Efficient Audio Stream Segmentation via the Combined T^2 Statistic and Bayesian Information Criterion

Bowen Zhou, *Member, IEEE,* and John H. L. Hansen, *Senior Member, IEEE*

***Abstract*—In many speech and audio applications, it is first necessary to partition and classify acoustic events prior to voice coding for communication or speech recognition for spoken document retrieval. In this paper, we propose an efficient approach for unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion (BIC). The proposed method extends an earlier formulation by Chen and Gopalakrishnan. In our formulation, Hotelling's T^2-Statistic is used to pre-select candidate segmentation boundaries followed by BIC to perform the segmentation decision. The proposed algorithm also incorporates a variable-size increasing window scheme and a skip-frame test. Our experiments show that we can improve the final algorithm speed by a factor of 100 compared to that in Chen and Gopalakrishnan's while achieving a 6.7% reduction in the acoustic boundary miss rate at the expense of a 5.7% increase in false alarm rate using DARPA Hub4 1997 evaluation data. The approach is particularly successful for short segment turns of less than 2 s in duration. The results suggest that the proposed algorithm is sufficiently effective and efficient for audio stream segmentation applications.**

***Index Terms*—Audio segmentation, Bayesian information criterion, Hotelling's T^2-statistic, spoken document retrieval.**

I. Introduction

In MANY speech and audio applications, it is first necessary to partition and classify acoustic events prior to voice coding for communication or decoding in speech recognition for spoken document retrieval. A range of tasks must deal with continuous audio streams such as Broadcast news data that contain a wide variety of data types including clean speech, narrow-band speech, speech corrupted by music or background noises, and music segments. To efficiently operate with such audio streams, an audio parsing process will be required. Ideally, the procedure of audio parsing is to first identify speaker, channel and/or other environment changes in an audio stream, followed by a labeling process for each segment, and finally to cluster acoustically homogeneous segments. Typically, the *a priori* knowledge of the acoustic conditions and speakers in the audio stream is absent during the parsing process. Therefore, the parsing needs to be conducted in an *unsupervised* manner.

In general, the nonverbal information extracted from the audio parsing procedure allows for more accurate and specific subsequent processing for special interest audio segments. For example, for the task of target speaker tracking, reliable segmentation and clustering would provide both pure and sufficient data that could be used to improve a data-driven speaker identification/verification system. Another example is for a system that performs automatic audio transcription. In such tasks, parsing information can be used to localize the occurrences of a specific speaker, channel or environment so that data can be pooled for improved model adaptation and thereby boost transcription performance. This is particularly important for Large Vocabulary Continuous Speech Recognition (LVCSR) based spoken document transcription and retrieval systems, since audio parsing plays the crucial role of segmenting spoken documents to feed the subsequent automatic transcription process and extract nonverbal information to guide the retrieval task.

Motivated by these applications, unsupervised audio segmentation and clustering has been explored by several researchers in recent years [3], [6], [15]. In this paper, we investigate several statistical methods that can be applied to parse the audio document.

The remainder of this paper is organized as follows. Section II reviews a number of algorithms for audio segmentation and clustering proposed in previous studies. Section III describes the formulation of a new proposed segmentation algorithm via the Hotelling's T^2-Statistic and Bayesian Information Criterion (BIC). Performance evaluation is also presented. Section IV presents discussions, followed by a summary in Section V.

II. Background: Algorithms for Audio Segmentation

Due to the importance of audio parsing algorithms for speech processing, a number of approaches have been proposed in recent years. In this section, we will review a number of typical approaches.

Segmentation is the key process of audio parsing since the subsequent clustering process depends highly on the quality of the segments obtained. Various segmentation algorithms have been proposed in the literature [3], [6], [15], [16], [8]. A number of studies have also considered segmentation as part of Broadcast News transcription [3], [5], [8], [15], [16], [18], [19], of which [18] and [19] compared the T2-distance measure

Manuscript received September 21, 2001; revised September 15, 2003. This work was supported by NSF under Cooperative Agreement IIS-981748 and SPAWAR under Contract N66001-03-1-8905. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NSF. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Vary.

B. Zhou was with the Robust Speech Processing Group, University of Colorado, Boulder, CO 80302 USA. He is now with IBM, Yorktown Heights, NY 10598 USA.

J. H. L. Hansen was with the Robust Speech Processing Group, University of Colorado, Boulder, CO 80302 USA. He is now with the University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: john.hansen@colorado.edu).

Digital Object Identifier 10.1109/TSA.2005.845790

to model based BIC. Based on these underlying methods, the algorithms roughly fall into the following categories: 1) metric based, 2) Gaussian mixture model (GMM) based, 3) recognition based, and 4) model selection based.

A. Metric Based Segmentation

A straightforward method of audio segmentation is to detect acoustic change points based on spectral changes. The underlying assumption is that the data of different acoustic types possess different spectral shapes, and these differences can be sufficiently measured by the distances between the acoustic feature vectors. In practice, the spectral changes are identified at the maxima of the dissimilarity in terms of some metric between neighboring windows that shift along the audio stream. Here, a window is typically 2 s, which should be longer than a speech frame. As one can envision, the choice of an appropriate distance measure is essential to segmentation performance for this class of algorithms. Previous studies have introduced the use of the Generalized Likelihood Ratio [9] and the symmetric Kullback-Leibler distance (KL2) [15] as the distance metric. If the windowed observations are modeled by the multivariate Gaussian distributions $\mathbf{N}(\mu_1, \Sigma_1)$ and $\mathbf{N}(\mu_2, \Sigma_2)$, then the KL2 distance [4] between these two neighboring windows is defined by

$$KL2_{1,2} = \frac{1}{2}(\mu_1 - \mu_2)'(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) + \frac{1}{2}\mathrm{tr}(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I). \quad (1)$$

Such a distance measure is continually calculated between neighboring windows along the audio stream and a distance curve is formed. To avoid fluctuations due to noise-corrupted speech, this curve is often smoothed using a low-pass filter. The local peaks over the curve can be treated as candidate segmentation points. However, it is often difficult to determine the final segmentation points from these candidates since this requires suitable thresholding, which is often tuned from training data, but cannot guarantee stability and robustness for all test data.

