METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation

Shih-sian Cheng and Hsin-min Wang

Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
{sscheng, whm}@iis.sinica.edu.tw

Abstract

This paper presents a hybrid approach for audio segmentation, in which the metric-based segmentation with long sliding windows is applied first to segment an audio stream into shorter sub-segments, and then the divide-and-conquer segmentation is applied to a fixed-length window that slides from the beginning to the end of each sub-segment to sequentially detect the remaining acoustic changes. The experimental results on five one-hour broadcast news shows show that our approach outperforms the existing metric-based and model-selection-based approaches.

1. Introduction

Audio segmentation, which aims to find out acoustic changes within an audio stream (e.g. the change boundary between two speakers), has received more and more research efforts from researchers because of its importance to the pre-processing of audio recordings for applications such as audio indexing, automatic transcription, speaker tracking, etc.

Bayesian Information Criterion (BIC), which is a model selection criterion in the statistics literature, has been widely adopted for the detection of acoustic changes in an audio stream [1-7]. In the model-selection-based approach proposed by Chen [1], BIC was applied to test whether there is an acoustic change within a detection window. If there is a change point detected, the detection window is moved to this point to search the next change point. Otherwise, the detection window grows in size to have a larger search range. However, with the growing window, Chen's BIC scheme suffers from the high computation cost, especially for the audio stream composed of long homogenous segments. Two improved BIC-based approaches were therefore proposed to speed up the detection process in [4, 5]. To improve the performance, in [4], a variable window scheme and some heuristics were applied to the BIC framework while, in [5], the T^2 statistic was integrated with the BIC criterion.

On the other hand, metric-based segmentation has been found very fast but less effective than model-selection-based segmentation. Hence, we proposed a sequential metric-based approach by integrating the BIC into the modified metric-based segmentation framework in [9]. This approach performed nearly as well as the model-selection-based approaches at a lower computation cost. In this paper, we further study the divide-and-conquer (DAC) segmentation, which detects acoustic changes in a top-down manner, and develop a hybrid approach, in which the metric-based segmentation with long sliding windows is applied first to segment an audio stream into shorter sub-segments, and then the DAC segmentation is applied to a fixed-length window that slides from the beginning to the end of each sub-segment to sequentially detect the remaining acoustic changes. The hybrid approach is very efficient, and the experimental results on broadcast news data show that it outperforms the existing model-selection-based and metric-based approaches.

The rest of this paper is organized as follows: We first review the metric-based segmentation and the model-selection-based segmentation in Section 2. Then, the proposed approach for audio segmentation is introduced in Section 3. Finally, the experimental results are presented in Section 4, and conclusions are made in Section 5.

2. Reviews

2.1. The metric-based audio segmentation

In the metric-based audio segmentation, the dissimilarity of two consecutive windows with the same length is computed along the audio stream to form a distance curve. This distance curve was often low-pass filtered and the locations of peaks were chosen as acoustic change points by heuristic thresholds. The distance measures, such as Kullback-Leibler distance (KL, KL2) and Generalized Likelihood Ratio (GLR)[2], often come from the statistical modeling framework. The feature vectors in each of these two consecutive windows are assumed to follow some probability density (usually, the multivariate Gaussian), and the distance is represented by the dissimilarity of these two densities. Figure 1 shows the procedures of the metric-based segmentation.

2.2. Audio segmentation via BIC

Given a data set $X = \{x_1, x_2, \cdots, x_n\} \subset R^d$ and a model set $M = \{M_1, M_2, \cdots, M_k\}$, the model selection problem is to choose the model that best fits the distribution of X. Bayesian Information Criterion (BIC) is a model selection criterion and the BIC value of M_i is defined as:

$$BIC(M_i) = \log pr(X \mid \hat{\Theta}_i) - \frac{1}{2}\lambda \#(M_i) \log n, \qquad (1)$$

where $\lambda = 1$, $pr(X \mid \hat{\Theta}_i)$ is the maximum likelihood of X for model M_i, and $\#(M_i)$ is the number of parameters of M_i. The model with the highest BIC value will be selected. In real applications, the value of λ can be tuned as needed.

