



# **XBIC: Real-Time Cross Probabilities measure for speaker segmentation**

**Xavier Anguera (xanguera@icsi.berkeley.edu)**

**International Computer Science Institute(ICSI)  
Berkeley, CA, 94704**

**and**

**Department of Signal Theory and Communications,  
TALP Research Center**

**Technical University of Catalonia (UPC)  
Barcelona, Spain**

**TR-99-2004**

**August 2005**

## **Abstract**

In this paper a novel probability based measure is presented that shows good results for real time blind speaker segmentation. In such task there is no previous information about the identity or how many speakers there are. Similar to the Bayesian Information Criterion (BIC), the proposed measure indicates the similarity between two speech segments on either side of a given test point. By computing cross probabilities between both segments, an abrupt decrease of the measure value indicates the existence of an acoustic change point. A scrolling window implementation, in a similar way that is used in metric based techniques, is shown to give better results regarding speed and change detection. This measure allows building real-time systems.

Tests with Broadcast news data show a general improvement compared to the commonly used BIC method.



# 1 Introduction

Segmenting a speech utterance into the different speakers that appear in the conversation has many applications. There are applications in speech and speaker recognition, audio/speaker indexing and transcription techniques, among others. Speaker segmentation consists of finding the temporal positions where there is an acoustic change that indicates a change of speaker or background conditions.

We can find in the literature some methods applied to address speaker segmentation. Metric based techniques [1] define acoustic distance measures to evaluate the similarity between two adjacent windows of speech. Such windows are scrolled through the speech utterance and the resulting distances curve is evaluated to find speaker changes. Another method that is often used is the Bayesian Information Criterion (BIC) [2]. Given a working speech segment and a proposed change point, the segments at both sides of that point are modeled with one or two Gaussian models. The difference between both alternatives compared to the complexity of training more parameters is used to decide if that is a feasible changing point. The working window scrolls until the end of the utterance is found. In order to increase the performance and/or precision, some systems use a double pass algorithm with either of these methods used alone or combined. Another method uses iterative segmentation ([3],[4]) where speakers are added or deleted until the optimum number of speakers and segmentation is found.

In this paper we present a new measure and demonstrate how we have used it in speaker segmentation, giving very interesting results. Similar to the BIC formulation, it measures the dissimilarity between two adjacent segments connected with a test point. In this approach each segment is modeled with a Gaussian distribution without any restriction on the topology used. The evaluation of each point is performed by calculating the cross probabilities of each segment given the other segment's model. It is decided that it is an acoustic changing point if the value falls below a predefined threshold.

A double-pass scrolling-window system is proposed, similar to the acoustic distance measures techniques, to be used with the proposed measure. It strives for simplicity, computational efficiency and solving problems like the masking of acoustic changes after one has been detected, present in common BIC implementations. We call the system presented Cross-Probabilities-BIC (XBIC).

## 2 BIC theory background

The Bayesian Information Criterion (BIC) is a well known method for speaker segmentation ([2]). It allows the creation of real time systems.

Given  $\Theta = \{\theta(j) \in \mathbb{R}^d | j \in 1 \dots N\}$ , a sequence of N observation vectors with dimension d, that have been parameterized from the speech signal to be segmented.

The Bayesian information criterion for  $\Theta$  is a penalized log likelihood as follows:

$$BIC_{\Theta} = \mathcal{L} - \lambda P \tag{1}$$

Where P is a penalty term and  $\lambda$  is a free design parameter dependent on the data being modeled. By default it is set to 1.

Given a point  $\theta(i) \in \Theta$ , we can define 2 partitions from  $\Theta$ :  $\Theta_1 = \{\theta_1(1) \dots \theta_1(i)\}$  and  $\Theta_2 = \{\theta_2(i+1) \dots \theta_2(N)\}$  with lengths  $N_1$  and  $N_2$ .

In many of the systems present in the bibliography ([2], [5], [6], [7]), the data is modeled with a full single Gaussian of dimension  $d$ . Therefore the likelihood  $\mathcal{L}$  becomes:

$$\mathcal{L} = -\frac{1}{2}N\log|\Sigma| + NC \quad (2)$$

Where  $|\Sigma|$  is the determinant of the covariance matrix and  $C$  is a constant,  $-\frac{1}{2}d(1 + \log(2\pi))$ .