B. Segmentation Using Gaussian Mixture Models

Gaussian mixture modeling (GMM) is widely used to model observations with unknown probability density functions (pdf) since any Riemann integrable distribution can be approximated by Gaussian mixtures, and the mathematical framework offers a rich set of computational techniques for using Gaussian distributions. With the given GMM, the observation vector $\mathbf{o}_t$ at time t is distributed as

$$p(\mathbf{o}_t \mid \Lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{o}_t \mid \Lambda_m) \quad (2)$$

where

$$p_m(\mathbf{o}_t \mid \Lambda_m) = \mathbf{N}(\mathbf{o}_t; \mu_m, \Sigma_m) = \frac{1}{(2\pi)^{n/2}\left|\Sigma_m\right|^{1/2}} e^{-(1/2)(\mathbf{o}_t - \mu_m)'\Sigma_m^{-1}(\mathbf{o}_t - \mu_m)}. \quad (3)$$

Here, n is the dimension of the audio feature vector, M is the number of mixtures, w_m is the mixture weight of the mth component of the mixture Gaussian with the constraint that $\sum_{m=1}^{M} w_m = 1$, and μ_m and Σ_m are the mixture mean and covariance. To apply a GMM for audio segmentation [3], it is assumed that different sets of Gaussian mixture model parameters can be estimated for the C classes of different acoustic conditions from the training data. For the incoming audio stream, the observation features are classified into one of C classes in a maximum likelihood manner

$$c_t = \mathrm{argmax}_{c \in 1,2,\ldots,C} p_c(\mathbf{o}_t \mid \Lambda^c) \quad (4)$$

where $p_c(\mathbf{o}_t \mid \Lambda^c)$ is the likelihood of the current observation generated from acoustic class c. The segmentation decision is made at locations where the acoustic class changes. It can be seen that such methods require pre-trained GMM's and *a priori* knowledge to define the acoustic classes before test data is processed. Therefore, this scheme is less practical for segmenting audio streams containing complex acoustic conditions. Generally, GMM processing is used more widely as a pre-processing step of the segmentation process, to identify speech and nonspeech turns [8], or to further distinguish wideband and telephone speech [16].

C. Recognition-Based Segmentation

Hain *et al.* [16] proposed a segmentation system for Broadcast News transcription, which extracts more complicated information from a multi-pass decoding process to perform the segmentation. In this method, the original audio stream is first decoded using a conventional Viterbi search over a network of only 4 states, each of which models the "wideband speech", "telephone speech", "music or nonspeech" and "speech and music." An inter-class transition penalty is used to prevent frequent transfer between states and thus to produce longer segments. To improve frame classification accuracy, the underlying acoustic models of these 4 states are dynamically adapted using Maximum Likelihood Linear Regression (MLLR). Next, pure nonspeech segments are discarded, and others are decoded through another round of gender-dependent phone recognition. The phone recognizer contains 45 context independent phone models per gender plus a silence/noise model with a null language model. The output is a phone sequence with male, female or silence tags. The phone tags are ignored and the phone sequence with the same gender label are merged. A set of heuristic rules are further applied to smooth the gender boundaries. Finally, the change points between genders are marked and thus the audio stream is segmented by gender transitions. It can be seen that this method requires a relatively complicated flow process for segmentation. Furthermore, the general disadvantage is that it is unable to detect speaker transitions between two speakers of the same gender when there is no significant intervening silence. This will be problematic for subsequent tasks such as speaker tracking.

D. Segmentation as a Model Selection

An alternative approach is proposed by Chen and Gopalakrishnan [6]. In their study, the segmentation problem is reformulated as a model selection task between two nested competing models. This method employs the Bayesian Information Criterion (BIC) as the model selection criterion, illustrating several desirable properties such as robustness, threshold independence and optimality. BIC [14] is a penalized maximum likelihood model selection criterion that has been widely used in statistical data processing. With such a scheme, the segmentation decision is derived by comparing BIC values. Other advantages of this scheme include that no prior knowledge concerning acoustic conditions is required and no prior model training is needed. In comparison to previously described GMM- or metric-based methods, this class of model-selection-based approaches is distinctive for its sound mathematical foundation. More information regarding BIC and its application in the segmentation is covered in Section III. However, the original BIC scheme developed in [6] is extremely computationally expensive with quadratic complexity and therefore has limitations for real-time applications. In this study, we formulate a more effective approach which can significantly reduce the computational requirements as well as provide more reliable segmentation and clustering performance.

