IDIAP RESEARCH REPORT

BIC Revisited for Speaker Change Detection

Jitendra Ajmera [1, 2] Iain A. McCowan [1]
Hervé Bourlard [1, 2]

IDIAP–RR 02-39

October 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[1] IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P. O. Box 592, CH-1920 Martigny, Switzerland, {jitendra, mccowan bourlard}@idiap.ch
[2] EPFL, Lausanne

IDIAP Research Report 02-39

BIC Revisited for Speaker Change Detection

Jitendra Ajmera Iain A. McCowan Hervé Bourlard

October 2002

Abstract. This paper presents a novel approach for detecting speaker changes in an audio stream. Like previous approaches, two neighboring windows of relatively small size are moved over the audio signal. The similarity between the contents of the two windows is computed using a similarity measure. In this paper, we propose a log-likelihood based criterion which can be used without the need for a threshold/penalty term. This is achieved by comparing the likelihoods of the models with the same complexities. We present an intuitive relationship of the proposed criterion with Bayesian information criterion (BIC), which is essentially a penalized log-likelihood ratio. The criterion was tested on HUB-4 1997 evaluation data and the results show that we achieve a performance comparable to the optimal BIC system.

1 Introduction

Segmentation of audio data is of interest for a broad class of applications, like surveillance, meeting summarization or indexing of broadcast news. The audio may be segmented according to different criteria. In [1], we presented an approach for speech/music segmentation, while in this paper, we address the task of segmenting the audio data in terms of homogeneous speaker segments. In particular, we concentrate on the problem of detecting speaker turns.

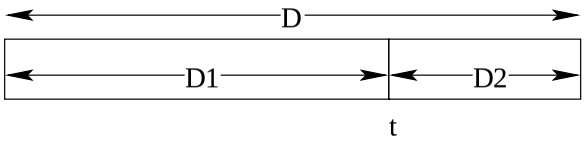

Figure 1: Two neighboring windows with data D_1 and D_2 around time t, where we want to decide if a change point exists or not.

Much research has been devoted to the task of speaker change detection. The most commonly used approaches have been metric based where the problem is formulated like this: Two neighboring windows of relatively small sizes are moved over the audio signal as shown in Figure 1. These windows are dynamically changing with 't' as explained in section 3. For every time instant 't', the similarity between two neighboring windows (having data D_1 and D_2 in Figure 1) is computed using a similarity measure. Then, the local maxima of this measure exceeding a threshold indicate speaker turns.

Various metric based algorithms differ in the similarity measure they employ. Several similarity measures have been investigated for this purpose including the symmetrical Kullback-Liebler (KL) divergence [2, 3], the likelihood ratio test [4, 3] and the Bayesian information criterion (BIC) [5, 6, 7]. These measures are then thresholded to make a decision regarding potential change point at time t. In the case of BIC, the threshold implicity comes from a penalty term. This penalty term involves a parameter λ which has to be tuned for different applications.

In this paper, we propose a new criterion (similarity measure) which removes the need for the penalty/threshold term. This is achieved by comparing likelihoods of models with the same complexity (same number of parameters). An intuitive relationship of this criterion with BIC is presented, and its behaviour is analysed in two extreme theoretical cases.

To assess the performance of the proposed criterion, we compare it to an equivalent BIC system. Of course, performance not only depends on the criterion, but also different parameters related to the process of shifting and comparing windows across time. As the contribution of this paper is the introduction of a new criterion, the rest of the framework (described in Section 3) is left unchanged in comparing the two systems. Experiments were conducted on the HUB-4 1997 evaluation set, with results showing that our (penalty/threshold free) criterion achieves a performance comparable to the optimal BIC system (having tuned parameter λ).

The remainder of this paper is organized as follows: Section 2 presents the proposed criterion and relates it to the BIC. In section 3, we discuss the speaker change detection framework in which the two criterion have been employed. Finally, the evaluation criterion and results are presented in section 4.