2.2.1. *One-change-point- detection*

In the one-change-point detection procedure proposed by Chen[1], it was assumed that there was at most one change point in X and the following hypothesis test of Gaussian process was performed sequentially on $x_i, i = 1, 2, \cdots, n$:

$$H_0 : x_1, x_2, \cdots x_n \sim N(\mu, \Sigma)$$
$$H_1 : x_1, x_2, \cdots x_i \sim N(\mu_1, \Sigma_1); x_{i+1}, \cdots, x_n \sim N(\mu_2, \Sigma_2).$$

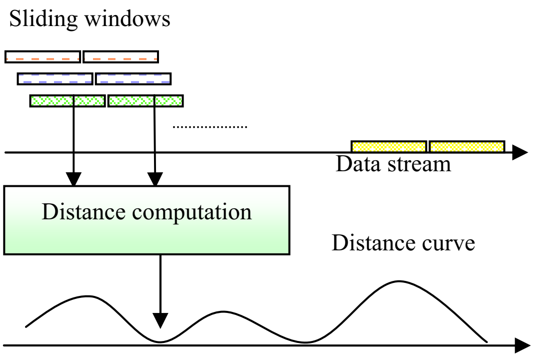


Figure 1: The procedures of metric-based segmentation.

The BIC difference of these two hypotheses at x_i is $\Delta BIC(i) = BIC(H_1) - BIC(H_0)$. If $\{\max_i \Delta BIC(i)\} > 0$, the time index corresponding to the maximal ΔBIC value was selected as the change point; otherwise, there was no change point in *X*.

2.2.2. Detection of multiple change points

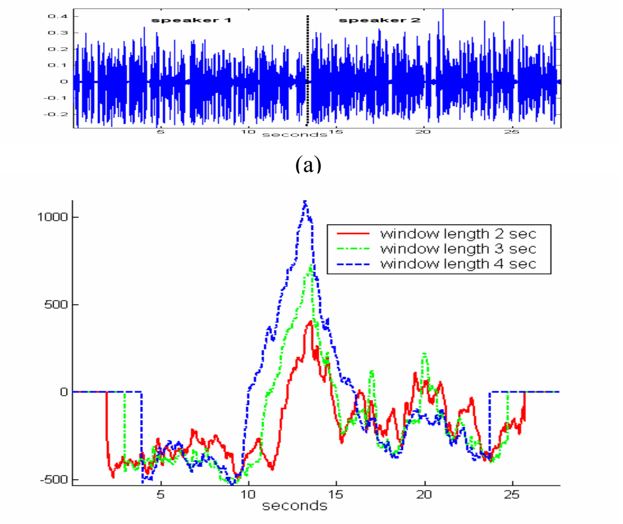
As for detecting multiple change points in an audio stream, the one-change-point detection process was sequentially applied in a growing detection window [1]. Starting from the beginning of the audio stream, a short detection window was used at first. If there was a change point detected, a new detection window was put at this point to search the next change point. Otherwise, the current detection window was enlarged, and the one-change-point detection process was applied again.

3. The proposed approach

3.1. The metric-based audio segmentation via BIC

In the metric-based segmentation, usually, a short window (typically 2 seconds) is adopted for dealing with homogenous segments of various lengths. In the previous studies, this approach has been found fast but less effective because the window is too short to obtain robust distance statistics. The threshold was usually designed to get a lower missed detection rate, at the cost of higher false alarm rate, because these falsely-detected change points could be eliminated by performing clustering and merging or other compensation approaches on the segments yielded by the initial segmentation[2][7]. Although the metric-based segmentation inevitably has a high missed detection rate when a long window is used, it has the advantage of low false alarm rate because the long window provides sufficient data samples for the distance measurement. In our hybrid approach, the metric-based segmentation with long sliding windows is first applied to detect the change points corresponding to long homogenous segments. Though, the change points that do not hold the above condition are very likely to be missed in this step, they can be detected by the following DAC segmentation.

