BLIND CHANGE DETECTION FOR AUDIO SEGMENTATION

Mohamed Kamal Omar, Upendra Chaudhari, Ganesh Ramaswamy

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
mkomar,uvc,ganeshr@us.ibm.com

ABSTRACT

Automatic segmentation of audio streams according to speaker identities, environmental and channel conditions has become an important preprocessing step for speech recognition, speaker recognition, and audio data mining. In most previous approaches, the automatic segmentation was evaluated in terms of the performance of the final system like the word error rate for speech recognition systems. In many applications like online audio indexing, and information retrieval systems, the actual boundaries of the segments are required. Therefore we present an approach based on the cumulative sum (CuSum) algorithm for automatic segmentation which minimizes the missing probability for a given false alarm rate. In this paper, we compare the CuSum algorithm to the Bayesian information criterion (BIC) algorithm, and a generalization of the Kolmogorov-Smirnov's test for automatic segmentation of audio streams. We present a two-step variation of the three algorithms which improves the performance significantly. We present also a novel approach that combines hypothesized boundaries from the three algorithms to achieve the final segmentation of the audio stream. Our experiments on the 1998 Hub4 broadcast news show that a variation of the CuSum algorithm significantly outperforms the other two approaches and that combining the three approaches using a voting scheme improves the performance slightly compared to using the a two-step variation of the CuSum algorithm alone.

1. INTRODUCTION

Many audio resources like broadcast news contain different kinds of audio signals like speech, music, noise, and different environmental and channel conditions. The performance of many applications based on these streams like speech recognition and audio indexing degrades significantly due to the presence of the irrelevant portions of the audio stream. Therefore segmenting the data to homogeneous portions according to type (speech, noise, music, etc.), speaker identity, environmental conditions, and channel conditions has become an important preprocessing step before using them [1], [2], [3], [4], [5], [6].

The previous approaches for automatic segmentation of audio data can be classified into two categories: informed and blind. Informed approaches include both decoder-based and model-based algorithms. In decoder-based approaches, the input audio stream is first decoded using speech and silence models [7]; then the desired segments can be produced by using the silence locations generated by the decoder. In model-based approaches, different models are built to represent the different acoustic classes expected in the stream and the input audio stream can be classified by maximum likelihood selection and then locations of change in the acoustic class are identified as segmental boundaries [3]. In both cases, models trained on the data representing all acoustic classes of interest are used in the automatic segmentation. The informed automatic segmentation is limited to applications where enough amount of training data is available for building the acoustic models. It can not generalize to unseen acoustic conditions in the training data. Also approaches based solely on speech and silence models mainly detect silence locations that are not necessarily corresponding to boundaries between different acoustic segments. In this paper, we will focus on blind automatic segmentation techniques which do not suffer from these limitations and therefore serve a wider range of applications.

Blind change detection avoids the requirements of the informed approach by trying to build models of the observations in a neighborhood of a candidate point under the two hypothesis of change and no change and using a criterion based on the log likelihood ratio of these two models for automatic segmentation of the acoustic data. Examples of this approach are [1], [2], [4], and [5]. In [6], the combination of an informed approach and a blind approach was considered. Most of the previous approaches had the goal of providing an input to a speech recognition, or a speaker adaptation system. Therefore they provided the evaluation of their systems based on comparisons of the word error rates achieved by using the automatic and the manual segmentation not the accuracy of the generated boundaries using the automatic segmentation [4], [3], [7]. Exceptions of this trend include [5] and when the main focus is data indexing like in [6].

In many applications like on-line audio indexing and information retrieval, the goal of the automatic segmentation algorithm is to detect the changes in the input audio stream and to keep the number of false alarms as low as possible. Unfortunately all of the current techniques for automatic blind segmentation like using the Kullback-Liebler distance, the generalized likelihood ratio distance [2], or the Bayesian information criterion [5] try to optimize an objective function that is not directly related to minimizing the missing probability for a given false alarm rate. If we define the missing probability as the probability of not detecting a change within a reasonable period of time of a valid change in the stream, then minimizing the missing probability is equivalent to minimizing the duration between the detected change and the actual change, namely the detection time.

