AUDIO SEGMENTATION VIA TRI-MODEL BAYESIAN INFORMATION CRITERION

Yunfeng Du[1,2], Wei Hu[2], Yonghong Yan[1], Tao Wang[2], Yimin Zhang[2]

[1] Institute of Acoustics, Chinese Academy of Sciences, Beijing, P.R.China
[2] Intel China Research Center, Beijing, P.R.China
ydu@hccl.ioa.ac.cn, wei.hu@intel.com

ABSTRACT

This paper addresses the problem of audio segmentation in practical media (e.g. TV series, movies and etc.) which usually consists of segments in various lengths with quite a portion of short ones. An unsupervised audio segmentation approach is presented, including a segmentation-stage to detect potential acoustic changes, and a refinement-stage to refine these candidate changes by a tri-model Bayesian Information Criterion. Experiments show that the proposed approach has good detectability of short segments and the novel tri-model BIC effectively improves the overall segmentation performance.

Index Terms **-** audio segmentation, acoustic change detection, tri-model Bayesian Information Criterion, data balance ratio

1. INTRODUCTION

Audio segmentation is also often called as acoustic change detection [1] which partitions the audio stream into homogenous segments by detecting changes of speaker identity, acoustic class or environmental condition. It is an essential step for audio clustering and classification as well as speaker clustering and tracking in many circumstances, thus plays an important role in various applications such as multimedia indexing, spoken document retrieval and speech recognition.

The current approaches of audio segmentation can be categorized into two major groups: the model based approach initializes a set of models for different acoustic classes from training corpus to classify the input audio stream so as to locate the changes [2]. However, in many cases, the pre-knowledge of speakers and acoustic classes are often not available. Therefore, unsupervised metric-based approaches are desirable in many applications. In metric-based approach, changes are determined by threshold on the basis of a distance computation for the input audio stream. Most of the distance measures come from statistical modeling framework, e.g. Kullback-Leibler distance, generalized likelihood ratio and others [3]. The hybrid of these two approaches is also applied. It incorporates a metric-based method as pre-segmentation and a clustering procedure to obtain the training data, and then performs a model based re-segmentation [4].

The BIC-based approach is first proposed in [1], which utilizes a sliding variable-size window to determine acoustic changes based on a model-selection criterion. It can be recognized as a special metric-based approach since the penalty term of BIC operates as a threshold. Based on BIC, a two-stage segmentation is proposed in [5], which first segments the audio stream by a distance measure, then refines these changes by BIC sequentially. This two-stage framework consisting of segmentation and refinement (false alarm compensation) demonstrates the effectiveness, and has been widely adopted by many works in recent years [6, 7, 8, 9].

In the previous works [1, 3, 4, 5, 6, 7, 8], the evaluated audio often consists of relative long acoustic segments (>2s or 3s), and the short segments (1-3s) are often neglected because they are difficult to be detected. However, the presences of short segments are usually frequent in practical media such as TV series, movies, phone conversations, and even broadcast news e.g. interview. Thus, to detect the short segments is a main challenge when applying the audio segmentation into real applications.

In this paper, an unsupervised audio segmentation approach is presented with emphasizing to detect the short segments. A novel tri-model BIC is also proposed. Section 2 details the algorithm. Section 3 introduces the theory of the proposed tri-model BIC. Experiments and analysis are presented in section 4. Section 5 gives the conclusion.

2. SYSTEM FRAMEWORK

The proposed approach is mainly consisted of five modules: pre-processing, feature extraction, segmentation, refinement, and post-processing. The pre-processing and post-processing modules are alternative regarding to the property of the input audio data which is down-sampled into 16 kHz with uniform format of 16bits, mono channel.

In analog to [5], the algorithm is mainly based on a two-stage analysis: the first-stage is a metric-based segmentation which uses a distance computation to determine the candidates of acoustic-changes in the audio stream; and the second-stage is a criterion-based refinement which utilizes a tri-model Bayesian Information Criterion to validate or discard these candidates.

Before feature extraction (a feature set of 14 MFCCs with log-energy computed in 20ms frame with 10ms overlapping is applied), a VAD (Voice Activity Detection) procedure can be used as pre-processing to remove silence or breathing from the audio data in order to facilitate the segmentation procedure. However, this will lead to additional computational cost, so it is alternative regarding to the proportion of silence in the audio data as well as the needs of the oriented applications.