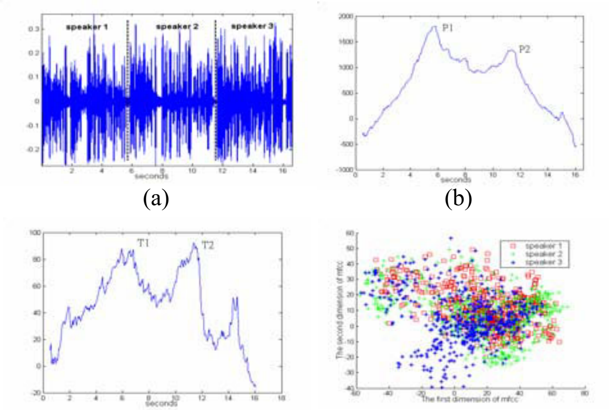
Following our previous work in developing a sequential metric-based segmentation approach [9], we use the ΔBIC value as the distance measure, in which the hypothesis H_0 models the feature vectors of the two consecutive windows as a multivariate Gaussian while each window is modeled by a distinct Gaussian in H_1. The local peaks of the ΔBIC curve with a peak width larger than 0.4 times of the window length are detected as the change points. The width of a peak is defined as the time span of its neighboring points with ΔBIC larger than 0. Figure 2 shows an example of metric-based segmentation using 2-, 3-, and 4-second sliding windows and ΔBIC as the distance measure. We can see that the ΔBIC curve is more reliable when a longer window is used.



(a)

(b)

Figure 2: (a) An audio stream consisting of two different speakers. (b) The distance (ΔBIC) plot of the metric-based segmentation on the audio stream in (a) with 2-, 3-, and 4-second sliding windows.



(a)

(b)

(c)

(d)

Figure 3: (a) An audio stream composed of three different speaker segments. (b) The ΔBIC plot based on 24-dimensional MFCC vectors. (c) The ΔBIC plot based on 2-dimensional MFCC vectors. (d) The distribution of the 2-dimensional MFCC vectors.

3.2. The divide-and-conquer audio segmentation via BIC

What happens if we apply the one-change-point detection process proposed by Chen [1] to an audio stream consisting of multiple homogeneous segments from different sources? The answer is one of the change points (perhaps the most significant one) will be found. The basic idea is illustrated in Figure 3. Figure 3(a) depicts an audio segment that consists of three different speakers. By applying the one-change-point detection to this audio segment using 24-dimensional MFCC vectors as speech feature vectors, we got the ΔBIC plot in Figure 3(b). The two significant peaks, P1 and P2, tell us the following message:

Modeling {Speaker 1} and {Speaker 2, Speaker 3} as one Gaussian respectively or modeling {Speaker 1, Speaker 2} and {Speaker 3} as one Gaussian respectively is better than modeling all the three speakers as one Gaussian.

This message conforms to the assumption that the feature vectors of each speaker respectively distribute as a multivariate Gaussian distribution in the feature space. Figure 3(c) depicts the ΔBIC plot derived using 2-dimensional MFCC vectors, and Figure 3(d) plots the 2-dimensional MFCC vectors of these three speakers. We can see from Figure 3(d) that it is better to respectively model {Speaker 1, Speaker 2} and {Speaker 3} as one Gaussian, rather than modeling all these three speakers as one Gaussian. This observation conforms to the peak T2 in Figure 3(c).

According to the one-change-point detection process, the audio stream in Figure 3(a) is divided into two sub-segments at the time index corresponding to P1, and then the second change points will be detected when the one-change-point detection process is applied to the second sub-segment. In this way, we can design a recursive divide-and-conquer procedure to detect the change points of an audio stream. The details of the proposed DACSeg algorithm are illustrated in Figure 4.

```
procedure DACSeg($X$)
//input: $X$ , an audio stream.
//output: CP, a set of change points of $X$.
Begin
   1. detect whether there is an acoustic change point in $X$
      by performing one-change-point detection;
   2. if there is no change point in $X$ or the length of $X$ is
      less than two seconds
        CP = $\varnothing$ ; // empty set
        return;
      else
        let $t$ be the change point in $X$ and divide $X$ into two
        segments, $Y$ and $Z$, at time $t$;
        CP_$Y$ = DACSeg($Y$);
        CP_$Z$ = DACSeg($Z$);
        CP = $t \cup$ CP_$Y \cup$ CP_$Z$;
End
```

Figure 4: The divide-and-conquer segmentation procedure.