In this paper, we will use a variation of the CuSum algorithm which minimizes the detection time for a given false alarm rate to automatically segment an input audio stream [8]. We will show that this variation significantly outperforms the Bayesian information criterion algorithm, and a generalization of the non-parametric Kolmogorov-Smirnov's test, [9]. We will also present a two-step

variation of the three algorithms which improves the performance significantly. Finally, we will introduce two approaches for combining the results of the three algorithms to achieve better and more robust segmentation.

In the next section, the three criteria used for automatic segmentation and the implementation of the corresponding algorithms are given. In section 3, the algorithm used in combining the output of the three systems to generate the final segmentation is presented. The experiments performed to evaluate the different strategies are described in section 4. Finally, Section 5 contains a discussion of the results and future research.

2. PROBLEM FORMULATION

The goal of our work is to search for a proper segmentation of a given audio signal such that each resulting segment is homogeneous and belongs to one of the different acoustic classes like speech, noise, and music and to a single speaker and a single channel. In this section, we will describe the actual implementation of the three algorithms and the assumptions made to make the estimation of the segmentation points efficient. In the three algorithms, each frame of data is represented by a feature vector of the cepstrum coefficients. Given an observation sequence of length n, the detection of a change is equivalent to accepting the hypothesis H_1 of change for time $r \leq n$ when testing it against the hypothesis H_0 of no change (i.e. $r \geq n$). The following algorithm is used to detect the change points in the input audio stream using any of the three criteria:

1. Initialize the first observation index f with zero and the last l with n_0.
2. Detect if there's a change using one of the three algorithms for the input sequence of observations.
3. If no change is detected set $l = l + n_0$.
   else set $f = r$ and $l = r + n$, where r is the location of the change detected.
4. If $(l - f > 3n_0)$ $f = l - 3n_0$.
5. If not end of audio stream go to 2.
6. End.

In the following, we will give details that are specific to the implementation of each algorithm.

2.1. Change Detection Using the CuSum Algorithm

Under the assumption that the sequence of the log likelihood ratios, $\{l_i\}_{i=1}^{n}$, is an i.i.d process, the CuSum algorithm is optimal in the sense of minimizing detection time for a given false alarm rate [10]. This assumption is valid for many interesting processes like some random processes that are modeled by Markov chains or some autoregressive processes [11]. In the CuSum algorithm, the likelihood ratio of the conditional PDFs of the observations under both the hypothesis H_1 of change for time $r \leq n$ and the hypothesis H_0 is estimated, then the maximum of the sum of the log likelihood ratio of a given sequence of observations is compared to a threshold to determine whether a boundary exists between two segments of the observation sequence. Given n observations, we compare

$$c_n = \max_r \sum_{k=r}^{n} l_k, \tag{1}$$

where l_k is the log likelihood ratio of the observation k to a threshold λ [8].

The CuSum algorithm assumes that the conditional PDFs of the observations under both the hypothesis H_1 of change for time $r \leq n$ and the hypothesis H_0 of no change (i.e. $r \geq n$) are known. In most automatic segmentation applications, this is not true. Therefore, we train a two-Gaussian mixture using the n observations in the given sequence. We initialize the two Gaussian components such that the mean of one of them corresponds to the mean of few observations in the beginning of the sequence of observations and the mean of the other corresponds to the mean of few observations in the end of the observations sequence. The automatic segmentation using the CuSum algorithm is then reduced to a binary hypothesis testing problem. The two hypothesis of this problem are

$$H_0 : z_{r^*}, \cdots, z_n \sim N(\mu_0, \Sigma_0),$$

and

$$H_1 : z_{r^*}, \cdots, z_n \sim N(\mu_1, \Sigma_1),$$

where $r^* = \arg\max_r \sum_{k=r}^{n} l_k$, l_k is the log likelihood ratio estimated using the two Gaussian components $N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$.