III. Segmentation via T^2-BIC

A. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is a model selection criterion that was first proposed by Schwarz [14] and widely used in the statistical literature. The problem of model selection is to choose one among a set of candidate models $\mathbf{M}_i$, $i = 1, 2, \ldots, m$, and corresponding model parameters θ_i to represent a given data set $D = (D_1, D_2, \ldots, D_N)$. These candidate models may be nested or nonnested. The BIC of model M_i for the given data is defined as

$$\mathrm{BIC}(M_i) = \log P(D_1, D_2, \ldots, D_N \mid M_i) - \frac{1}{2} d_i \log N \quad (5)$$

where d_i is the number of *independent* parameters in the model parameter set, and $P(D_1, D_2, \ldots, D_t \mid M_i)$ is the maximized data likelihood for the given model. In BIC, the term $(1/2) d_i \log N$ is subtracted from the log-likelihood to penalize for model complexity, where BIC favors the model which maximizes the BIC values. The procedure was originally derived in [14] as a large-sample Bayesian inference for the case of independent, identically distributed (i.i.d.) observations and linear models by assuming that the prior probabilities of all models were equal. The results apply much more widely than this, however, and in essence are valid for any regular statistical model [12] (i.e., one in which the Maximum-Likelihood Estimator (MLE) is asymptotically normal with the mean as the true value and the variance matrix is set equal to the inverse expected Fisher information matrix). Therefore, BIC can be used to compare models with differing parameterizations, differing number of components, or both. In the case of only two competing models, the BIC difference can be seen as an approximation to the logarithm of the Bayes factor [12].

For the model selection process, BIC can be interchanged with another well-known criterion (e.g., Akaike information criterion (AIC) [1]). However, some important differences make them distinct. Generally, BIC is asymptotically consistent as a selection criterion, which means that given a family of models, including the true model, the probability that BIC will choose the correct model approaches one as the sample size $N \to \infty$. However, AIC behaves differently and tends to choose the more complex models as $N \to \infty$. On the other hand, for finite samples, BIC often selects the simple model due to its heavy penalty against complexity. This observation suggests that, for applications dealing with varied sample sizes such as the type of audio segmentation we will consider, it is reasonable to adjust the penalty of complexity especially when sample size is small.

In recent years, BIC has attracted more attention in the speech community and has been applied in HMM training tasks such as mixture size selection [5] and decision tree state tying [7], [13].

B. BIC Segmentation

Let us denote $X = x_i \in R^d, i = 1, 2, \ldots, N$ as the sequence of framed-based cepstral vectors extracted from an audio stream in which there is at most one segment boundary. We wish to consider if there is a boundary at frame $b \in (1, N)$. If we suppose that each acoustic homogeneous speech block can be modeled as one multivariate Gaussian process $X \sim N(\mu, \Sigma)$, the segmentation issue can be cast as a model selection problem between the following two nested models [6]

$$\begin{aligned} M_1 : X = & x_1, x_2, \ldots, x_N \sim N(\mu, \Sigma) \\ \text{and } M_2 : & x_1, x_2, \ldots, x_b \sim N(\mu_1, \Sigma_1); \\ & x_{b+1}, x_{b+2}, \ldots, x_N \sim N(\mu_2, \Sigma_2). \end{aligned}$$

That is, the first model assumes that all samples are independent and identically distributed to a single Gaussian, and the second model assumes the first b samples are drawn from one Gaussian while the last $N - b$ samples are drawn from another Gaussian. We can see that the i.i.d. condition does not hold for M_2, but as stated in Section III-A, regularity conditions of the assumed Gaussian distribution (i.e., the assumed Gaussian distribution is regular) support the BIC's application in this context. Under this expression, if BIC favors M_1 then the data is assumed homogeneous, otherwise a break should occur within this block of data.

It can be shown that given the assumption of a normal distribution $N(\mu, \Sigma)$, the likelihood of observation data $x_1, x_2, \ldots, x_N$ is maximized when $\mu = \hat{\mu}$ and $\Sigma = \hat{\Sigma}$, where

$$\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad (6)$$

and

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})'. \quad (7)$$

According to (5), the BIC values of these two models can be computed as

$$\begin{aligned} \mathrm{B\bar{I}C}(M_1) = & -\frac{d}{2} N \log 2\pi - \frac{N}{2} \log |\hat{\Sigma}| - \frac{N}{2} \\ & -\frac{1}{2} \lambda \left(d + \frac{1}{2} d(d+1) \right) \log N \quad (8) \end{aligned}$$

and similarly

$$\begin{aligned} \mathrm{B\bar{I}C}(M_2) = & -\frac{d}{2} N \log 2\pi - \frac{b}{2} \log |\hat{\Sigma}_1| - \frac{N-b}{2} \log |\hat{\Sigma}_2| \\ & -\frac{N}{2} - \lambda \left(d + \frac{1}{2} d(d+1) \right) \log N \quad (9) \end{aligned}$$

where $\hat{\Sigma}$, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are ML covariance estimations from corresponding sample data, λ is the penalty factor to compensate for small sample size cases,[1] and d is the cepstral feature dimension. Next, the BIC difference between the two models can be computed as a function of break point b

$$\begin{aligned}\Delta\mathrm{BIC}(b) &= \mathrm{B\bar{I}C}(M_2) - \mathrm{B\bar{I}C}(M_1)\\ &= \frac{1}{2}(N\log|\hat{\Sigma}| - b\log|\hat{\Sigma}_1| - (N-b)\log|\hat{\Sigma}_2|)\\ &\quad - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log N. \end{aligned} \tag{10}$$

According to the BIC rule, segmenting this audio stream into two parts at frame b will be favored if $\Delta\mathrm{BIC}(b) > 0$. The final segmentation point decision can be achieved via MLE

$$\hat{b} = \mathrm{argmax}_{1<b<N;\Delta\mathrm{BIC}(b)>0}\Delta\mathrm{BIC}(b). \tag{11}$$

For an audio stream that contains multiple segmentation boundaries, a sequential detection algorithm was proposed in [6] which investigates a moving window that sweeps through the audio stream. The window starts from the beginning of the stream with a width of 1 s. Inside the current window, the BIC test of (10) is evaluated for every $1 < b < N$ to determine if a boundary exists. The window is extended forward by 1 s if no boundary is found, or a new window is started from the detected boundary as the next window.

