

CLOSED-FORM EXPRESSIONS VS. BIC: A COMPARISON FOR SPEAKER CLUSTERING

Themost Stafylakis^{1,2}, Xavier Anguera³, Vassilis Katsouros¹ and George Carayannis^{1,2}

¹Institute for Language and Speech Processing, Greece, ²National Technical University of Athens, Greece,

³Telefonica Research, Barcelona, Spain

{themosst,vsk,gcara}@ilsp.athena-innovation.gr & xanguera@tid.es

ABSTRACT

In this paper, the use of closed-form expressions is compared to the BIC approximation, with respect to speaker clustering. We first show that the particular BIC setting which is commonly used in this task, namely the approximation of the marginal - with respect to the model parameters - and conditional - with respect to the latent variables - likelihood, belongs to an exponential family, and hence admits a closed-form expression by attaching conjugate priors. We then formalize the role of the tuning parameter as a hyperparameter of the prior and finally we explain the several proposed setting - global, local and segmental - based on the strength of the prior. Experiments are carried out for the speaker clustering task and improvement over the BIC approximation is reported.

Index Terms— Clustering methods, Bayesian methods, Speaker recognition

1. INTRODUCTION

The tasks of speaker segmentation and clustering, jointly termed as speaker diarization have largely been dominated by the use of the Bayesian Information Criterion (BIC), [1]. The BIC offered an intuitive and effective baseline test to obtain a fast point-estimate about the partition of a dialogue into speakers. The step-by-step approaches to speaker diarization (i.e. the strategy of applying a speaker change detector to segment the audio stream into speaker turns and then group the segments according to their statistical similarities, usually using Agglomerative Hierarchical Clustering) offer a good compromise between computational effort, modularity and accuracy. Many algorithms utilize this strategy in order to initialize more complicated strategies, for examples using MAP-adapted GMMs, [2].

The BIC however is only an approximation to the integrated log-likelihood, [3]. Therefore, it would be interesting to investigate if any gain in accuracy can be attained by using its closed-form expression instead. To come up with this expression, one needs to attach the conjugate prior of the Normal distribution (i.e. the Normal-Inverse Wishart distribution) and integrate out the parameters of the emission probabilities, i.e. $\varphi_k = (\mu_k, \Sigma_k)$, $k = 1, \dots, K$, assuming a partition into K speakers. Such an approach will also allow one to include the tuning parameter λ as an extra hyperparameter that controls the amount of virtual observations, commonly termed as the strength of the prior. Note that the inclusion of λ is compulsory in order to obtain a speaker-level partition of the data. This is because the distribution of a speaker is far from being Normal, at least in the MFCC and relative domains. The huge misspecification of the model leads to an enormous overestimation of the true number of speakers by the BIC.

The closed-form expression also allows us to examine from a

Bayesian perspective the penalty terms of the several BIC variants with respect to the prior. In fact, many such issues relative to the use of BIC still lack of a coherent analysis and are usually considered as heuristics. Novel approaches, based on Dirichlet processes ([4]) and inferential paradigms such as Variational Bayes ([5]), have created higher standards in terms of statistical reasoning compared to the step-by-step BIC-based algorithms. An examination of the BIC in a more Bayesian reasoning is one of the main aims of this paper. We show that by deriving the closed-form expression, both the tuning parameter and several particular settings (global, local and segmental, [6]) can now be explained more formally, and can fit into a coherent Bayesian framework. The implied graphical model of these settings will also be examined. Finally, the normal distributions will be modeled as the emission probabilities of an HMM, and a prior over the partition space will also be considered.

The rest of the paper is organized as follows. In Sect. 2, a Bayesian formulation of the problem is presented, along with the closed-form expression. In Sect. 3, the several BIC settings and the role of the tuning parameter is discussed with respect to the closed-form expression. In Sect. 4, we provide some experiments on speaker clustering based on the ESTER benchmark, while our conclusions and future work directions are given in Sect. 5.