2.2. Change Detection Using the BIC Algorithm

The Bayesian information criterion is based on the log likelihood ratio of two models representing the two hypothesis of having two-class or one-class observation sequence. It adds a penalty term to account for the difference in the number of parameters of the two models [12]. The parameters of both models are estimated using the maximum likelihood criterion. Given n observations, the Bayesian information criterion BIC approach compares

$$b_n = \sum_{k=1}^{n} l_k - \frac{1}{2}(d_1 - d_2)\log(nM), \tag{2}$$

where d_1 and d_2 are the number of parameters of the two models, and M is the dimension of the observation vector [5], [12].

We implemented the BIC algorithm using the same assumptions given in [5]. So the conditional PDF of the observations under the hypothesis H_1 of change consists of two Gaussian PDFs. Both Gaussian PDFs are trained using maximum likelihood estimation. One of them is trained using the observations before the hypothesized boundary and the other is trained using observations after it. The conditional PDF of the observations under the hypothesis H_0 of no change is modeled with a single Gaussian PDF trained using maximum likelihood estimation from using all the n observations. Detecting a change at time r using the BIC algorithm is then reduced to a binary hypothesis testing problem. The two hypothesis of this problem are

$$H_0 : z_1, \cdots, z_n \sim N(\mu_0, \Sigma_0),$$

and

$$H_1 : z_1, \cdots, z_{r-1} \sim N(\mu_1, \Sigma_1);$$
$$z_r, \cdots, z_n \sim N(\mu_2, \Sigma_2),$$

where $N(\mu_0, \Sigma_0)$ is the Gaussian model trained using all the n observations and $N(\mu_1, \Sigma_1)$ is trained using the first r observations and $N(\mu_2, \Sigma_2)$ is trained using the last $n - r$ observations. Since the model of the conditional PDF under the hypothesis H_1 of change depends on the location of the change, reestimation of the

model parameters is required for each new hypothesized boundary within the sequence of observations of length n. This problem is avoided in our CuSum algorithm implementation, as in this case both models are independent of the location of the hypothesized boundary.

2.3. Change Detection Using the Kolmogorov-Smirnov's Test

The Kolmogorov-Smirnov's test is a nonparametric test of change in the input data [9]. It compares the maximum of the difference of the empirical CDFs of the data before and after the hypothesized change point to a threshold to determine whether this point is a valid boundary point between two distinct classes. In other words, to test the validity of a boundary at observation k, the test compares

$$S_n = \sup_z |F_k(z) - G_{n-k}(z)|, \tag{3}$$

where

$$F_k(z) = \frac{1}{k}\sum_{j=1}^{k}\Theta(z - z_j), \tag{4}$$

$$G_{n-k}(z) = \frac{1}{n-k}\sum_{j=k+1}^{n}\Theta(z - z_j), \tag{5}$$

and $\Theta(.)$ is the unit step function, to a threshold α [8].

The Kolmogorov-Smirnov's test was designed for one-dimensional observations. To generalize for observation vectors of dimension M, we assume that the elements of the observation vector are statistically independent and replace the criterion of the Kolmogorov-Smirnov's test with the following one

$$S_n = \sup_m \sup_s |F_k^m(z_s^m) - G_{n-k}^m(z_s^m)|, \tag{6}$$

where

$$F_k^m(z_s^m) = \frac{1}{k}\sum_{j=1}^{k}\Theta(z_s^m - z_j^m), \tag{7}$$

and

$$G_{n-k}^m(z^m) = \frac{1}{n-k}\sum_{j=k+1}^{n}\Theta(z_s^m - z_j^m), \tag{8}$$

for $m = 1, \cdots, M$, and the range of values of each dimension is quantized to a fixed number of bins, $\{z_s^m\}_{s=1}^S$ to be used in calculating the empirical CDFs.