C. Integrating T^2-Statistic With BIC

It can be seen from the previous section that the BIC-based segmentation algorithm has quadratic complexity. Although the speed can be improved by performing the search over a grid (say, at every 30 frames for a frame speed of 100 frames/s), the computational cost is still extensive since we need to evaluate the determinants of *two full covariance matrices for every possible break point b in a window*. Also, since the mean and covariance for these distributions must be estimated, the error in the estimation, especially for the covariance matrix, will be high for shorter duration acoustic events. Therefore, we are motivated to propose and derive a faster and more efficient approach to detect the possible boundary through the T^2-Statistic.

Hotelling's T^2-Statistic is a multivariate analog of the well-known t-distribution [2]. One application of Hotelling's T^2-Statistic is to test the null hypothesis that the mean of one normal population is equal to the mean of another where the covariance matrices are assumed equal but unknown.

In terms of segmentation, the problem can be stated as follows: for a given audio stream $X = x_i \in R^d$, $i = 1, 2, \ldots, N$, determine if the two samples, one containing the frames $[1, b]$ and the second contains $[b+1, N]$, are homogeneous. If the covariance of the audio sample is assumed to be common and unknown, the two samples are homogeneous if and only if they are drawn from the same underlying normal distribution. With this, the segmentation problem can be viewed as testing the hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$, where μ_1, μ_2 are the means of these two samples respectively. As derived in [2], the likelihood ratio test is given by the following T^2-Statistic

$$T^2 = \frac{b(N-b)}{N}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \tag{12}$$

where Σ is the common covariance matrix. The T^2 value is distributed as T^2 with $N-2$ degrees of freedom [2]. The critical region is

$$T^2 \geq \frac{(N-2)d}{N-d-1}F_{d,N-d-1}(\alpha) \tag{13}$$

with a significance level α, where $F_{d,N-d-1}(\alpha)$ is the F-point with d and $N-d-1$ degrees of freedom and significance level α.

One can directly employ (13) to form a metric-based segmentation scheme. However, simply applying this measure involves the issue of setting reliable thresholds, which is often difficult for varying acoustic environments. On the other hand, we note that the T^2 value defined in (12) can also be used as a distance measure of two samples. Obviously, the smaller the value of T^2, the more similar the two sample distributions. Within the given audio stream, the location where the maximum value of T^2 is achieved is the candidate point of interest that forms the two most dissimilar samples, which suggests an acoustic event change *may* be present. Fig. 1 illustrates an example of the T^2-Statistic as a function of the segmentation point b, testing over a speech stream which lasts 37 s. The audio stream contains a speaker change point from speaker A to speaker B at the time location of 27 s, as indicated by the dotted line in the figure. We can see that the peak position of the T^2-Statistic value reveals clearly the location of the change frame.

The above observations thus motivates the integration of the T^2-Statistic into the BIC-based segmentation. In this scheme, the T^2-Statistic, here referred to as $T^2(b)$, is first evaluated using (12) along the grid for the window of the audio stream under consideration. Next, the point $\hat{b}$ where the global maximum T^2 is achieved is identified, and the BIC rule according to (10) is **only** applied at this point $\hat{b}$ to either accept or reject the segmentation hypothesis for this window of audio [20].

There are several advantages supporting this simple combination. First, by pre-selecting the candidate break points through the T^2-Statistic, we avoid the computation of two full covariance matrices at other points, and thus remove an order of $(N+2)d^2$ multiplications, compared to the original BIC algorithm for each sliding window. Therefore, integrating the T^2-Statistic with BIC is more efficient for boundary identification using a sequential detection manner, yet maintains the benefits of BIC such as threshold independence and a firm mathematical foundation, in contrast to making the segmentation decision simply using (13). Second, the ML-based break point choice according to (11) tends to be unreliable when the sample size is small or the break point occurs adjacent to the window boundary. When a break occurs within a small-sized window, insufficient data often makes the second-order statistics biased and thus produces incorrect break point decisions. On the other hand, the evaluation of the T^2-Statistic only requires the first-order statistics and hence is more robust for small sample cases. Thus, pre-selecting candidate break points via T^2 can prevent some mis-locations in BIC segmentation.

[1] Since BIC tends to favor a simple model when the sample size is small, setting a smaller λ will increase the probability that a complicated model is selected.