For some applications we would like to have more flexibility in choosing the kind of models to use. The likelihood can be written as:

$$\mathcal{L} = \mathcal{P}(\Theta|\lambda) = \sum_{k=1}^N \log p(\theta(k)|\lambda) \quad (3)$$

When making a decision whether there is an acoustic change at point  $\theta_i$ , we consider two hypotheses: two independent models best fit the data on both sides of the change point  $\theta_i$  versus one single model fitting all of the data. The best hypothesis is chosen by evaluating

$$\Delta BIC = BIC_{\Theta} - (BIC_{\Theta_1+\Theta_2}) \quad (4)$$

As it is seen in [4],  $\Delta BIC$  is formulated as the ratio between the log probabilities of both hypotheses in the following way:

$$\Delta BIC(i) = \mathcal{P}(\Theta|\lambda) - [\mathcal{P}(\Theta_1|\lambda_1) + \mathcal{P}(\Theta_2|\lambda_2)] - \frac{1}{2}\Lambda K \log N \quad (5)$$

Where  $\lambda_1$ ,  $\lambda_2$  and  $\lambda$  represent the models for partitions  $\Theta_1$ ,  $\Theta_2$  and  $\Theta$  respectively.  $K$  is the difference in the number of parameters between  $\lambda_1$  and  $\lambda_2$  and  $\Lambda$  is a design constant.

It is decided that there is an acoustic changing point if  $\Delta BIC > 0$ , meaning that two models better fit the data than one.

### 3 Probabilistic distance measure for Hidden Markov Models

Let us show that the development of BIC into probabilistic terms is related in some sense with the distance introduced by L. Rabiner in [8],[9] as a probabilistic distance measure for Hidden Markov Models. Rabiner defined a distance measure between two existing Markov models as a combination of the likelihoods of two sets of artificially generated data evaluated by the two models.

Given two HMM models defined by  $\lambda_1 = (A_1, B_1, \pi_1)$  and  $\lambda_2 = (A_2, B_2, \pi_2)$  we consider that each of them is able to generate a data set  $\Theta_1 = \{\theta_1(1), \dots, \theta_1(N_1)\}$ ,  $\Theta_2 = \{\theta_2(1), \dots, \theta_2(N_2)\}$ .

The distance, noted  $D(\lambda_i, \lambda_j)$ , between the two different models is defined as:

$$D(\lambda_i, \lambda_j) = \frac{1}{N_j} (\mathcal{P}(\Theta_j|\lambda_i) - \mathcal{P}(\Theta_j|\lambda_j)) \quad (6)$$

Where in general:

$$\mathcal{P}(\Theta_i|\lambda_j) = \sum_{k=1}^{N_i} \log p(\theta_i(k)|\lambda_j) \quad (7)$$

As the distance  $D(\lambda_i, \lambda_j)$  is not symmetric, we need also to take into account its counterpart  $D(\lambda_j, \lambda_i)$ . The probability distance is finally:

$$D_{rab} = \frac{D(\lambda_i, \lambda_j) + D(\lambda_j, \lambda_i)}{2} \quad (8)$$

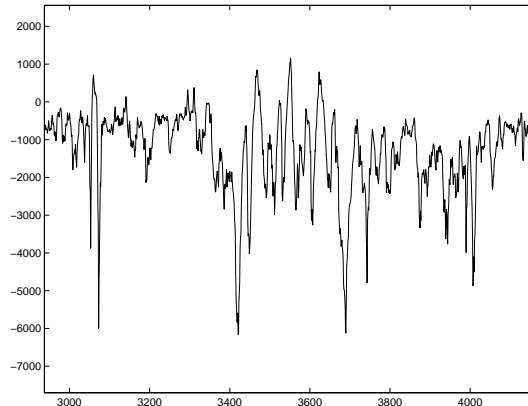


Figure 1: *XBIC distance plot for a segment with different speakers*

## 4 Cross Probabilities Method

Meanwhile the distance in (8) was defined to compare a pair of existing models by using artificially generated data, we propose to compare the two sequences of existing observation vectors by calculating the distance between two models trained with them.