2 Proposed Criterion

The goal of the criterion in speaker change detection is to determine the similarity between the contents of windows having data D_1 and D_2 (Figure 1) in order to decide if a change point at time 't' exists or not.

This can be formulated as a problem of hypothesis testing where the two hypotheses H_0 and H_1 are:

- H_0: A change point is hypothesized at time t in Figure 1, and thus the data in the two neighboring windows is believed to come from two different sources (speakers). Accordingly, the probability density functions (PDF) of data D_1 and D_2 are modeled by two individual Gaussian distributions with parameters θ_1 and θ_2 respectively.
- H_1: It is hypothesized that there is no change point at time t and thus the PDF of the complete dataset $D = (D_1 \cup D_2)$ is modeled by a Gaussian mixture model (GMM) with two Gaussian components. The parameters θ of this model are trained on data D using the expectation-maximization (EM) algorithm to maximise the likelihood, i.e. $L(D|\theta)$.

The log-likelihood of the data in the two hypotheses are compared and accordingly a change point is considered at time t if:

$$\log L(D_1|\theta_1) + \log L(D_2|\theta_2) > \log L(D|\theta). \tag{1}$$

In the form of a similarity measure (used later in this paper), we can define ΔL as:

$$\Delta L = \log L(D_1|\theta_1) + \log L(D_2|\theta_2) - \log L(D|\theta) \tag{2}$$

A change point is considered at time t, if $\Delta L > 0.0$. ΔL is then computed for all time instants in the active window and the final change point is decided where ΔL is maximum.

A similar criterion for speaker clustering was used in [8, 9], where the decision about merging of two clusters is made without the need for any thresholding. In [8, 9] this is also achieved by comparing the likelihoods of two models with the same complexity. The criterion proposed here has been adapted from this previous work for the task of speaker change detection.

2.1 How does it work?

In the following, we analyze the behaviour of proposed criterion in two extreme theoretical cases.
Case 1: Let us assume that the subsets D_1 and D_2 come from two entirely different speakers and hence have very distinct probability density functions (PDF). In this case, it is reasonable to assume that the EM algorithm for the GMM training will converge to the same Gaussian distributions as in hypothesis H_0 (with parameters θ_1 and θ_2) weighted with weights w_1 and w_2, $w_1 + w_2 = 1.0$.

If the PDFs are very distinct, we can assume that:

$$L(x_i|\theta_2) \ll L(x_i|\theta_1) \qquad \forall x_i \in D_1 \tag{3a}$$

$$L(x_j|\theta_1) \ll L(x_j|\theta_2) \qquad \forall x_j \in D_2 \tag{3b}$$

Under these assumptions, the $\log L(D|\theta)$ in Eq. (2) becomes :

$$\log L(D|\theta) \approx N_1 \log w_1 + \log L(D_1|\theta_1) + N_2 \log w_2 + \log L(D_2|\theta_2), \tag{4}$$

where, N_1 and N_2 are the number of points in the subsets D_1 and D_2 respectively. Also in this case,

$$w_1 \approx \frac{N_1}{N} \qquad \text{and} \qquad w_2 \approx \frac{N_2}{N}. \tag{5}$$

Using Eq. (4) and (5), ΔL in Eq. (2) can be written as:

$$\Delta L \approx -N_1 \log \frac{N_1}{N} - N_2 \log \frac{N_2}{N} \tag{6}$$

It is then easy to see that:

$$0.0 < \Delta L \leq N \log 2.0 \tag{7}$$

Since $\Delta L > 0.0$, the criterion will favour the hypothesis that there exists a change point at time t.

