectors exacted at $10ms$ frame rate.

NIST provided hand-segmentation of this data according to different categories: clean prepared speech, clean spontaneous speech, telephone-quality speech, speech with background music and speech with background noise. As commented in [4], it is very hard to come up with a standard for analyzing the errors in segmentation since segmentation can be very subjective; even two people listening to the same speech may segment it differently. Nevertheless, we analyze the performance of our detection by comparing with the hand-segmentation provided by NIST.

We first examine whether our detected changing points were true, i.e. the Type-I errors. Among the 462 detected changes, there were 19 (4.1%) errors which happened in the middle of speaker turns. Our BIC criterion seems sensitive in pure music region. There were 14 (3.0%) detected changes in the middle of pure music segments; we did not count them as errors since first one can argue that the music tune changed in those areas, second the pure music segments were discarded by the classifier and did not affect the recognition accuracy. There were 20 (4.3%) detected changes slightly biased from the true changes. The biases were less than 1 seconds, as shown in panel (a) in Figure 3. We did not count these as errors since they came so close to the true changes. The bias might be caused by contamination of the statistics for estimating the Gaussian models by outliers, or by the statistics from the previous turn if the previous change point was missed in the detection. Usually these errors can be fixed, for example, by moving to the nearest silence. It is also possible to refine the boundary by finer analysis in the detected region.

We also examine whether true changing points were missed in our detection, i.e. the Type-II errors. In the NIST segmentation, there were 620 changes. Totally 207 (33.4%) changes were missed. 154 (25.0%) errors were caused by short turns with duration less than 2 seconds. Examples

Type-I Error	4.1%		
Type-II Error	33.4%	$< 2s$	25.0%
		$> 2s$	8.4%

Table 1. Change detection error rates

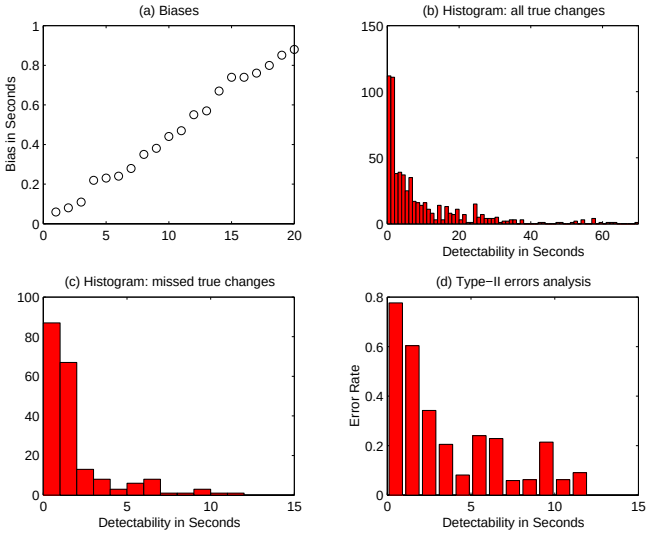

Figure 3. Error analysis of change detection

of these short turns are sentences made up of only brief phrases such as "Good morning" and "Thank you". About 50 of these short turns contained voices from more than one speaker. They were labeled as "excluded regions" by NIST and were not included in the final scoring of the recognition system but were included in determining the change detection accuracy. Figure 3 analyzes the Type-II errors in detail. Panel (b) shows the histogram of the detectability of the true changes; there were 223 true changes with detectability less than 2 seconds. Panel (c) shows the histogram of the detectability of the true changes which were missed in the detection; it is clear that most of the errors came from low detectabilities less than 2 seconds. Panel (d) describes the Type-II error rates according to different degrees of detectability: when detectability is below 1 second, as the type-II error rate is 78%, most such changing points were missed; as the detectability increases, the Type-II error drops.

4. CLUSTERING VIA BIC

In this section, we describe how to apply the BIC criterion in clustering. Let $\mathcal{S} = \{s_i : i = 1, \cdots, M\}$ be the collection of signals we wish to cluster; each signal is associated with a sequence of independent random variables $\mathcal{X}^i = \{x_j^i : j = 1, \cdots, n_i\}$. In the context of speech clustering, $\mathcal{S}$ is a collection of audio segments; $\mathcal{X}^i$ can be the cepstral vectors exacted from the i'th segment. Denote $N = \sum_i n_i$ as the total sample size of the vectors $\mathcal{X}^i$.

Let $\mathcal{C}_k = \{c_i : i = 1, \cdots, k\}$ be the clustering which has k clusters. We model each cluster c_i as a multivariate Gaussian distribution $N(\mu_i, \Sigma_i)$, where μ_i can be estimated as the sample mean vector and Σ_i can be estimated as the sample covariance matrix. Thus the number of parameters

for each cluster is $d + \frac{1}{2}d(d+1)$. Let n_i be the number of samples in cluster c_i. One can show that

$$BIC(\mathcal{C}_k) = \sum_{i=1}^{k}\{-\frac{1}{2}n_i \log|\Sigma_i|\} - \lambda P \qquad (7)$$

where the penalty

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1))\log N$$

and the penalty weight $\lambda = 1$. We choose the clustering which maximizes the BIC criterion.