If we assume that the two segments have the same length ( $N_1 = N_2$ ) and rearranging the terms in equation (8) we can define:

$$D'_{rab} = (\mathcal{P}(\Theta_1|\lambda_2) + \mathcal{P}(\Theta_2|\lambda_1)) - (\mathcal{P}(\Theta_1|\lambda_1) + \mathcal{P}(\Theta_2|\lambda_2)) \quad (9)$$

Equation (9) is similar to expression (5) but it does not have a penalty term, which just moves to 0 the decision threshold of the hypothesis test. In both methods we are defining a hypothesis test between two terms. The second term ( $\mathcal{P}(\Theta_1|\lambda_1) + \mathcal{P}(\Theta_2|\lambda_2)$ ) is common in both equations and relates to how well the data is modeled by two independent models.

The first term from eq. (9) and (5) measures how well both segments are related to each other. In the BIC formulation this is done by evaluating how a single model containing all the data performs. In (9) it tests how well each segment's model can represent the other segment's data. In both cases, the more similar the two segments are, the bigger the resulting probability will be.

Given a speech segment, the distance proposed in (9) has a value close to 0 for points within similar acoustic regions and becomes negative when they are dissimilar. Minimums in this distance measure show places where acoustic changes are most provable. In such change points the value measures decreases abruptly mainly due to the cross probabilities, being the second term residual. As our interest is on finding acoustic changes, we can simplify the equation and define the XBIC measure as:

$$XBIC(i) = (P(\Theta_1|\lambda_2) + P(\Theta_2|\lambda_1)) \quad (10)$$

When evaluating each speech segment with the opposite model it measures how acoustically close are both segments. bigger values (negative, close to 0) represent more similar segments and the smaller the value, the more probable is that they belong to a different speaker.

In figure (1) we show the XBIC(i) measure for a speech segment with different speakers where segments  $\Theta_1$  and  $\Theta_2$  have been scrolled through the speech segment and the measure calculated

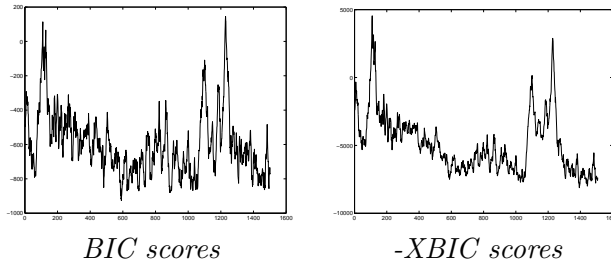


Figure 2: Score plots for a segment containing different speakers

in each point. The existence of low value regions in the plot indicates candidates to be acoustic changing points. By defining an appropriate threshold we conclude that there is an acoustic change if  $XBIC(i) < Thr_{XBIC}$

In figure (4) we plot the BIC and XBIC measure for a speech segment containing several speaker changes, possible change points get magnified using XBIC. We therefore can reduce the false detection of changes, which is a well known problem for the BIC approach.

## 5 Segmentation System Architecture

It is desirable for many systems based on BIC and metric distances to be implemented in real time. In order to test the new measure, a sequential architecture is proposed which resembles the metric based systems. This architecture uses the XBIC distance and achieves increased accuracy using a two passes algorithm.

As seen in figure (3), as the signal enters the system it is parameterized with MFCC coefficients. The probability distance algorithm is then computed in two steps, in a similar way to [7], in order to ensure a good tradeoff between speed and accuracy. A first pass calculates the XBIC measures until the acoustic change criteria is met. Then a second pass looks around that point to find the exact change point. In both passes the decision point joins two segments of equal length- T frames. In the second pass T can be reduced, to focus in on the region around the proposed change point.

To meet the acoustic change criteria the XBIC measure must fall below a predefined threshold and the value must be an absolute minimum among neighboring measures. This is done to minimize the amount of false detections due to local minima.

A distance  $D_{fast}$  and  $D_{slow}$  is applied between measures for the first and second passes respectively, which increases the quickness of the algorithm while no possible changing points are found ( $D_{fast} \gg D_{slow}$ ). The second pass is computed in a window of length  $\pm D_{fast}$  frames around the possible change point to find the exact location.