3. AN ALGORITHM FOR COMBINING THE THREE SYSTEMS

Since the three approaches for automatic segmentation of the audio data described before use different criteria and different modeling of the conditional PDFs of the observations under both hypothesis of valid change or no change. It is reasonable to expect these algorithms to employ complementary information for automatic change detection and therefore combining the three approaches can improve the overall performance and robustness of the automatic change detection system. To combine the three approaches, we use each of them separately to generate a set of potential change points. Then the values of the three measures used in the three algorithms for detection of the change are evaluated at every point of the three sets. Then based on either a voting scheme or a likelihood ratio test of two models trained on the values of the three measurements of manually segmented data near and far from a valid change respectively, the set of valid change points are selected from the collection of the three sets. The steps of the algorithm are

1. Initialize the first observation index f with zero and the last l with n_0.
2. Detect if there's a change using the three algorithms for the input sequence of observations.
3. Generate a list of the candidate points from the union of the output of the three algorithms.
4. Calculate the values of the measurements of the three algorithms at every point of the candidate list.
5. Remove the invalid changes from the list using a voting scheme or a likelihood ratio test.
6. If the candidate list is empty set $l = l + n_0$.
   else set $f = r$ and $l = r + n_0$, where r is the location of the last change in the candidate list.
7. If $(l - f > 3n_0)$ set $f = l - 3n_0$.
8. If not end of audio stream go to 2.
9. End.

4. EXPERIMENTS

We tested the three approaches of the CuSum algorithm, the BIC algorithm, and the generalized Kolmogorov-Smirnov's test on the automatic segmentation of the broadcast news Hub4 1998 evaluation data. The data is sampled at 16 KHZ and windowed to frames of 20 ms duration with overlap of 10 ms. Nineteen cepstrum coefficients are calculated for each frame. We selected the initial number of observations to be tested for a change, n_0, to be 300 frames for the CuSum and the BIC algorithms and 400 frames for the generalized Kolmogorov-Smirnov's test. For the generalized Kolmogorov-Smirnov's test, we divided the range of the values of each dimension of the observation vector to five bins and the two empirical CDF's are compared in each of these five bins to find the maximum. The size of the testing data is 5.5 hours which contained approximately 625 homogeneous segments. For the three algorithms, we tried adding a verification step in which the objective criterion is calculated at the candidate change points using the knowledge obtained in the first step of the previous and the next change point and then compared to a new threshold. The thresholds in our experiments are chosen empirically to minimize, on 2-hours held-out data, the objective function

$$O = MP + 0.1 * FA, \tag{9}$$

where MP is the missing probability and FA is the false alarm probability, and under the constraint that the false alarm probability is less than 0.1. The missing probability is calculated by assuming a change is missed, if no change was detected within one second from it. Table 1 shows the results for one-step automatic segmentation using the three algorithms. It shows that our implementation of the CuSum algorithm significantly outperforms the

Algorithm	Missing Prob.	FA Prob.
Kolmogorov-Smirnov	37.6	8.6
BIC	35.9	9.4
CuSum	27.9	8.9

Table 1. One-Step Automatic Segmentation

Algorithm	Missing Prob.	FA Prob.
Kolmogorov-Smirnov	35.9	8.3
BIC	32.6	9.1
CuSum	16.8	4.9

Table 2. Automatic Segmentation with a Verification Step

BIC and generalized Kolmogorov-Smirnov's tests. The BIC algorithm works better than the generalized Kolmogorov-Smirnov's test, although the latter tends to give lower false alarm probability. Table 2 shows that adding a verification step after the initial segmentation improves significantly the performance of both the BIC and the CuSum algorithms.

We tested also the combination algorithm described in the previous section. Table 3 shows that the combination based on the voting scheme significantly outperforms that based on the likelihood ratio test using models trained using manually segmented data. It shows also that this voting scheme slightly outperforms the best single automatic segmentation system which uses our implementation of the CuSum algorithm with a verification step.