4.1. Hierarchical Clustering via greedy BIC

As one can imagine, it is often very costly to search globally for the best BIC value, since clustering has to be performed to obtain different numbers of clusters. However, for hierarchical clustering methods, it is possible to optimize the BIC criterion in a greedy fashion.

Bottom-up methods start with each signal as one initial node, then successively merge two nearest nodes according to a distance measure. Let $\mathcal{S} = \{s_1, \cdots, s_k\}$ be the current set of nodes; suppose s_1 and s_2 are the candidate pair for merging, and the merged new node is s. Thus we are comparing the current clustering $\mathcal{S}$ with a new clustering $\mathcal{S}' = \{s, s_3, \cdots, s_k\}$. We model each node s_i as a multivariate Gaussian distribution $N(\mu_i, \Sigma_i)$. It is clear from (7) that the increase of the BIC value by merging s_1 and s_2 is

$$BIC = n\log|\Sigma| - n_1\log|\Sigma_1| - n_2\log|\Sigma_2| - \lambda P \qquad (8)$$

where $n = n_1 + n_2$ is sample size of the merged node, Σ is the sample covariance matrix of the merged node, the penalty

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1))\log N$$

and the penalty weight $\lambda = 1$.

Our BIC termination procedure is that two nodes should not be merged if (8) is negative. Since the BIC value is increased at each merge, we are searching for an "optimal" clustering tree by optimizing the BIC criterion in a greedy fashion.

Note that we merely use our criterion (8) for termination. It is possible to use our criterion (8) as the distance measure in the bottom-up process. However, in many applications, it is probably better to use more sophisticated distance measures. It is also clear that our criterion can be applied to top-down methods.

4.2. Speaker Clustering on the Hub4 1996 evaluation data

The data set consists of the clean prepared and the clean spontaneous portion of the HUB4 1996 evaluation data [2], hand-segmented into 824 short segments. Cepstral coefficients were extracted as feature vectors $\mathcal{X}^i$ for each segment. We used the log likelihood ratio distance measure; Bottom-up clustering was performed with maximum linkage, with the BIC termination criterion (8).

The true number of speakers is 28; the BIC termination criterion chose 31 clusters. For each cluster, we define the purity as the ratio between the number of segments by the dominating speaker in that cluster and the total number of segments in that cluster. Figure 3 shows the purities of each cluster. Clearly our algorithm results in not only clusters with high purity, but also the appropriate number of clusters.

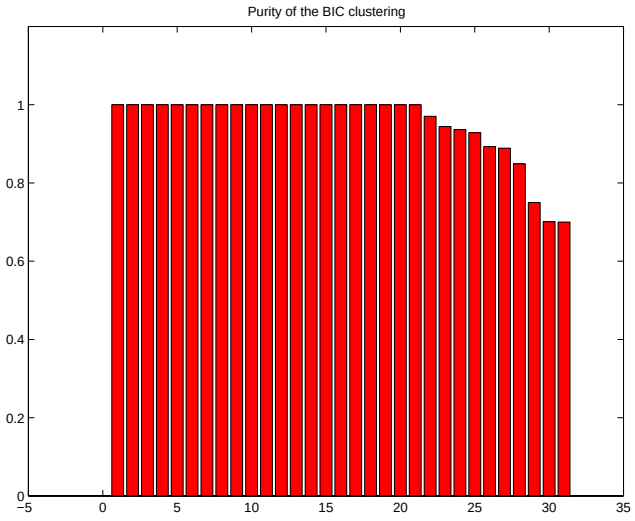

Figure 4. Clustering Purities

	Prepared	Spontaneous
Baseline	18.8%	27.0%
MLLR w/o clustering	18.7%	26.9%
MLLR w/ ideal clustering	17.5%	24.8%
MLLR w/ BIC clustering	17.5%	24.6%

Table 2. MLLR adaptation enhanced by BIC clustering

Speaker clustering can enhance the performance of unsupervised adaptation. The reason is that most of the 824 segments here are quite short, around $2 \sim 3$ seconds. Without speaker clustering, unsupervised adaptation techniques such as MLLR [9] has small improvements due to lack of data. Good speaker clustering can bring the segments of the same speaker together thus improving the performance of unsupervised adaptation. We started from a baseline system which had about 90k Gaussians. The decoding results were scored according to two conditions: clean prepared and clean spontaneous. As shown in Table 2, the baseline error rates were 18.8% and 27.0% for the two conditions respectively. Without clustering, MLLR reduced the error rates by only 0.1%. With our clustering, MLLR reduced the error rates by 1.3% for the clean condition and by 2.4% for the spontaneous condition. Table 2 also shows the error rates of MLLR using the ideal clustering by the true speaker identities. It is clear that our speaker clustering enhanced the performance of MLLR as much as the ideal clustering.

