# A NEW PENALTY TERM FOR THE BIC WITH RESPECT TO SPEAKER DIARIZATION

*Themos Stafylakis*[1,2]*, Georgios Tzimiropoulos*[3]*, Vassilis Katsouros*[1] *and George Carayannis*[1,2]

[1]Institute for Language and Speech Processing, Greece, [2]National Technical University of Athens, Greece,
[3]Imperial College London, UK
{themosst,vsk,gcara}@ilsp.gr & georgios.tzimiropoulos@imperial.ac.uk

**ABSTRACT**

In this paper we revise the penalty term of the Bayesian Information Criterion (BIC). Based on our previous approach to penalize each cluster only with its corresponding effective sample size - which we called the Segmental-BIC - we examine a new formula of the penalty term. The criterion we derive has the appealing property of the Segmental-BIC, that is it approximates the evidence of overall partitions while leading to an autonomous pairwise dissimilarity measure. We tested our new criterion on two speaker diarization benchmarks and we report significant increase in accuracy.

***Index Terms***— Bayesian Information Criterion, Cluster Analysis, Speaker Diarization

## 1. INTRODUCTION

We concern the problem of text-independent Speaker Diarization (SD), i.e. the problem of automatically grouping an audio document (broadcast news, meetings, etc.) into speakers, without knowing a priori the identities and the number of the participants or using the transcript. The task is of great importance in many areas of speech processing, including speaker - adaptive speech recognition, speaker recognition in broadcast news (BN), enrichment of the transcription with speaker-level information and others.

Like many other areas in speech processing, the use of Bayesian Statistics provides us with a solid paradigm for formulating our prior beliefs and draw inference about the quantities of interest. The Bayesian Information Criterion (BIC) [1] is an elegant reference test for model comparison and hypothesis testing and as such it has been adopted from the SD community as a fundamental criterion for estimating the partition and the number of speakers. A fundamental property of the BIC is its capacity to approximate the evidence of overall partitions, using a specific type of priors - the unit-information priors, [2]. For many inferential tasks, such a prior is a reasonable choice. For instance, in density estimation using finite mixture models, the BIC given good results with respect to the generalization performance [3].

However, the introduction of the Local-BIC and the significant increase in the SD accuracy it achieved, showed that the original formulation of the BIC in [4], (i.e. the Global-BIC) was far from being optimal for the SD task. The Local-BIC is an autonomous pairwise dissimilarity measure, i.e. the corresponding $\Delta$BIC formula is completely defined by the sufficient statistics of the two clusters being examined and their sizes. Nevertheless, the Local-BIC exists only in $\Delta$BIC formula, meaning that it cannot approximate the evidence of overall partitions. One can only utilize it to obtain a point-estimate for the partition, using algorithms that are based on pairwise distances. As a result, it cannot be regarded as a means to draw inference about the partition.

To conbine the strengths of the two approaches, the authors proposed in [5] a new variant, the Segmental-BIC. The idea is to redefine the priors of the BIC, so that the corresponding $\Delta$BIC becomes autonomous. The results show that the Segmental-BIC is at least comparable to the Local-BIC and superior to the Global-BIC, especially in cases where the purity of the clusters counts more that their coverage. However, the results we demonstrate in this paper show that even with the baseline formulation of the SD, i.e. change-point detection followed by the Agglomerative Hierarchical Clustering (AHC) stage with single-Gaussian densities (see [6]), much better performance can be attained. The proposed criterion is a new variant of the Segmental-BIC and is based on the analysis pioneered by Sin and White in [7] about the properties that a criterion should meet in order to be consistent.

The outline of the paper has as follows. In Section 2, a brief review of the BIC is given, and the use of the BIC in SD is discussed. The Segmental-BIC is presented in Section 3, where the refinement of the penalty term is introduced. Finally, the criteria are tested against two benchmark tests and the results are given in Section 4, followed by some future work directions.