Case2: Let us assume that D_1 and D_2 come from the same source (speaker), and have very similar real PDFs. In an extreme case, we can assume that $\theta_1 \approx \theta_2$. If we replace θ_1 and θ_2 by θ', the left hand side of Eq. (1) can be written as:

$$\begin{aligned} & \log L(D_1|\theta_1) + \log L(D_2|\theta_2) \\ \approx \quad & \log L(D_1|\theta') + \log L(D_2|\theta') \\ \approx \quad & \log L(D|\theta'). \end{aligned} \tag{8}$$

Moreover, since a GMM can model data D better than a single Gaussian, one can write that :

$$\log L(D|\theta') \leq \log L(D|\theta) \tag{9}$$

Using Eq. (8) and (9) in (2), it can be inferred that $\Delta L \leq 0.0$ and hence the criterion will favour the hypothesis that there does not exist a speaker change at time t. The equality sign will hold true in the case when the true PDFs of the two subsets are identical mono-Gaussian distributions.

2.2 Relation to BIC

The BIC is a model selection criterion originally proposed by Schwarz [10]. It is a likelihood criterion penalized by the model complexity, which is the number of parameters in the model. It formally coincides with Rissanen's minimum description length (MDL) [11].

The BIC was used for speaker change detection in [5, 6]. The problem in that case was formulated in a similar way with the difference that in the hypothesis H_1 (that there is no change point at time t), the data D is instead modeled by another single Gaussian distribution with parameters θ'. Following the previous definitions, the BIC at time t is defined as:

$$BIC(t) = \log L(D_1|\theta_1) + \log L(D_2|\theta_2) - \log L(D|\theta') - \lambda \frac{\Delta K}{2} \log N \tag{10}$$

where ΔK is the difference in the number of parameters of the two models, N is the number of points in D and λ is the penalty factor. The factor λ should ideally be 1.0 but in practice, this factor needs to be tuned for a given application [6, 7, 3].

An intuitive relationship of the proposed criterion with BIC comes from the fact that we compare likelihoods of models having the same number of parameters. This makes $\Delta K = 0$ in Eq. (10), in turn eliminating the need for the penalty factor.

3 Speaker Change Detection Framework

As mentioned previously, metric based approaches to speaker change detection involve shifting of two neighboring windows along the audio stream. We follow the basic algorithm presented in [5] and also incorporate some implementation details presented in [6]. The algorithm now runs as follows:

```
1. initialize the interval [a, b] a =0, b = MIN_WINDOW;
2. find the change point in [a, b] according to the proposed criterion.
3. if(no change in [a, b])
      b = b+ MORE_FRAMES;
   else if(t is the changing point)
      a = t+1, b=a+NEW_SPEAKER_FRAMES;
4. if(b-a > MAX_WINDOW)
      a=b-MAX_WINDOW;
5. go to (2)
```

The main contribution of this paper is to introduce a new criterion for making a decision about a change point, i.e. step 2 in the above algorithm. However, the other parameters, such as `MORE_FRAMES`, `MAX_WINDOW`, etc, can also affect the performance of a system [6]. In the following experiments, in order to compare the proposed criterion directly with the BIC, we changed only the criterion in the two systems (step 2), keeping all other parameters fixed.

4 Experimental Setup

The HUB-4 1997 evaluation set was used to test the performance of the proposed criterion. The HUB-4 database consists of nearly 3 hours of broadcast news data, totalling 624 acoustic changes. However, we restricted our task to only speaker change detection, giving only 515 change points to detect.

Feature vectors used were 24-dimensional mel frequency cepstral coefficients (MFCC) extracted every 10ms. While full covariance matrices could be used for the proposed criterion, here we only used diagonal covariance matrices in order to minimise the computational complexity and allow a real-time implementation.

When full covariance matrices are used in the case of BIC, the ΔK in Eq. (10) is equal to $d + \frac{1}{2}d(d+1)$, where d is the dimension of the feature vectors. In the case of diagonal covariance matrices, we instead have $\Delta K = 2d$.