If the pre-processing module is applied, a corresponding post-processing module should implement to insert the silence back in order to align the time. If a computed-change is close to a silent

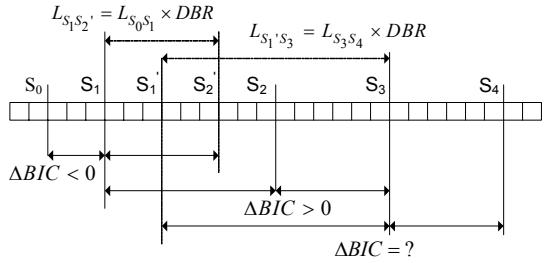

Fig.1 Algorithm of BIC refinement in the second stage

part within a certain distance (0.5s), it is moved to the nearest boundary of that silence segment.

2.1. Segmentation

The first-stage relies on a computation of distance between two adjacent analysis windows of the same size (1s or 2s) shifted by a fixed step (0.1s) along the whole feature data of the input audio. In each analysis window, a single multi-dimensional Gaussian process is estimated from the feature data via maximum likelihood estimation. The symmetric Kullback-Leibler (KL2) distance [10] is chosen here. This process results in a graph of distances with respect to time. The graph is smoothed by a low-pass filtering operation, and then all the "significant" local maxima that represent potential change points are acquired by searching the graph.

The threshold condition proposed in [9] is utilized: A local maximum is regarded as significant if it satisfies the condition $|max-min_l|>\alpha\sigma$ or $|max-min_r|>\alpha\sigma$, where α is a fraction, σ is the standard deviation of the distance graph, min_l and min_r are the left and the right minima around the peak *max* respectively; and no higher local maximum near it within a certain distance (0.5s).

2.2. Refinement

In the second-stage, a ΔBIC value is computed for each potential change point detected in the first-stage to validate or discard this point. For example in [5], given $\{s_1, ..., s_n\}$ is the set of candidate change points found in the first stage, a ΔBIC value is computed for each pair of windows $[s_{i-1}, s_i]$, $[s_i, s_{i+1}]$. If the value is negative, a change point is identified at time i. If not, the point s_i is discarded from the candidate set, so that the ΔBIC value is now computed for the new pair of windows $[s_{i-1}, s_{i+1}]$, $[s_{i+1}, s_{i+2}]$.

The refinement approach in [5] is an iterative implementation. However, if an actual change point is missed in the first-stage or wrongly discarded in the second-stage, the segment containing this point would contaminate the following iterations, especially for the long segment, since all the data of this segment would be adopted in the next iteration. Therefore, we propose a data balance criterion between the two segments around a candidate point: if the data ratio of the longer segment to the shorter one exceeds a limitation denoted as Data Balance Ratio (DBR), the residual data far from the candidate point in the longer segment would be excluded in the current iteration. The computation is carried on with DBR limitation on the validated indexes sequentially as illustrated in Fig.1. This approach has two advantages: first, it can decrease the probability of data contamination in long segments. Second, it can improve the computational efficiency since the residual data needn't be computed. The DBR value is set as 4 in our algorithm.

3. TRI-MODEL BIC

3.1. Traditional Bi-model BIC

The Bayesian Information Criterion (BIC) is a likelihood criterion penalized by the model complexity [1]. It states that the quality of a model M to represent a data sequence $X = \{x_1, ..., x_n\}$ is given by

$$BIC(M) = \log L(x_1, \ldots, x_n \mid M) - \frac{\lambda}{2} K(M) \log N$$

with $L(x_1, ..., x_n \mid M)$ representing the likelihood of model M estimated from X via maximum likelihood principle and $K(M)$ representing the complexity of model M, equal to the number of free parameters of the model. λ is a penalty weight, theoretically equal to 1; however, it is a tunable parameter as threshold practically.