## 2. BIC, UNIT INFORMATION PRIORS AND SPEAKER DIARIZATION

### 2.1. Information Criteria and the rationale for the BIC

Suppose we are given a sample of $N$ observation vectors $\mathcal{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}]^T$, $\mathbf{x}^{(i)} \in \Re^d$ and we want to infer the true underlying model from a predefined set $\mathcal{M}_j : \theta \in \Theta_j \subseteq \Re^{\mathcal{P}_j}$, where $\mathcal{P}_j$ denotes the number of free parameters under $\mathcal{M}_j$. If none of them is the true one (i.e. we are dealing with misspecified models) the analysis is still valid; we should choose the one that has asymptotically the minimum Kullback-Leibler divergence between the Data Generating Process and the model. If two or more have the same divergence asymptotically, we should choose the most parsimonious. The rationale for the BIC is to approximate the integrated-likelihood (or the evidence) of the model $p(\mathcal{X}|\mathcal{M}_j) = \int_{\Theta_j} p(\mathcal{X}|\theta, \mathcal{M}_j)\pi(\theta|\mathcal{M}_j)d\theta$

by expanding it as a quadratic around its MAP-mode $\tilde{\theta}$, a technique known as the Laplace approximation, [2]. The approximation of $S = \log p(\mathcal{X}|\mathcal{M}_j)$ yields

$$S \approx \frac{\mathcal{P}_j}{2}\log 2\pi - \frac{1}{2}\log|\tilde{\mathcal{H}}_\theta(\tilde{\theta})| + l(\tilde{\theta}|\mathcal{X}) + \log\pi(\tilde{\theta}|\mathcal{M}_j) \quad (1)$$

where $\tilde{\mathcal{H}}_\theta(\tilde{\theta})$ is the Hessian of $-\log p(\mathcal{X}|\theta,\mathcal{M}_j)\pi(\theta|\mathcal{M}_j)$ with respect to $\theta$, evaluated at $\tilde{\theta}$. As $N$ grows, the MAP-mode attains the ML-mode $\hat{\theta}$, assuming some regularity conditions with respect to the prior. By separating the terms that scale with $N$ from the ones that do not, the above yields $S \approx SIC + T$, where

$$SIC = l(\hat{\theta}|\mathcal{X}) - \frac{\mathcal{P}_j}{2}\log N \quad (2)$$

and

$$T = \frac{\mathcal{P}_j}{2}\log 2\pi - \frac{1}{2}\log|\mathcal{J}_\theta(\hat{\theta})| + \log\pi(\hat{\theta}|\mathcal{M}_j) \quad (3)$$

where we used the decomposition of the observed information with the expected information $\mathcal{J}_\theta(\hat{\theta})$ as $|\hat{\mathcal{H}}_\theta(\hat{\theta})| = N^{\mathcal{P}_j}|\mathcal{J}_\theta(\hat{\theta})|$

$$(\mathcal{J}_\theta(\hat{\theta}))_{ij} = \int p(\mathbf{x}|\hat{\theta})\left(-\frac{\partial^2\log p(\mathbf{x}|\hat{\theta})}{\partial\theta^i\partial\theta^j}\right)d\mathbf{x} \quad (4)$$

has been utilized. The term $SIC$ stands for the Schwarz Information Criterion. Formally, the BIC is defined as twice the $SIC$.

### 2.2. Unit information priors and the tuning parameter

When the prior is not specified, the approximation of $S$ is $O(1)$. If however the unit-information priors are assumed,

$$\theta \sim \mathcal{N}(\hat{\theta}, \mathcal{J}_\theta^{-1}(\hat{\theta})) \quad (5)$$