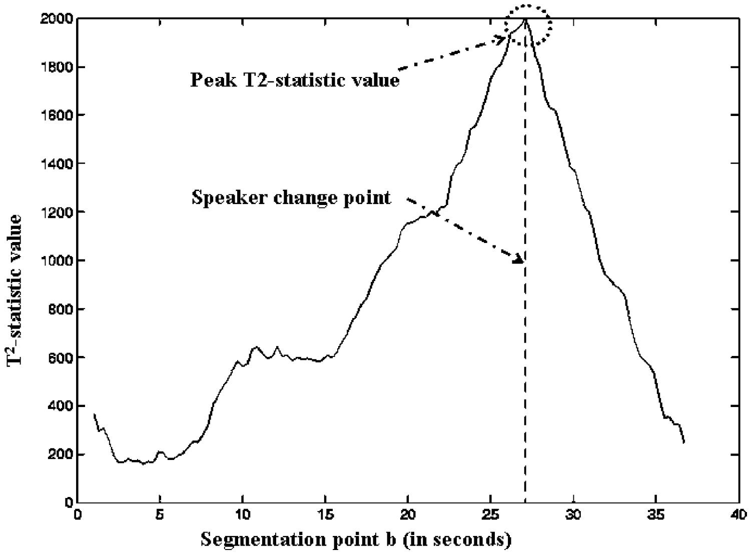

Fig. 1. The values of T^2-Statistic as a function of break point b over a speech stream of 37 s, which contains a speaker change positioned at the time location indicated by the dotted line.

D. Further Speed and Efficiency Gains

In addition to the integration of the T^2-Statistic within our algorithm, we consider several further improvements, among them a variable-size window increasing scheme and skip-frames test, which have been discussed in [17] for pure segmentation. Next, we discuss how we extend and integrate them into the T^2-Statistic test in this study. The amount and purity of the data within each window is a vital element for making reliable statistical decisions. In the sequential segmentation algorithm, the width of the current window has a significant impact on the pre-selection of candidate break points via the T^2-Statistic and subsequent BIC decision. If the window is too wide in duration to allow more than one change point to be contained, then the assumption of model selection is not valid. If the window is too short, a lack of data will cause poor Gaussian estimation and result in an incorrect segmentation decision. Moreover, these errors will contaminate the Gaussian statistics of the subsequent window, thus affecting the detection of the next segment boundary. Therefore, we employ a heuristic dynamic window increasing scheme.

We begin with a window width of $W_0 = 200 \text{ frames}^2$. Across the audio stream, if no break is found within the previous window W_{i-1}, the current window width W_i is set as

$$\begin{aligned} W_i &= W_{i-1} + \Delta W_i \\ &\text{where} \\ \Delta W_i &= \begin{cases} 100 & \text{if } W_{i-1} < 500 \\ \Delta W_{i-1} + 50 & \text{otherwise.} \end{cases} \end{aligned}$$

The motivation for this heuristic increasing scheme is that we need to be more careful when the window size is relatively small, while we can scan the audio at a higher speed when the data window increases in size, since reliable statistics can be expected. Moreover, the current window width W_i is also controlled by the T^2-Statistic peak position of the previous window. If the peak appears close to the end window boundary within a threshold in the previous window, this may suggest an approaching break, we therefore reset $\Delta W_i = 50$. By employing such an adjustable increasing window step, we are better able to capture small segment breaks, and scan the stream at a much faster rate if no breaks occur during homogeneous data.

The second efficiency improvement is gained from the frame skipping test. The point is that not all frames within a window need to be considered as a candidate boundary, especially when the current window is large. For example, the data segments close to the window boundary can be excluded from T^2 testing since we cannot obtain robust Gaussian estimation with such limited data. Furthermore, for long windows (say, greater than 1000 frames), it is less likely that a frame break occurs in the beginning part of the current window since it is difficult for such a break to survive from the previous segmentation test in the previous window. Therefore, we also exclude these frames for T^2-Statistic testing.

Another proposed improvement can be achieved by dynamically computing the entire window covariance matrix Σ, which is used by both the T^2-Statistic and BIC test. Assume that we have not yet detected a break for some period of time, so the window length continues to grow. We can compute the current window covariance by combining the last window and new extended data statistics. In such a manner, we can avoid repeated computation of the covariance from overlapping data between consecutive windows. We can see this if we consider the current window width to be $W_i = W_{i-1} + \Delta W_i$, where we have the covariance matrix for the current window as follows:

$$\begin{aligned} \Sigma_i(k,l) = \frac{1}{W_i}\{&W_{i-1}(\Sigma_{i-1}(k,l) + \mu_{i-1}(k)\mu_{i-1}(l)) \\ &+\Delta W_i(\Sigma_{\Delta i}(k,l) + \mu_{\Delta i}(k)\mu_{\Delta i}(l))\} - \mu_i(k)\mu_i(l) \end{aligned} \quad (14)$$

where $0 \le k, l \le d$, μ_{i-1}, Σ_{i-1}, $\mu_{\Delta i}$, and $\Sigma_{\Delta i}$ are the means and covariance matrices of the previous window and the new added data, respectively, and μ_i is the current entire window mean

$$\mu_i(k) = \frac{1}{W_i}\{\mu_{i-1}(k)W_{i-1} + \mu_{\Delta i}(k)\Delta W_i\}. \quad (15)$$

[2] Note that we use a frame rate of 100 frames/s, with each frame of 20 ms with an overlap of 50%.