The problem of determining if there is a change at point i in X can be converted into a model selection problem. The alternative models are: (1) A single-segment model M_0 which assumes that X is generated by a single Gaussian process, that is $\{x_1, ..., x_n\} \sim N(\mu_0, \Sigma_0)$. (2) A two-segment model M_1 which assumes that X is generated by two distinct Gaussian processes, that is $\{x_1, ..., x_i\} \sim N(\mu_1, \Sigma_1)$ and $\{x_{i+1}, ..., x_n\} \sim N(\mu_2, \Sigma_2)$. The BIC values for the two models and the difference between the two BIC values are:

$$BIC(M_0) = \log L(x_1, \ldots, x_n \mid \mu_0, \Sigma_0) - \frac{\lambda}{2} K(M_0) \log N$$

$$BIC(M_1) = \log L(x_1, \ldots, x_i \mid \mu_1, \Sigma_1) + \log L(x_{i+1}, \ldots, x_n \mid \mu_2, \Sigma_2) - \frac{\lambda}{2} K(M_1) \log N$$

$$\Delta BIC = BIC(M_0) - BIC(M_1) =$$

$$\frac{N_1}{2} \log |\Sigma_1| + \frac{N_2}{2} \log |\Sigma_2| - \frac{N}{2} \log |\Sigma_0| + \frac{\lambda}{2} \left(d + \frac{1}{2} d(d+1)\right) \log N$$

N, N_1, N_2 are the number of data vectors in the complete sequence, the subset $\{x_1, ..., x_i\}$, and the subset $\{x_{i+1}, ..., x_n\}$ respectively. d is the dimension of the data vector. A negative value of ΔBIC indicates that the two-segment model fits the data sequence X better, means that there is a change at point i. We denote this traditional approach as bi-model BIC.

3.2. Tri-model BIC

As described above, the value of ΔBIC determines whether there is a change at point i in the data sequence $X = \{x_1, ..., x_n\}$. However, there is a defection in the BIC formulation of the two-segment model: it separates the data sequence into two heterogeneous parts for modeling, but recognizes them as two homogeneous parts in data amount for penalty, which is somewhat paradoxical. As a result, the penalty term $P = \lambda/2 \cdot K(M) \cdot \log N$ is independent of the point i where a change is assumed to occur. It is obvious that P will keep the same across all positions from x_1 to x_n . This implies that different two-segment models of X will be penalized by the same value, and does not adapt to the position of the assumed change, which is somewhat unreasonable.

Therefore, we propose another concept that considers the two-segment model as two independent single-segment models to describe the occurrence of a change at point i in X and uses three models to formulate the problem of determining an assumed change. The three models are: (1) A single-segment model M_0 which assumes that $X = \{x_1, ..., x_n\}$ is generated by a single Gaussian process $N(\mu_0, \Sigma_0)$. (2) A single-segment model M_1 which

assumes that $X_1 = \{x_1, ..., x_i\}$ (the first part of X) is generated by a single Gaussian process $N(\mu_1, \Sigma_1)$. (3) Another single-segment model M_2 which assumes that $X_2 = \{x_{i+1}, ..., x_n\}$ (the remainder part of X) is generated by a single Gaussian process $N(\mu_2, \Sigma_2)$. The BIC values for the three models are:

$$BIC(M_0) = \log L(x_1, \ldots, x_n \mid \mu_0, \Sigma_0) - \frac{\lambda}{2} K(M_0) \log N$$

$$BIC(M_1) = \log L(x_1, \ldots, x_i \mid \mu_1, \Sigma_1) - \frac{\lambda}{2} K(M_1) \log N_1$$

$$BIC(M_2) = \log L(x_{i+1}, \ldots, x_n \mid \mu_2, \Sigma_2) - \frac{\lambda}{2} K(M_2) \log N_2$$

Now, the formula of ΔBIC can be deduced based on the definitions of the three models:

$$\Delta BIC = BIC(M_0) - BIC(M_1) - BIC(M_2)$$

$$= \frac{N_1}{2} \log |\Sigma_1| + \frac{N_2}{2} \log |\Sigma_2| - \frac{N}{2} \log |\Sigma_0|$$

$$- \frac{\lambda}{2} (d + \frac{1}{2} d(d+1)) (\log N - \log N_1 - \log N_2)$$

A negative value of ΔBIC indicates that the quality of modeling the data as a whole sequence by a single Gaussian process is less than the overall quality of modeling the data as two individual sequences by two independent Gaussian processes. Thus, a change could be considered as occurring at point i when the $\Delta BIC<0$. We denote this approach as tri-model BIC.

4. EXPERIMENT AND ANALYSIS

The algorithm is evaluated on three types of corpuses: TV series, broadcast news and phone conversation. Unlike the previous works that only mark relative long segments (>2s or 3s) as detectable units, our work pays more attention to the detectability of short segments (1-3s). Therefore, more precise resolution in segment boundary location is required in our approach.