In our experience, the one-change-point detection process is more likely to produce a false alarm detection in a long homogenous segment than in a short one with the same penalty factor (i.e., λ) of BIC. This is because a single multivariate Gaussian is not flexible enough to well model the huge samples in a long homogenous segment. To avoid this situation, we can perform the DACSeg procedure sequentially in a fixed-length detection window that slides from the beginning to the end of the audio stream. We illustrate this procedure, which is called SeqDAC in the rest of this paper, in Figure 5. If the length of the target audio stream is not longer than the fixed-length detection window, the SeqDAC degenerates to the DACSeg, which detects the acoustic changes of an audio stream in a global way.

3.3. The hybrid approach

In detecting the acoustic changes of a long audio stream (e.g. a one-hour broadcast news show), we can first divide it into shorter sub-segments using the metric-based segmentation with long sliding windows. This step will produce a very low false alarm rate. Then, we can apply the SeqDAC procedure to further detect the acoustic changes in each sub-segment.

```
procedure SeqDAC($X$, $W$, $\eta$)
//input: $X$, an audio stream.
        $W$, the length of the detection window.
        $\eta$, the detection window will shift $W\eta$ if no change
          point is detected.
//output: CP, a set of change points of $X$.
Begin
   CP = $\varnothing$ ;
   timeBeg=1; timeEnd = timeBeg+$W$;
   repeat:
      1. if timeEnd > length of $X$
           timeEnd = length of $X$;
           detectionWind = $X$ (timeBeg : timeEnd);
           CP_temp = DACSeg(detectionWind);
           CP=CP $\cup$ CP_temp ;  goto End;
     2. detectionWind = $X$ (timeBeg : timeEnd);
        CP_temp = DACSeg(detectionWind );
        if CP_temp is empty // no change detected
           timeBeg = timeBeg + $W\eta$;
           timeEnd = timeBeg+$W$;
        else
          let timeBeg be the latest time index in CP_temp;
          timeEnd = timeBeg+$W$;
          CP=CP $\cup$ CP_temp ;  goto 1;
 End
```

Figure 5: The sequential DAC segmentation procedure.

4. Experiments

4.1. Data description and parameterization

We used five one-hour shows selected from the MATBN2002 Mandarin Chinese broadcast news corpus [8] for evaluation. The acoustic change points associated to manually annotated speaker turns, speech-silence boundaries and changes of background sounds are used as the ground truth. There are 2245 such change points in total.

About the parameterization of the evaluation data, a 20ms Hamming window shifted with a step of 10ms is used to evaluate 24 mel-frequency cepstral coefficients (MFCCs) as the speech features.

4.2. Performance evaluation

The detection tasks can be viewed as involving a tradeoff between two error types: missed detection (MD) and false alarm (FA). In this study, an actual change point t is considered missed if there is no detected change point within [t-1,t+1](a 2-second window centered on t), and a detected change point $\hat{t}$ is counted as a false alarm if there is no actual change point within$[\hat{t}-1,\hat{t}+1]$. Following the definition given in [2], the missed detection rate (MDR) and false alarm rate (FAR) are computed by

$$\text{MDR} = 100 \times \frac{\text{number of MD}}{\text{number of actual change points}} \%$$

$$\text{FAR} = 100 \times \frac{\text{number of FA}}{\text{number of actual change points} + \text{number of FA}} \%$$

4.3. Experimental results

We have conducted experiments based on five one-hour broadcast news shows to compare our approach with several existing approaches, including the metric-based approach, the model-selection-based approaches proposed by Chen[1] and Zhou[5], and our previously proposed sequential metric-based approach[9].

In the metric-based method, both KL2 and GLR were used as the distance measure, the length of the sliding windows was two seconds, and the decision mechanism proposed by Delacourt[2] was adopted, in which all the time indices corresponding to "significant" peaks were considered as change points. Different thresholds were tested to yield the missed detection rate and false alarm rate pairs.