Once an acoustic change point has been decided by both steps, this is output and the algorithm continues with the distance computation one frame after the detected point.

In order to compare the use of fixed versus variable length segments we have also implemented the XBIC measure following [7]. Given a fixed analysis window, evaluation points define different length segments. If a speaker change is detected a new fixed window is defined starting at that change point. If no change is found, the fixed window is enlarged and the process is restarted.

In order to compare BIC with XBIC, we have implemented BIC using this same system.

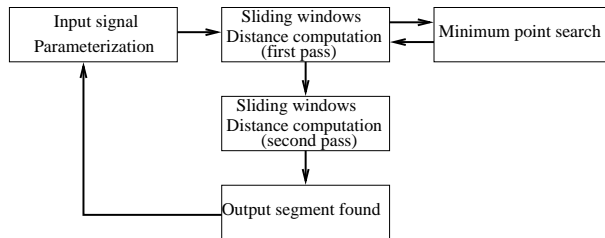


Figure 3: *Speaker segmentation system architecture using XBIC*

## 6 Experimental Results

### 6.1 Experimental Setup

Two databases have been used to test the proposed XBIC measure and the system in which it is implemented: the 1996 HUB-4 Evaluation Test material and the 1997 HUB-4 Evaluation material. The first one consists of more than two hours of English Broadcast News in a variety of different acoustic conditions, with almost 400 speaker changes split in four files from four shows in different radio stations. The second one has 512 speaker changes, with similar conditions to the 1996 test.

For all cases the input signal has been parameterized using 32 MFCC coefficients (16 static + 16 dynamic), extracted every 10ms with a 25ms window size. In all systems, GMM models with one Gaussian and full covariance were used.

For both implementations, we selected step sizes of  $D_{fast} = 0.1s$ ,  $D_{slow} = 0.01s$ . The segment length is  $T_{1st\_pass} = 4$  seconds and  $T_{2nd\_pass} = 2$  for the fixed size system, and all to  $T_{min} = 4$  seconds for the variable size window.

For both systems using XBIC and BIC, several thresholds and  $\lambda$  values have been selected in order to plot curves representing various possible cases.

### 6.2 Evaluation and Results

Two kinds of error measures have been computed, false detection (FD) and false rejection (FR):

$$\%FD = \frac{\# \text{ false\_detections}}{\text{total\_amount\_of\_detections}} \quad (11)$$

$$\%FR = \frac{\# \text{ missed\_detections}}{\text{total\_amount\_of\_true\_changing\_points}} \quad (12)$$

In some publications these values are inverted and instead they use the recall as  $RCL = 1 - FR$ , and the precision as  $PRC = 1 - FD$ .

In order to summarize these two metrics into one, the F measure is defined as:

$$F = \frac{2 * PRC * RCL}{PRC + RCL} \quad (13)$$

The script used for the evaluation is called eval\_seg and is also used in COST278 evaluations (see [10]), with the same evaluation conditions.

Our first experiments are directed to compare the two alternative system implementations using the XBIC measure. Both systems have been tested with both databases and the EER (Equal Error

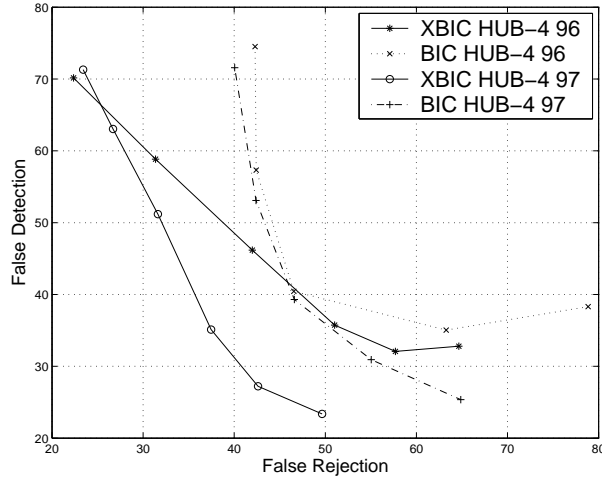


Figure 4: *DET curves for BIC and XBIC for both databases used in evaluation*

Rate) point (for which False Rejection and False Detection errors are the same) has been computed. The F value is computed for these cases.