the approximation becomes $S = SIC + O(N^{-1/2})$, [2]. The interpretation is to form a data-dependent prior utilizing the amount of information contained in a single observation. Note, also that the centering of the prior at the ML estimate is rather inadequate for small sample sizes. In the SD task, a better approach would be to center the prior of the parameters of each cluster at a (possibly pre-trained) model that corresponds to a representative speaker of the same macro-class (i.e. of the same gender - bandwidth - acoustic environment) with the clusters in question.
Furthermore, note that if $N^{1-\lambda}$ observations are utilized, the penalty term of the $SIC$ becomes $\frac{\lambda\mathcal{P}_j}{2}\log N$. Hence, the tuning parameter can be interpreted as a hyperparameter that controls the variance of the sample size dependent prior. By placing $\lambda > 1$, we stretch the variance of the prior as $N$ grows so that the observations overwrite the prior more quickly. For small sample sizes, the prior is more informative (i.e. of lower variance) because the ML estimate is in general a poor estimator for the true value of the parameters. In such cases, a more informative prior may prevent the MAP-estimate from attaining unrealistic values, due to the small coverage of the range of phonemes and/or the speech abnormalities. As we show next, the new version of the Segmental-BIC assumes a prior that becomes nearly flat as $N$ grows with a much higher rate, yet it always remains proper, i.e. it integrates to one.

### 2.3. Speaker Diarization and the use of the BIC

The Global-BIC (7) is a form of BIC suited to cluster analysis. It is based on the *classification* integrated log-likelihood (see [3])

$$l(\hat{\varphi};\mathbf{s}|\mathcal{X}) = \sum_{i=1}^{N}\log f(\mathbf{x}^{(i)}|\hat{\varphi}_{\mathbf{s}^{(i)}}) \quad (6)$$

that is the log-likelihood conditioned on a single partition $\mathbf{s}$.

$$BIC^G = 2l(\hat{\varphi};\mathbf{s}|\mathcal{X}) - \lambda\mathcal{P}_j^{in}\log N \quad (7)$$

We use the notation $\varphi_k = (\mu_k, \Sigma_k)$, $k = 1,\ldots,K$, $K = \max(\mathbf{s})$ to denote the space of the *internal* parameters $\varphi \in \Phi \subseteq \Re^{\mathcal{P}_j^{in}}$. The *external* parameters $\alpha \in \mathcal{A} \subseteq \Re^{\mathcal{P}_j^{ex}}$ correspond to the state transition probabilities if a HMM topology is assumed. They do not appear in the formula due to the conditioning on $\mathbf{s}$ and the use of flat priors over the space of allowed partitions, i.e. all the partitions that comply with the minimum state occupancy duration constrains and have ascending ordering of the labels for each new speaker entry (i.e. baseform labeling). Note finally that $\mathcal{P}_j = \mathcal{P}_j^{in} + \mathcal{P}_j^{ex}$, $\mathcal{P}_j^{in} = KP$ and $P = d + d(d+1)/2$ if single-Gaussians are used.
From the corresponding $\Delta BIC$ of the Global-BIC,

$$\Delta BIC^G = 2\log GLR_{ab} - \lambda P\log N \quad (8)$$

where

$$GLR_{ab} = \frac{p(\mathcal{X}_a|\hat{\varphi}_a)p(\mathcal{X}_b|\hat{\varphi}_b)}{p(\mathcal{X}_{a\cup b}|\hat{\varphi}_{a\cup b})} \quad (9)$$

denotes the Generalized Likelihood Ratio between two utterances $\mathcal{X}_a$ and $\mathcal{X}_b$, one may observe that despite the hard clustering scheme, its orientation remains the density estimation, i.e. infer $K$ that generalizes best to unseen data. The dissimilarity measure between two fixed clusters decreases with the overall sample $N$, which is a typical behaviour of complexity criteria that aim to favour compact and robust representations of a data set. However, the task we are concerned with is rather different. We want to estimate the speaker-oriented partition; the true number of speakers is estimated indirectly. Note also that the maximization of the classification integrated likelihood instead of the posterior of the partition is a Bayesian procedure that implies uninformative priors over the space of partitions and not over $K$. Multiple experiments (e.g. [8]) have shown that an autonomous dissimilarity measure over the space of internal parameters, like the Local-BIC,

$$\Delta BIC^L = 2\log GLR_{ab} - \lambda P\log(n_a + n_b) \quad (10)$$

is far more accurate in terms of Diarization Error Rate (DER). The Local-BIC, however, suffers from several limitations explained above and in [5]. Hence, it cannot be considered as optimal.