In the proposed method, each broadcast news show was pre-segmented into shorter sub-segments by the metric-based segmentation with two consecutive sliding windows of four seconds. The distance measure is ΔBIC with λ=1.2 and the peak selection criterion discussed in Section 3.1 is adopted. In our preliminary experiments, this criterion slightly outperformed Delacourt's criterion. The metric-based pre-segmentation yielded a high missed detection rate of 46.9% at a low false alarm rate of 2.4%. After the pre-segmentation, the SeqDAC procedure, in which W is twenty seconds and η is 0.25, was applied to further detect the change points in the sub-segments. To yield missed detection rate and false alarm rate pairs, the penalty factor, λ, in BIC was varied from 0.5 to 1.4 with a step of 0.1 in Metric-SeqDAC and the model-selection-based approaches, while λ varied from 0.5 to 1.2 with a step of 0.05 in the sequential metric-based approach.

Figure 6 depicts the performance curves of the proposed approach and the other existing approaches. We can see an obvious performance gap between the conventional metric-based approaches ("Metric-based-KL2" and "Metric-based-GLR") and the model-selection-based approaches (Chen's and Zhou's approaches). The performance of our previously proposed sequential metric-based approach is between that of the metric-based approaches and the model-selection-based approaches. The proposed approach (Metric-SeqDAC) obviously outperforms all the other approaches.

The proposed approach is very efficient for the broadcast news task because of two reasons: First, many of the change points have been detected in the metric-based pre-segmentation step, and this step runs very fast. Second, in broadcast news data, usually, there are only several change points in a detection window that is not too long (e.g. 20 seconds). Therefore, the SeqDAC will not be the bottleneck of the hybrid approach, since the recursive tree of the DACSeg procedure won't go deep.

5. Conclusions

We have proposed a hybrid approach for audio segmentation. Our approach first applies the metric-based segmentation with long sliding windows to pre-segment a long audio stream into shorter sub-segments, and then applies the sequential divide-and-conquer segmentation to each sub-segment to detect the remaining change points. Instead of searching the acoustic change points in a bottom-up manner, which was widely adopted in previous studies, the divide-and-conquer procedure searches acoustic change points in a top-down manner. Our approach is very efficient and the experimental results on broadcast news have demonstrated the good performance.

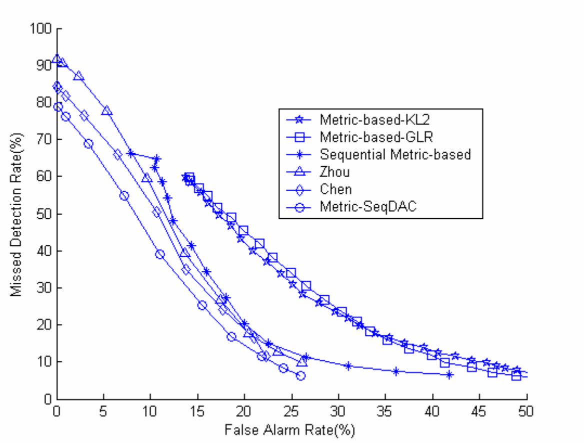


Figure 6: The performance curves of the proposed approach and the other existing approaches.

6. References

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.*

[2] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, 32(1-2):111-126, 2000.

[3] P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia, "On the use of the Bayesian Information Criterion in multiple speaker detection," *Proceedings of Eurospeech2001.*

[4] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian Information Criterion," *Proceedings of Eurospeech1999.*

[5] B. W. Zhou, and John H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion," *Proceedings of ICSLP2000.*

[6] M. Cettolo and M. Vescovi, "Efficient audio segmentation algorithms based on the BIC," *Proceedings of ICASSP2003.*

[7] X. Zhong, M. Clements, and S. Lim, "Acoustic change detection and segment clustering of two-way telephone conversation," *Proceedings of Eurospeech2003.*

[8] H. M. Wang, "MATBN 2002: a Mandarin Chinese broadcast news corpus," *Proceedings of ISCA \& IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.*

[9] S. S. Cheng and H. M. Wang, "A sequential metric-based audio segmentation method via the Bayesian Information Criterion," *Proceedings of Eurospeech2003.*