4.3. Discussion

Jin et al. of BBN [7] proposed a similar automatic speaker clustering algorithm. They also used the log likelihood ratio distance measure proposed in Gish et al. [6], however, with the distances between consecutive segments scaled down by a parameter α. They performed hierarchical clustering; for any given number k, the clustering tree was pruned to ob-

tain k tightest clusters. A heuristic model selection criterion

$$\sum_{j=1}^{k} n_j^{\alpha} |\Sigma_j^{\alpha}| * \sqrt{k} \tag{9}$$

was then used to search through the space of (α, k) for the best clustering. They applied this algorithm to cluster the HUB4-96 evaluation data for the purpose of unsupervised adaptation. Similar to our results above, this automatic clustering enhanced the unsupervised adaptation as much as the ideal clustering according the the true speaker identities.

This heuristic model selection criterion (9) resembles the BIC criterion (7): they both penalize the likelihood by the number of clusters. However, the BIC criterion has a solid theoretical foundation and seems more appropriate. Indeed the number of speaker clusters found in [7] is considerably less than the truth. Moreover, extra information such as the adjacency of the segments was utilized in [7].

Siegler et al. of CMU [10] proposed another speaker clustering algorithm. They chose the symmetric Kullback-Leibler metric as the distance measure, and performed hierarchical clustering. The clusters were obtained by thresholding the distances. Unlike our method and the BBN clustering, this clustering is not fully automatic: the thresholding level was tuned in a delicate fashion: it had to be small enough such that the clusters created were made up of segments from only one speaker and yet large enough to improve the performance of the unsupervised adaptation.

5. CONCLUSION

We presented a maximum likelihood approach to detecting changing points in a independent Gaussian process; the decision of a change is based on the BIC criterion. The key features of our approach are:

- Instead of making local decision based on distance between two adjacent sliding windows of fixs sizes, we expand the decision windows as wide as possible so that our final decision of change points can be more *robust*.
- Our approach is thresholding-free. The BIC criterion can be viewed as thresholding the log likelihood distance, with the thresholding level automatically chosen as $\lambda \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N$ where N is the size of the decision window and d is the the dimension of the feature space.

We also proposed to apply the BIC criterion as a termination criterion in the hierarchical clustering. Our change detection algorithm can successfully detects acoustic changing points with reasonable detectability ($> 2s$); Our experiments on clustering demonstrated that the BIC criterion is able to choose the number of clusters according to the intrinsic complexity present in the data set and produce clustering solution with high purity.

We applied our algorithms on the Hub4 1997 evaluation data [3]. Table 3 shows the recognition error rates. Our segmentation was only 0.6% worse than the NIST hand-segmentation. After clustering, the unsupervised adaptation further reduced the error rate by 2.7%.

	Error Rate
NIST hand-segmentation	19.8%
IBM segmentation	20.4%
adaptation after clustering	17.7%

Table 3. Segmentation and clustering in Hub4 1997 task

We comment that the penalty weight λ in the BIC criterion could be tuned to obtain various degrees of segmentation and clustering. A smaller weight would result in more changes and more clusters. In this paper, we simply choose $\lambda = 1$ according to the BIC theory.

REFERENCES

[1] H. Akaike, "A new look at the statistical identification model", IEEE Trans. Auto. Control, vol 19, pp 716-723, 1974.

[2] R. Bakis et al., "Transcription of broadcast news shows with the IBM large vocabulary speech recognition system", Proceedings of the Speech Recognition Workshop, pp 67-72, 1997.

[3] S. Chen et al., "IBM's LVCSR System for Transcription of Broadcast News Used in the 1997 Hub4 English Evaluation", Proceedings of the Speech Recognition Workshop, 1998.

[4] H. Beigi and S. Maes, "Speaker, channel and environment change detection", Proceedings of the World Congress on Automation, 1998.

[5] D. Foster and E. George, "The risk inflation factor in multiple linear regression", Technical Report, Univ. of Texas, 1993.

[6] H. Gish and N. Schmidt, "Text-independent speaker identification", IEEE Signal Processing Magazine, pp 18-21, Oct. 1994.

[7] H. Jin, F. Kubala and R. Schwartz, "Automatic speaker clustering", Proceedings of the Speech Recognition Workshop, pp 108-111, 1997.

[8] F. Kubala et al., "The 1996 BBN Byblos Hub-4 transcription system", Proceedings of the Speech Recognition Workshop, pp 90-93, 1997.

[9] 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, pp 171–186.

[10] M. Siegler, U. Jain, B. Ray and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio", Proceedings of the Speech Recognition Workshop, pp 97-99, 1997.

[11] G. Schwarz, "Estimating the dimension of a model", The Annals of Statistics, vol. 6, pp 461-464, 1978.

[12] K. Shinoda et al., "Speaker adaptation with autonomous model complexity control by MDL principle", Proceedings of ICASSP, pp 717-720, 1996.

[13] W.S. Wei, Time Series Analysis, Addison-Wesley, 1993.

[14] P. Woodland, M. Gales, D. Pye and S. Young, "The Development of the 1996 HTK broadcast news transcription system", Proceedings of the Speech Recognition Workshop, pp 73-78, 1997.