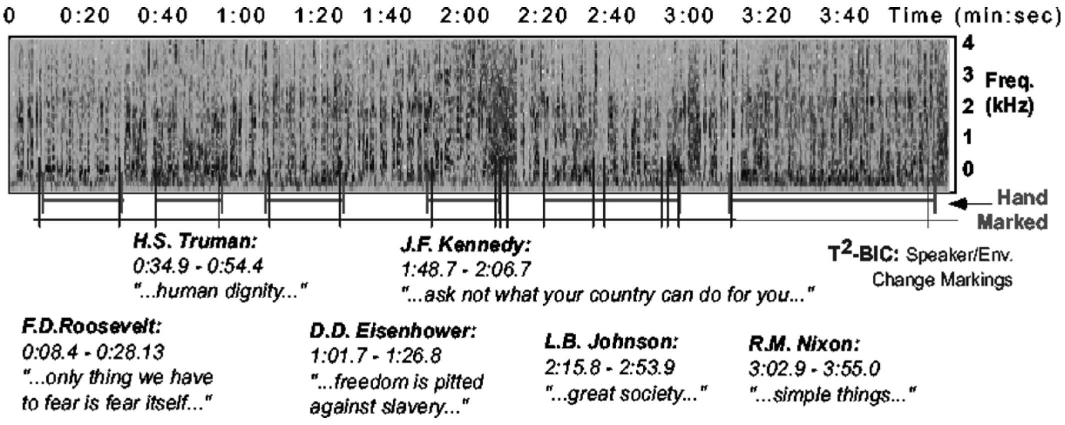

Fig. 2. Historian discussing U.S. history with audio examples from six U.S. Presidents included. Hand marked speaker change locations are shown, along with T^2-BIC detected speaker/environmental changes.

These proposed improvements will now be integrated into the T^2-BIC formulation.

E. Implementation Issues

The implementation of the proposed algorithm is described as follows.

```
1. Initialization: set the working window $[S, E]$, with the start frame $S = 100$, and the end frame $E = 200$;
2. Compute the statistic $T^2(b)$ for possible $S < b < E$;
3. $\hat{b} = \text{argmax}_{S<b<E} T^2(b)$
4. Calculate $\Delta\text{BIC}(\hat{b})$ as in (10);
5. If $\Delta\text{BIC}(\hat{b}) > 0$
    $S = \hat{b} + 100$, $E = S + 100$
   else,
    $E = E + \Delta W$.
6. If $E \geq \text{Length}(audio)$, stop, otherwise go back to Step 2.
```

Here, the heuristic frame-skipping and dynamic window adjusting schemes are applied with Step 2 and 6. It can be seen that the test window continues to expand in a similar manner to that described in Chen *et al.* [6]. The advantage of this growing window over a sliding window of fixed size (e.g., [15]) is that more robust estimation can be obtained as further data is incorporated. However, this window expanding scheme also makes the segmentation search more costly, and having the disadvantage of error broadcasting (i.e., the detection errors in the previous window will impact subsequent windows).

Several observations concerning the T^2-Statistic should be noted here. The first concerns the choice of the common covariance matrices regarding the T^2-Statistic computation in (12). It is suggested from experiments that the common covariance matrices should be locally estimated using the data from two samples. Our segmentation experiments show that a local covariance matrix can allow for more accurate break point selection than a "universal" within-speaker matrix. We argue that the reason is that a local covariance matrix can capture more local speaker information than an averaged global one. Second, the feature used for computing the T^2-Statistic is a re-ordered 24-dimension feature set selected from the standard 39-dimension MFCC-based feature vector used for speech recognition.

F. Experimental Setup

The experimental results reported in this paper are all evaluated on audio streams sampled at 16 kHz. The feature sets used for both T^2-Statistic and BIC computation are identical, which is a 24-dimension feature set selected from the standard 39-dimension Mel Frequency Cepstral Coefficient (MFCC) vector used for speech recognition. The 24-dimensional feature set includes the frame energy, 12 static cepstral coefficients ($c_1, \ldots, c_{12}$), and the first 11 first-order derivatives ($\Delta c_1, \ldots, \Delta c_{11}$). We use a frame rate of 100 frames/s, where each frame is 20 ms in duration with an overlap of 50% between adjacent frames.

G. Experimental Results

To obtain a direct impression of how the proposed T^2-BIC segmentation scheme works on real world audio streams, we first evaluate the algorithm on sample audio streams from one of our ongoing project [10], the National Gallery of the Spoken Word (NGSW), which contains more than 60 000 hours of recordings from the past 100 years. An extracted sample is shown in Fig. 2, where a spectrogram is shown of a radio announcer providing an historical overview of U.S. history, with injected audio clips from six U.S. presidents. The sample audio clips are corrupted by various background noise sources which depend on the recording conditions for each president. As we can see from the spectrogram, the integrated presidential clips contain some short nonspeech breaks such as audience applause. The automatic detected speaker/acoustic changes are shown along with the hand-labeled change locations. On the average, speaker changes were identified within 90.7 ms, with some additional environmental changes automatically detected in presidential speeches due to audience applause. The results here suggest that the T^2-BIC based segmentation algorithm may be robust to background noise and also sensitive to nonspeech acoustic changes.