4.1 Evaluation Criterion

Such a change detection system has two possible types of error, Type-I and Type-II. Type-I errors occur if a true change is not spotted within a certain window (1 second in our case). Type-II errors occur when a detected change does not correspond to a true change in the reference (false alarm). Type I and II errors are also measured using precision (PRC) and recall (RCL) respectively, which are defined as:

$$PRC = \frac{\text{number of correctly found changes}}{\text{total number of changes found}} \tag{11a}$$

$$RCL = \frac{\text{number of correctly found changes}}{\text{total number of correct changes}} \tag{11b}$$

In order to compare the performance of different systems, the F-measure is often used and is defined as:

$$F = \frac{2.0 * PRC * RCL}{PRC + RCL} \tag{12}$$

The F-measure varies from 0 to 1, with a higher F-measure indicating better performance.

4.2 Results

The results using the proposed criterion and the BIC are presented in Table 1. It is clear from the results that the performance of the BIC depends heavily on the value of λ (penalty factor). If this value is too high (greater than 7.0 in this case), the algorithm avoids many false alarms (higher PRC), but at the cost of deleting many genuine changes (lower RCL). On the other hand, if the λ is too low (less than 6.0 in this case), the algorithm generates too many false alarms (lower PRC), in addition to detecting most of the genuine changes (higher RCL). This demonstrates that λ in the BIC serves as an implicit data-dependent threshold.

Conversely, the proposed criterion is free of any such threshold or penalty factor. The performance using the proposed criterion is comparable to that of the BIC with optimal λ (= 6.0 or 7.0), and better when compared to other values of λ, such as the theoretically motivated case of $\lambda = 1$.

We note that the proposed criterion has been used across other data sets in different applications, giving similarly robust results.

Criterion	RCL	PRC	F
Proposed	0.65	0.68	0.67
BIC ($\lambda = 1.0$)	0.81	0.22	0.35
BIC ($\lambda = 4.0$)	0.77	0.46	0.58
BIC ($\lambda = 5.0$)	0.74	0.57	0.64
BIC ($\lambda = 6.0$)	0.71	0.66	0.68
BIC ($\lambda = 7.0$)	0.66	0.71	0.68
BIC ($\lambda = 8.0$)	0.60	0.73	0.66

Table 1: Results of the proposed criterion and BIC (with different values of λ) on HUB-4 1997 evaluation data. The results are presented in terms of recall (RCL), precision (PRC) and F-measure. As expected, a higher value of λ results in more deletions, and less false alarms (higher PRC and lower RCL). The best results for the BIC (with $\lambda = 6.0$ or 7.0), are comparable to the results obtained using the proposed criterion.

5 Conclusion

A new criterion for speaker change detection has been proposed in this paper. An intuitive relation of the criterion with the BIC was presented. In contrast to other metric based approaches, the proposed criterion does not require a penalty/threshold term to make decisions. This is achieved by comparing likelihoods of models with same number of parameters. Two theoretical sample cases were presented to illustrate the behaviour of the criterion. The proposed criterion was tested on the HUB-4 1997 evaluation data and results were compared with those obtained using the BIC. The results show that the proposed threshold-free criterion gives comparable performance to that of the optimal BIC system.

Acknowledgement

This work was supported by the Swiss National Science Foundation through project no. 2100-65067.01 on "AudioSkim".
We would also like to thank Prof. Jean-Pierre Martens for his co-operation in the evaluation process.

References

[1] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speech/music segmentation using HMM," *ICASSP*, 2002.

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

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

[4] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker idenitification," *ICASSP*, 1991.

[5] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," Tech. Rep., IBM T.J. Watson Research Center, 1998.

[6] Alain Tritschler and Ramesh Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," *Eurospeech*, pp. 679–682, 1999.

[7] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111–126, Sept. 2000.

[8] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," *ICSLP*, 2002.

[9] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved unknown-multiple speaker clustering using HMM," Tech. Rep., IDIAP RR 02-23, 2002.

[10] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[11] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE transaction on information theory*, vol. IT-30, no. 4, July 1984.