4.1. Data

- *Dcj31: Excerpt of Dae-Jang-Geum Korean TV Series (Episode 31, 10 minutes), 76 target-changes (speaker and acoustic-class changes), speech with complex background audio.*
- *Hub4: Subset of Hub4 1997 Mandarin Broadcast News (LDC98S73, 30 minutes), 84 target-changes (speaker-changes), speech including little spontaneous speech.*
- *Swb1: Subset of Switchboard 1997 English Phone Conversation (LDC97S62, 10 minutes), 44 target-changes (speaker-changes), pure spontaneous speech.*

Dcj31 and Swb1 are preprocessed by VAD, while Hub4 is not so as to test the robustness of the algorithm with audio stream contaminated by silence or breathing. A statistics of short segments in each dataset is presented in Table 1. It could be seen that the evaluated datasets have quite a lot of short segments.

4.2. Methodology

The algorithm is evaluated by recall rate (RCL), precision (PRC) and F-measure [3] to determine the best segmentation performance. Recall rate is stressed more than precision in the experiments since false alarms can be compensated by following procedures such as clustering or classification. The analysis window and α are set as "1s, 15%", "2s, 30%" and "1s, 30%" for the dataset of Dcj31, Hub4 and Swb1 respectively so as to recall most target-changes with relatively fewer false alarms in the first-stage.

In most previous works, for instance, [7] interprets if there is a gap (e.g. silence, breathing or noise) between two heterogeneous segments, the corresponding target-change is allowed to be at any place in that gap. However, our work requires the target-change to correspond with one of the two boundaries of a segment; all the computed-changes in the gap are considered as false alarms. It is obvious that our evaluation methodology is more rigorous.

	short segments / total segments	boundaries of short segments / total target-changes
Dcj31	32/77 = 41.6%	50/76 = 65.8%
Hub4	16/85 = 18.8%	25/84 = 29.8%
Swb1	10/45 = 22.2%	18/44 = 40.9%
overall	58/207 = 28.0%	93/204 =45.6%

Table 1. The statistics of short segments in the datasets

4.3. Result

Both the traditional bi-model BIC and the proposed tri-model BIC are implemented respectively. The penalty weight λ (with a tuning step of 0.05) is set to maximize the F-measure to achieve the best segmentation performance of each approach on each dataset. The corresponding F-measure, recall rate and precision are presented in Table 2. A statistics of recall rates of short segments and other target boundaries (excluding the boundaries of short segments) is presented in Table 3.

	model	λ	F	RCL	PRC
Dcj31	tri-	1.15	0.734	81.6%	66.7%
	bi-	0.85	0.703	76.3%	65.2%
Hub4	tri-	1.75	0.683	82.1%	58.5%
	bi-	1.35	0.660	77.4%	57.5%
Swb1	tri-	1.30	0.647	75.0%	56.9%
	bi-	1.00	0.638	68.2%	60.0%
Over-all	tri-	-	0.694	80.4%	61.0%
	bi-	-	0.671	75.0%	60.7%

Table 2. The best segmentation performances via tri-model BIC and bi-model BIC on the datasets

From Table 2 and Table 3, it could be seen that tri-model BIC almost wins bi-model BIC in all aspects. In the overall statistics, tri-model BIC leads to more than 7% relative improvement in recall rate without decrease in precision compared with bi-model BIC, and gets about 5% and 9% relative improvements in detecting short segments and other target boundaries respectively.

These facts show that tri-model BIC is explicitly superior to bi-model BIC in the segmentation task. This can be interpreted via the difference between the two ΔBIC formulations, which mainly focus on the penalty term (threshold): the penalty term with better adaptability will achieve better segmentation performance as the tri-model BIC does.

	model	RCL of short segments	RCL of other boundaries
Dcj31	tri-	84.0%	76.9%
	bi-	78.0%	73.1%
Hub4	tri-	64.0%	89.8%
	bi-	68.0%	81.4%
Swb1	tri-	61.1%	84.6%
	bi-	55.6%	76.9%
Overall	tri-	74.2%	85.6%
	bi-	71.0%	78.4%

Table 3. The detectabilities of short segments and other boundaries via tri-model BIC and bi-model BIC on the datasets

4.4. Performance analysis

The segmentation results via tri-model BIC on the evaluated datasets are further analyzed in comparison with the results without DBR limitation in the refinement stage as Table 4 presented.