5. RESULTS AND DISCUSSION

In this paper, we examined three approaches for blind automatic segmentation of audio streams. Our implementation of two of these approaches, namely the CuSum algorithm and the generalized Kolmogorov-Smirnov test, is novel and for the first time applied to automatic segmentation of audio streams. We also presented a two-step variation of the algorithms which improved the perfomnace significantly. Finally, we presented also two approaches for combining the scores from the three systems to achieve better performance and more robust segmentation of the audio stream.

The results for our tests show that our implementation of the CuSum algorithm significantly outperforms other blind automatic segmentation techniques of audio data like the BIC algorithm and the generalized Kolmogorov-Smirnov approach. It has the advantage also of not having to reestimate the conditional models at each potential segmentation point within the same window. The better performance can be attributed partially to the fact that the CuSum algorithm tries to minimize the detection time for a given false alarm rate. This objective is more suited to automatic segmentation applications than the objectives of both the BIC and the generalized Kolmogorov-Smirnov approaches.

Algorithm	Missing Prob.	FA Prob.
CuSum	16.8	4.9
Voting Combination	15.6	7.3
Likelihood Ratio Combination	16.7	6.9

Table 3. Automatic Segmentation using Systems Combination

Combining the scores of the three systems using voting schemes is significantly better than the likelihood ratio approach using models trained near the change points and others far from the change points. Combination using voting is slightly better than using the CuSum algorithm alone with a verification step but at the expense of increasing the processing time of the data.

Further investigation of the effect of the type of the input features and the models for the conditional PDFs on the automatic segmentation performance will be our main goal in future research. We will consider also many other alternatives for combining the scores of the three algorithms.

6. REFERENCES

[1] H. Beigi, S. Maes "Speaker, Channel, and Environment Change Detection," *Proceedings of the World Congress on Automation*, pp. 18–22, 1998.

[2] H. Gish, N. Schmidt, "Text-independent speaker identification," in *IEEE Signal Processing Magazine*, pp. 18–21, 1994.

[3] J. L. Gauvain, L. Lamel, "Audio Partitioning and Transcription for Broadcast Data Indexation," in *Multimedia Tools and Applications*, vol. 14, no.2, pp. 187–200, 2001.

[4] M. Siegler, U. Jain, B. Ray, R. Stern, "Automatic Segmentation, Classification, and Clustering of Broadcast New Audio," in *DARPA Speech Recognition Workshop Proc.*, pp. 97–99, 1997.

[5] S. S. Chen and P. S. Gopalakrishnan, "Speaker Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion," in *DARPA Speech Recognition Workshop Proc.*, 1998.

[6] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, "Strategies for Automatic Segmentation of Audio Data," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1423–1426, 2000.

[7] B. Ramabhadran, J. Huang, U. Chaudhari, G. Iyengar, H. J. Nock, "Impact of Audio Segmentation and Segment Clustering on Automated Transcription Accuracy of Large Spoken Archives," in *Proc. of EuroSpeech*, pp. 2589–2593, 2003.

[8] M. Basseville, I Nikiforov, *Detection of Abrupt Changes-Theory and Application*, Prentice-Hall, April 1993.

[9] J . Deshayes, D Picard, " Off-line Statistical Analysis of Change-Point Models Using Non-Parametric and Likelihood Methods" in *Detection of Abrupt Changes in Signals and Dynamical Systems*, Springer-Verlag, 1986.

[10] A. N. Shirayev, "The Problem of the Most Rapid Detection of a Disturbance in a Stationary Process," in *Soviet Math. Dokl.*, no. 2, pp. 795–799, 1961.

[11] G. V. Moustakides, "Quickest Detection of Abrupt Changes for a Class of Random Processes," in *IEEE Transactions On Information Theory*, vol. 44, no. 5, September 1998.

[12] R. E. Kass, A. E. Rafetry *Bayes Factors,* Technical Report no. 254, Department of Statistics, University of Washington, July 1994.