Evaluation of the proposed algorithm was further performed on a Sun Ultra-60 workstation using the DARPA Hub4 1997 Evaluation corpus which contains three hours of broadcast programs. The evaluation result is determined by comparing the automatically detected acoustic event changes with hand-segmentation provided by NIST. By definition, we assume that an acoustic event break point is true if the bias from the hand-labeled break point is less than 1 s. Moreover, we do not count frames where the same speaker simply changes his speaking style (for example, from "spontaneous" to "planned/read" speech) as break points. In addition, we do not ignore false alarms during music segments, which based on music type can possess a number of changing acoustic events. Table I presents competitive results between the original BIC algorithm and the proposed T^2-BIC segmentation algorithm with variable-sized windowing and skip-frame test, where the penalty factor in (10) is set to $\lambda = 1.4$. The table shows the false alarm rate for each algorithm, which reflects a 5.7% increase for the new method. It should be noted that an increase in the false alarm rate is less critical when compared to actually missing acoustic event changes (i.e., false positives are less troubling than false negatives). First, the former only produces a larger number of smaller segments for a particular acoustic event. Furthermore, the additional false positive acoustic breaks could be successfully merged during the clustering stage. On the other hand, missing an acoustic event change will produce an acoustic data distribution that is impure (e.g., most likely a bimodal pdf), and such a boundary is not recoverable. The resulting impure pdf will seriously affect the performance of any subsequent speech processing such as model adaptation or speaker tracking. Next, the percentage of missed acoustic turns is shown, which shows a 6.7% reduction using the proposed algorithm. While there was a 1.5% reduction in missed turns for acoustic events greater than 2 s in duration, there was an impressive 5.2% reduction for events of 2 s or less in duration. This would be an important factor in formulating an audio search algorithm for short duration segments, or partitioning incoming acoustic data for proper model adaptation for speech recognition.

The reduced number of missed turns achieved by T^2-BIC should contribute to the T^2-based pre-selection procedure due to its more reliable choice of break points for small sample size cases as pointed out in Section III-C. Also, variable-sized windowing enables the slow expanding process for small-sized windows, and thus lowers the miss rate of short duration segments.

Finally, in Table I, we show the computation CPU time required to perform segmentation analysis of the entire Hub4 1997 corpus. While traditional BIC requires 2160 min of CPU time, the new T^2-BIC scheme with variable-sized windowing and skip-frames tests is able to perform the task in 21 min, a computational speed improvement of 100. This represents the most significant advantage of the proposed algorithm.

TABLE I
SEGMENTATION RESULTS ON THE HUB4 1997 EVALUATION DATA

Algorithm	False Alarms	Missed turns		CPU Time Used (minutes)
		< 2 secs.	≥ 2 secs.	
BIC	10.8%	29.3%		2,160
		21.5%	7.8%	
T^2-BIC	16.5%	22.6%		21
		16.3%	6.3%	

IV. DISCUSSION

As discussed in Section III-G, the proposed audio parsing technique is effective in detecting acoustic changes and thus in segmenting and clustering for Broadcast News data, where acoustic conditions often change frequently. However, for some audio document, such as lectures from a single speaker, there may be no obvious acoustic change points occurring for long periods of time. The resulting very long audio segments will present difficulties for subsequent automatic transcription processing. To deal with such cases, an iterative T^2-BIC scheme can be considered as follows. Initially, the λ in (10) is set to a relatively high value, such as 1.5, to exclude pseudo segmentations when segmenting long audio files. If any large segments remain after the current pass of T^2-BIC segmentation, the value of the associated λ is decreased by 0.1 and another round of T^2-BIC is applied to this specific large segment. This process is repeated until no segments are longer than the upper limit that the subsequent recognizer can handle. A simple energy-based silence detector is then employed in a guided manner to locate possible silence frames near the T^2-BIC break points. These silence frames are picked as the final segmentation points. This silence location process helps reduce breaks within sentences or phrases.

T^2-BIC might also be used in conjunction with other segmentation methods. For example, the broad approach of HMM-based speech vs. nonspeech detection [11] is efficient to segment audio streams based on the turns between speech and silence. Further, T^2-BIC can be used to detect speaker or other acoustic (background noise or channel) changes during continuous speech segments. In this scheme, the silence-removed segments produced by HMM-based speech vs. nonspeech detection can help operate T^2-BIC on audio streams of significantly reduced duration, and obtain improved segmentation accuracy and speed.

Finally, further improvements may be obtained from a more careful selection of the feature vectors. For audio stream segmentation, the feature vector should be robust to the phoneme content delivered in the speech, but should be sufficiently sensitive to changes of speaker, channel conditions and acoustic environments. In other words, audio segmentation is rather different from regular tasks of speech recognition. Therefore, it is doubtful that if MFCC's are the most appropriate features for such tasks.

V. SUMMARY

In this paper, we have considered the formulation of an efficient algorithm for audio stream segmentation and clustering. The novel formulation is based on the T^2-statistic and Bayesian Information Criterion. It is shown that the proposed formulation can segment the 3 hours of Hub4 1997 evaluation data within 21 min of CPU time while only missing 22.6% of the acoustic event changes (compared to a previous BIC approach that missed 29.3% of the acoustic event changes and required 2160 min of CPU time). It is important to note that the ability to achieve reliable speaker change detection in an audio stream is dependent on the duration of each speaker turn. The proposed algorithm has been applied successfully for audio indexing tasks [10] for its notable efficiency.

Acknowledgment

The authors thank S. Chen for providing more extensive information about their original BIC segmentation results [6], and the anonymous reviewers who provided constructive feedback for this paper.

References

[1] H. Akaike, "A new look at the statistical identification model," *IEEE Trans. Automat. Contr.*, vol. 19, pp. 716–723, 1974.

[2] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, NY: Wiley, 1958.

[3] R. Bakis, S. Schen, P. Gopalakrishnan, R. Gopinath, S. Maes, and L. Polymenakos, "Transcription of broadcast news shows with the IBM large vocabulary speech recognition system," in *Proc. IEEE ICASSP-97: Int. Conf. Acoust., Speech, and Signal Proc.*, Munich, Germany, 1997, pp. 711–714.

[4] M. Basseville, "Distance measures for signal processing and pattern recognition," *Eur. J. Signal Process.*, vol. 18, no. 4, pp. 349–369, Dec. 1989.