F measure	HUB-4 96	HUB-4 97
fixed segments	55.84	63.68
variable segments	51.47	48.88

Table 1: XBIC measure results using different segment lengths

As we can see in table (1), for both databases the fixed window approach outperforms the variable window system. The global average improvement a 19,4%. Such difference was expected as (9) was defined under the hypothesis that both signal segments had the same length.

The next step is to compare the XBIC measure using the proposed system architecture and the BIC algorithm with variable size segments. We can see in figure (4) the DET curves for both methods and for both databases taken into consideration.

In both databases the XBIC method outperforms the BIC algorithm in almost all considered evaluation points. In the EER point the F values using XBIC are 3% higher in HUB-4 96 and 18% higher in HUB-4 97 than using BIC. This indicates a substantial dependence of segmentation algorithms with the nature of the data. For the HUB-4 96 database both systems behave almost the same way, but for HUB-4 97 data, XBIC is clearly better.

We can see also how the DET curve for XBIC is more linear than the one from BIC, which increases the difference between the two for values other than the EER case. In practice, this allows systems to be built using different working points, still with similar characteristics to the EER.

The measure of XBIC involves less computational cost than BIC. XBIC only trains one model on each segment, whereas BIC needs also to train the compound model.

By using the proposed system instead of a variable-size segments system, we obtain two main improvements. On the one hand is the speed and simplicity of the system. In the proposed architecture the algorithm advances with fixed steps and it doesn't back up unless a potential change point is found. In the variable-size-segments system, until the potential change point is

found the algorithm keeps backing up and repeating the measures in a slightly modified window. This significantly increases the computation time.

On the other hand, once a change point is found, the proposed architecture restarts the measures right after the change point, being able to detect new change points immediately. By using BIC with a standard system, after a change point is found a fixed segment size is set for the first segment, masking any change points within that region. Furthermore, if any exists within that region, it affects the detection of other changes.

## 7 Conclusions

In this paper we present a novel measure (XBIC) to perform real-time blind speaker segmentation. Similar to the BIC method, a hypothesis test is performed evaluating whether one or two Gaussian models best represent the data from two adjacent segments.

The XBIC measure takes the decision by calculating the cross-probabilities between each data segment and the model trained with data from the other segment. For a given acoustic change point there is an abrupt decrease in the measure. This is easy to detect using a predefined threshold.

We have implemented the XBIC measure using a system with two fixed length segments scrolling through the input speech with a two passes algorithm. It is a simple system which gives good results.

Tests on Broadcast news data show that the proposed system behaves similarly or better than BIC and has a more linear error curve. This shows that XBIC is an interesting alternative to the common BIC based systems as it opens new possibilities and solves existing problems like masking regions after detected speaker changes and excessive false detections.

## References

- [1] J.W. Hung, H.M. Wang, and L.S. Lee, "Automatic metric based speech segmentation for broadcast news via principal component analysis," in *ICSLP'00*, Beijing, China, 2004.
- [2] S. Shaobing Chen and P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998.
- [3] X. Anguera and J. Hernando, "Evolutive speaker segmentation using a repository system," in *ICSLP'04*, Jeju Island, Korea, Oct. 2004.
- [4] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *ASRU'03*, US Virgin Islands, USA, Dec. 2003.
- [5] L. Perez-Freire and C. Garcia-Mateo, "A multimedia approach for audio segmentation in tv broadcast news," in *ICASSP'04*, Montreal, Canada, May 2004, pp. 369–372.
- [6] S.E. Tranter and D.A Reynolds, "Speaker diarization for broadcast news," in *ODISSEY'04*, Toledo, Spain, May 2004.
- [7] P. Sivakumaran, J. Fortuna, and A.M. Ariyaeinia, "On the use of the bayesian information criterion in multiple speaker detection," in *Eurospeech'01*.

- [8] B.H. Juang and L.R. Rabiner, “A probabilistic distance measure for hidden markov models,” *AT&T Technical Journal* 64, AT&T, Feb. 1985.
- [9] L.R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb. 1989.
- [10] A. Vandecatseye, J-P Martens, et al., “The cost278 pan-european broadcast news database,” in *LREC’04*, Lisbon, Portugal, May 2004.