## 3. THE SEGMENTAL-BIC APPROACH

### 3.1. The key-idea of the Segmental-BIC

As described in [5], a way to merge the two variants is to attach the following prior to the parameters of the *k*th cluster

$$\varphi_k \sim \mathcal{N}\left(\hat{\varphi}_k, n_k^{\lambda-1}\mathcal{J}_{\varphi_k}^{-1}(\hat{\varphi}_k)\right) \quad (11)$$

where

$$(\mathcal{J}_{\varphi_k}(\hat{\varphi}_k))_{ij} = \int p(\mathbf{x}|\hat{\varphi}_k) \left( -\frac{\partial^2 \log p(\mathbf{x}|\hat{\varphi}_k)}{\partial \varphi_k^i \partial \varphi_k^j} \right) d\mathbf{x} \qquad (12)$$

The above prior leads to the following criterion

$$BIC^S = 2l(\hat{\varphi}; \mathbf{s}|\mathcal{X}) - \lambda P \sum_{k=1}^{K} \log n_k \qquad (13)$$

This principle of the Segmental-BIC is to utilize the same amount of information (i.e. $n_k^{1-\lambda}$ observations) to form the prior for clusters of fixed size, instead of documents of fixed-size. Doing so, the corresponding $\Delta$BIC formula

$$\Delta BIC^S = 2 \log GLR_{ab} - \lambda P \log \frac{n_a n_b}{n_a + n_b} \qquad (14)$$

becomes independent from $N$. Hence, the Segmental-BIC is a complexity criterion that approximates the evidence of overall partitions (like the global one), while preserving the pairwise distances (like the local one).

### 3.2. The refinement of the Segmental-BIC

We now refine the above penalty term. The analysis is based on [7] where the consistency of the information criteria is examined. Let us denote by $c_n$ the penalty term of the $\Delta$BIC formula. Let $\mathcal{H}_0$ be the null hypothesis that $\mathcal{X}_a$ and $\mathcal{X}_b$ belong to the same speaker. Under $\mathcal{H}_0$, the most general requirement for weak consistency is $P(n^{-1/2} c_n \rightarrow \infty) = 1$, where $n = n_a + n_b$, i.e. the penalty term should grow faster than $\sqrt{n}$. Under the alternative hypothesis, it should grow slower that linearly, so that the difference between the likelihoods dominates the results. In order to accomplish these requirements, while retaining the properties of the Segmental-BIC, we propose the following criterion

$$BIC^S_{SR} = 2l(\hat{\varphi}; \mathbf{s}|\mathcal{X}) - \lambda P \sum_{k=1}^{K} \sqrt{n_k} \log n_k \qquad (15)$$

which we name it the Segmental Square Root-BIC. Using straightforward calculations, one may verify that the corresponding $\Delta$BIC formula meets the demands discussed above.

The implied priors have as follows

$$\varphi_k \sim \mathcal{N}\left( \hat{\varphi}_k, n_k^{\lambda\sqrt{n_k}-1} \mathcal{J}^{-1}_{\varphi_k}(\hat{\varphi}_k) \right) \qquad (16)$$

meaning that the prior becomes nearly flat very quickly, yet it remains proper. Hence, the inherent in the BIC strategy of centering the prior at the ML estimate instead of the parameters of a pre-trained model becomes less important. The rate that the variance of the prior grows with $n_k$ is what counts, at least for moderate sample sizes.