	DBR	RCL	FAR1	FAR2	AMM
Dcj31	4	81.6%	17.2%	16.1%	0.128s
	∞	80.3%	20.0%	15.8%	0.130s
Hub4	4	82.1%	21.2%	20.3%	0.280s
	∞	79.8%	24.0%	20.7%	0.281s
Swb1	4	75.0%	32.8%	10.3%	0.216s
	∞	72.7%	35.6%	10.2%	0.235s
Over-all	4	80.4%	22.3%	16.7%	0.210s
	∞	78.4%	25.7%	17.1%	0.215s

Table 4. Performance analysis of the approach via tri-model BIC

The false alarms are evaluated by means of two types: FAR1 means there is a computed-change within a homogenous segment; FAR2 means there is a computed-change in the position where an acoustic-change exists but not a target-change. For instance, FAR2 in Dcj31 often represents as one of the boundaries of an instantaneous speech or music segment (<1s), as well as some environmental sound-effects. In Hub4, FAR2 is mostly caused by long silence (>1.5s) including gaps (>1s) between two speaker segments, and some background-audio during outdoor interview. In Swb1, FAR2 often represents as one of the boundaries of a short word (<1s) like "hmm", "yeah" or "ok" when two speakers speak simultaneously. The cause of FAR2 may derive from the first stage via the metric-based segmentation which is more sensitive to every change (intonation or environment) [5].

The refinement approach in [5] could be recognized as a special case of the proposed refinement approach when the DBR value tends to infinite. Actually, the DBR limitation significantly improves the computational speed in the experiments. It could be seen from Table 4 that the DBR limitation effectively improves RCL and decreases FAR1 but has little effect on FAR2 since there are indeed acoustic-changes.

But, from the view of detecting all acoustic-changes including silence [8], our approach seems to be more sensitive and accurate than the previous works, especially for the short segments. For instance, the recall rates of short segments reported in [1] and [6] are 30.9% and 37.6% respectively. However, from the overall statistics in Table 3, it could be seen that nearly 75% of short segments are successfully detected in our approach.

The average mismatch (AMM), which reflects the accuracy for the computed segment boundaries [8], indicates that our approach also has very good resolution (about 0.2s) in segment boundary location.

5. CONCLUSION

In this paper, we present an unsupervised audio segmentation approach which aims at processing real-world media such as TV series, broadcast news and phone conversations, which usually consist of segments in various durations. The proposed tri-model BIC approach shows better segmentation performance and higher resolution of segment boundary location than the previous works. The proposed data balance criterion, for the refinement stage of the approach, also proves to be effective by experiments.

6. ACKNOWLEDEMENT

This work is sponsored by Intel China Research Center, and partially supported by Chinese 973 Program (2004CB318106) and NSF of China (10574140, 60535030). Many thanks to Prof. Zhijian Ou (Tsinghua University) for providing part of the testing audio corpus.

7. REFERENCES

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

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

[3] K.Mori, and S.Nakagawa, "Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition", *ICASSP*, pp.413-416, April 2002.

[4] T.Kemp, M.Schmidt, M.Westphal and A.Waibel, "Strategies for automatic segmentation of audio data", *ICASSP*, vol.3, pp.1423-1426, 2000.

[5] P.Delacourt and C.Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communications*, vol.32, pp.111-126, 2000.

[6] L.Lu and H.Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis". *ACM Multimedia*, pp. 602- 610, Juan-les-Pins, France, 2002.

[7] A.Vandecatseye and J.Martnes, "A fast, accurate and stream-based speaker segmentation and clustering algorithm", *Eurospeech*, pp.941-944, Geneva, 2003.

[8] R.Huang and J.Hansen, "Advances in unsupervised audio segmentation for the broadcast news and NGSW corpora", *ICASSP*, pp.741-744, 2004

[9] P.Zochova and V.Radova, "Modified DISTBIC algorithm for speaker change detection", *Interspeech*, pp.3073-3076, 2005.

[10]M.A.Siegler, U.Jain, B.Raj, and R.M.Stern, "Automatic segmentation, classification and clustering of broadcast news audio", *DARPA Speech Recognition Workshop*, 1997