[5] S. Chen, E. Eide, M. Gales, R. Gopinath, D. Kanevsky, and P. Olsen, "Recent improvements to IBM's speech recognition system for automatic transcription of broadcast news," in *Proc. DARPA Broadcast News Transcription Workshop*, 1999.

[6] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, Feb. 1998, pp. 127–132.

[7] W. Chou and W. Reichl, "Decision tree state tying based on penalized Bayesian information criterion," in *Proc. IEEE ICASSP-99: Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 345–348.

[8] J. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, no. 1–2, pp. 89–108, 2002.

[9] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Mag.*, vol. 11, no. 4, pp. 18–32, 1994.

[10] J. H. L. Hansen, B. Zhou, M. Akbacak, R. Sarikaya, and B. Pellom, "Audio stream phrase recognition for a national gallery of the spoken word: 'One small step'," in *ICSLP–2000: Int. Conf. Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 1089–1092.

[11] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system," in *Proc. IEEE ICASSP-02: Inter. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, 2002, pp. 53–56.

[12] A. Raftery, "Bayesian Model Selection in Social Research. Technical Report," Dept. of Statistics, Univ. Washington, Seattle, 1994.

[13] W. Reichl and W. Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," in *Proc. IEEE ICASSP-98: Int. Conf. Acoust., Speech, Signal Process.*, May 1998, pp. 801–804.

[14] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[15] M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, Feb. 1997, pp. 97–99.

[16] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S. J. Young, "Segment generation and clustering in the HTK broadcast news transcription system," in *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998, pp. 133–137.

[17] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in *Proc. Eurospeech '99*, 1999, pp. 679–682.

[18] S. Wegmann, P. Zhan, and L. Gillick, "Progress in broadcast news transcription at Dragon systems," in *Proc. IEEE ICASSP-99: Inter. Conf. Acoust., Speech, Signal Process.*, May 1999, 1912.

[19] P. Zhan, S. Wegmann, and L. Gillick, "Dragon systems' 1998 broadcast news transcription system for Mandarin," in *Proc. DARPA Broadcast News Transcription Workshop*, 1998.

[20] B. Zhou and J. H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," in *Proc. ICSLP–2000: Int. Conf. Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 714–717.

Bowen Zhou (M'03) received the B.S. degree from the University of Science and Technology of China in 1996, the M.S. degree from the Chinese Academy of Sciences in 1999, and the Ph.D. degree from the University of Colorado at Boulder in 2003, all in electrical engineering.

He was a Research Assistant with the Robust Speech Processing Group-Center for Spoken Language Research (RSPG-CSLR) during his graduate studies at the University of Colorado. He was an invited speaker for IBM User Interface Technology Student Symposium in November 2002. He joined the Department of Human Language Technologies at IBM Thomas J. Watson Research Center, Yorktown Heights, NY, in March 2003. His current research interest includes automatic speech recognition, natural language understanding, speech-to-speech machine translation, spoken information retrieval, and machine learning.

Dr. Zhou has served as a reviewer for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

John H. L. Hansen (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1988 and 1983, respectively, and the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ, in 1982.

He is Professor with the Departments of Speech, Language, and Hearing Sciences, and Electrical and Computer Engineering, at the University of Colorado at Boulder. He also serves as Department Chairman of Speech, Language and Hearing Sciences. In 1988, he established the Robust Speech Processing Laboratory (RSPL), which is now the Robust Speech Processing Group at the Center for Spoken Language Research (CSLR), which he co-founded and serves as Associate Director. He was a faculty member with the Departments of Electrical and Biomedical Engineering, Duke University, Durham, NC, for 11 years before joining the University of Colorado in 1999. In the fall of 2005, he joined the University of Texas at Dallas, Richardson, as Professor and Department Chair of Electrical Engineering, where he will establish the Center for Robust Speech Systems (CRSS). His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust recognition and training methods for spoken document retrieval in noise, accent, stress/emotion, and Lombard effect, and speech feature enhancement in hands-free environments for human-computer interaction. He has served as a technical consultant to industry and the U.S. Government, including ATT Bell Labs, IBM, Sparta, Signalscape, BAE Systems, ASEC, VeriVoice, and DoD in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech/speaker analysis. He is the author of more than 172 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals* (New York: IEEE Press, 2000) and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000, ISBN: 92-837-1027-4), and co-editor of *DSP for In-Vehicle and Mobile Systems* (Norwell, MA: Kluwer, 2004). He also organized and served as General Chair for ICSLP-2002: International Conference on Spoken Language Processing, Denver, CO, Oct. 2002.

Dr. Hansen was an invited tutorial speaker for IEEE ICASSP-95 and the 1995 ESCA-NATO Speech Under Stress Research Workshop, Lisbon, Portugal, the 2004 IMI-COE Symposium (Nagoya, Japan). He has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01: Research Study Group on Speech Processing, 1996–1999), Chairman for the IEEE Comm. and Signal Proc. Society of N.C. (1992–1994), Advisor for the Duke University IEEE Student Branch (1990–1997), Tutorials Chair for IEEE ICASSP-96, Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1998), Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), Member of the Editorial Board for *IEEE Signal Processing Magazine* (2001–2003). He has also served as guest editor of the Oct. 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was the recipient of a Whitaker Foundation Biomedical Research Award, an NSF Research Initiation Award, and has been named a Lilly Foundation Teaching Fellow for "Contributions to the Advancement of Engineering Education."