## 4. EXPERIMENTAL RESULTS

The experiments are based on the 2002 NIST Rich Transcription set (NIST-02) and the ESTER SD benchmark. The algorithm we use is the step-by-step approach described in [9]. All the criteria are provided with the same segmentation file in order to focus on the AHC stage. Note that the results are better compared to those we reported in [5] due to a more precise tuning of the parameters of the segmentation stage. No Viterbi re-alignment is applied. We use 18-dimensional static mfcc augmented by the log-energy. The implementation is based on the *open-source* software provided by the LIUM Laboratory, [9].

To compare the criteria, we use the Overall Speaker Diarization Error Rate (DER, %) as well as the average cluster purity (*acp*) vs. average speaker purity (*asp*) trade-off. For details about these metrics we refer to [10]. The formula of the Segmental-BIC with Jeffreys' priors (denoted by Segmental-BIC$_c$) can be found in [5].

We first examine the NIST-02 set. It consists of 6 shows, of 10 minutes each and the *acp-asp* curves are illustrated in Fig. 1. The next experiment is based on the ESTER Speaker Diarization Benchmark. The benchmark consists of 32 shows from various France Radio Channels. The shows are divided to development (14 shows, about 8 hours total duration, denoted by ESTER-D) and test set (18 shows, about 10 hours total duration, denoted by ESTER-T). The *acp-asp* curves on the ESTER-D set are illustrated in Fig. 2.

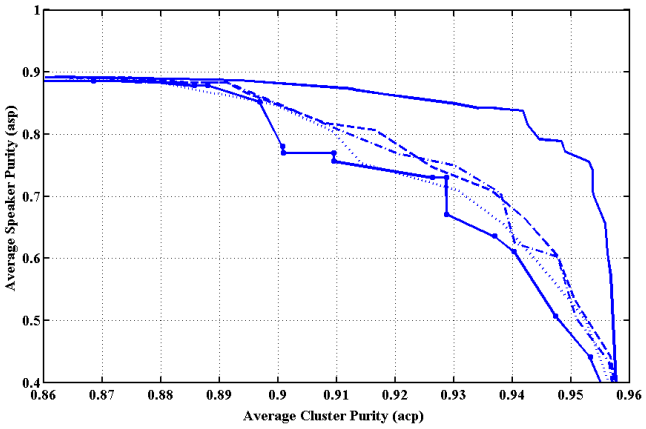


**Fig. 1**. *acp vs. asp on the 2002 NIST BN shows. Solid with dots: Global-BIC, Dotted: Local-BIC, Dash & dots: Segmental-BIC with normal priors, Dashed: Segmental-BIC with Jeffreys' priors, Solid: Segmental-SR-BIC*

The range of $\lambda$ for the criteria examined was $[0.9, 11.0]$ apart from the Segmental-SR-BIC, which scales in $[0.015, 0.200]$. The minimum overall diarization error rates for each set separately are shown in Table 1. Clearly, the new penalty term outperformed the other approaches.

In order to examine the repeatability of the results, we used the $\lambda$ that gave the best results on the ESTER-D for each criterion. The results are shown in Table 2, where the optimum value of $\lambda$ is also given. All the experiments demonstrate the superiority of the modified penalty term and justify the analysis of Sin and White about the rate the penalty should grow.

We should also mention that the operational points at which the minimum DER is attained differ across the criteria. The Local-BIC

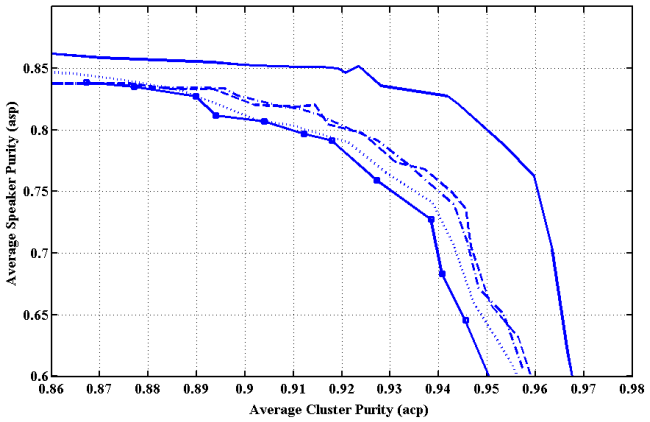


**Fig. 2**. *acp vs. asp on the ESTER development data. Solid with dots: Global-BIC, Dotted: Local-BIC, Dash & dots: Segmental-BIC with normal priors, Dashed: Segmental-BIC with Jeffreys' priors, Solid: Segmental-SR-BIC*

reaches its minimum DER usually by underestimating the true number of speakers, while the Segmental-SR-BIC by overestimating it. Hence, the Segmental-SR-BIC results are improvable, possibly by appending the MAP-adapted GMM-UBM scheme described in [8] or other approaches. On the contrary, the minimum DER of the Local-BIC is reached at operational points of low *acp*, meaning that it cannot be improved using further bottom-up clustering schemes.

**Table 1**. Minimum Overall Speaker Diarization Error Rate (%) for the three sets

| | NIST-02 | ESTER-D | ESTER-T |
|---|---|---|---|
| Global-BIC | 13.07 | 18.84 | 22.46 |
| Local-BIC | 12.99 | 17.37 | 17.47 |
| Segmental-BIC | 12.88 | 17.53 | 20.05 |
| Segmental-BIC$_c$ | 12.71 | 17.25 | 19.46 |
| Segmental-SR-BIC | 11.09 | 13.80 | 14.17 |

**Table 2**. Overall Speaker Diarization Error Rate (%) based on the tuning derived from the ESTER-D set

| | NIST-02 | ESTER-T | $\lambda$ |
|---|---|---|---|
| Global-BIC | 16.41 | 23.02 | 4.68 |
| Local-BIC | 14.03 | 18.21 | 5.05 |
| Segmental-BIC | 14.28 | 20.89 | 6.89 |
| Segmental-BIC$_c$ | 13.89 | 20.11 | 5.78 |
| Segmental-SR-BIC | 12.36 | 14.17 | 0.139 |

## 5. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a new penalty term of the BIC. After providing some intuition about the Segmental-BIC, we investigated the use of a penalty term that grows faster that logarithmically with the number of observation. The motivation was to retain the main principle of the Segmental-BIC and comply with the general constrains for consistency proposed by Sin and White. The experiments prove the superiority of the new criterion, both in terms of average cluster/speaker purity and Diarization Error Rate.
As a future work, we plan of incorporating the temporal information by attaching informative priors over the space of partitions, as well as testing it on SD-for-meetings benchmarks, too.

## 6. ACKNOWLEDGEMENTS

This work is funded by the Greek General Secretariat of Research and Technology under the program PENED-03/251.

## 7. REFERENCES

[1] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, 1978.

[2] R. E. Kass and L. Wasserman, "A Reference Bayesian test for nested hypotheses and its relation to the Schwarz criterion," *Journal of the American Statistical Association*, vol. 90, pp. 928–934, 1995.

[3] C. Fraley and A. Laftery, "How many clusters? Which clustering method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, vol. 41, no. 8, 1998.

[4] S. Chen and P. Gopalakrishnam, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[5] T. Stafylakis, V. Katsouros, and G. Carayannis, "Redefining the Bayesian Information Criterion for speaker diarisation," in *Proceedings of Interspeech*, 2009.

[6] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 1557–1565, 2006.

[7] C.-Y. Sin and H. White, "Information criteria for selecting possibly misspecified parametric models," *Journal of Econometrics*, vol. 71, no. 1-2, pp. 207–225, 1996.

[8] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Improving speaker diarization," in *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, November 2004.

[9] P. Deleglise, Y. Esteve, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based System for French Broadcast News," in *Proceedings of Interspeech, Lisbon, Portugal*, 2005.

[10] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved Unknown-Multiple Speaker clustering using HMM," IDIAP/EPFL, Tech. Rep., 2002.