2. MODELING, CLOSED-FORM EXPRESSIONS AND BIC

2.1. Modeling and statistical quantities

We denote the set of observation vectors that comprise an audio document by $\mathbf{y} = [\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)}]^T$, $\mathbf{y}^{(i)} \in \mathbb{R}^d$ while the corresponding latent variables by $\mathbf{s} = [s^{(1)}, \dots, s^{(n)}]^T$, $s^{(i)} = 1, 2, \dots$, i.e. the cluster indicators. Using the HMM terminology, \mathbf{s} is the state sequence that we are trying to estimate. (The terms cluster and state will be used interchangeably). To do so, we maximize the posterior of the joint distribution (\mathbf{s}, K) , where K is the order of the HMM.

We choose to maximize the evidence over this joint space for two main reasons. First or all, note that the diarization task is rarely treated as a model order selection task, where one typically tries to infer K by integrating out the latent variables along with the model parameters. On the contrary, it is regarded as a latent variable estimation task, since the official scoring method (Diarization Error Rate, DER) is defined on the space of state sequences $\mathbf{s} \in \mathcal{S}^n$ and not on the order of the model. Therefore, from a decision theoretic perspective, it would be consistent to define the loss function on the space of state-sequences and not directly on the model's order. Moreover, the model order selection task is an intractable one, and the consistency of BIC has never been proven for HMMs, [7]. The second reason is the ambiguity about the variable K . Consider a state sequence \mathbf{s} of K' speakers (i.e. $\max(\mathbf{s}) = K'$) and note that such a sequence can be generated by any HMM of K states, where $K \geq K'$. By view-

ing \mathbf{s} as a random draw from an K -order HMM, clearly there is no constraint that each draw will visit all the K states. Hence, a precise calculation of the posterior of \mathbf{s} should involve the averaging with respect to all compatible with \mathbf{s} orders of the HMM, which is not a trivial problem due to degeneracy of the orders $K > K'$. Therefore, we choose to maximize the posterior of the joint event (\mathbf{s}, K) , and consider only cases where the order of the HMM matches to the number of speakers in \mathbf{s} , i.e. $K = \max(\mathbf{s})$.

2.2. Posterior and integrated likelihood

Having clarified that we aim to maximize the posterior of the joint event (\mathbf{s}, K) , we decompose it as follows

$$P(\mathbf{s}, K | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{s}, K) P(\mathbf{s}, K) \quad (1)$$

where

$$p(\mathbf{y} | \mathbf{s}, K) = \int_{\Phi} p(\mathbf{y} | \varphi, \mathbf{s}, K) \pi(\varphi | \mathbf{s}, K) d\varphi \quad (2)$$

is the integrated likelihood of a K -order HMM, conditioned on the state sequence \mathbf{s} . The last term of the r.h.s. of (1) is the prior of (\mathbf{s}, K) , and can be further decomposed as follows

$$P(\mathbf{s}, K) = P(\mathbf{s} | K) P(K) \quad (3)$$

where the conditioning on the hyperparameters is kept implicit. Note that the set of parameters we integrate out in (2), usually termed as the internal parameters of the model, does not include the initial probabilities $\alpha_0 = \{\alpha_{0k}\}_{k=1}^K$ nor the transition matrix A of the HMM. This is due to the conditioning of the integrated likelihood on \mathbf{s} which makes \mathbf{y} conditionally independent from $A^+ = [\alpha^T, A^T]^T$ given \mathbf{s} . The latter parameters of the HMM, usually termed as the external parameters of the model, should be marginalized from (3) as follows

$$P(\mathbf{s} | K) = \int_{A^+} P(\mathbf{s} | A^+, K) \pi(A^+ | K) dA^+ \quad (4)$$

The above quantity should be considered as a prior over the state sequence and is usually being omitted from most of the step-by-step diarization algorithms. A fully Bayesian modeling however should include it, since it is a quantity that scales with n . Hence, we include it in our analysis, despite of being at least one order of magnitude below the magnitude of the internal parameters. A closed-form expression can be derived by attaching Dirichlet priors over each line of A^+ . Finally, we treat $\pi(K)$ as flat over a reasonable range $K \in [K_{\min}, K_{\max}]$, although a Poisson prior may also be considered, given n .

2.3. Priors and closed-form expressions

Since both $p(\mathbf{y} | \varphi, \mathbf{s}, K)$ and $P(\mathbf{s} | A^+, K)$ belong to an exponential family, the parameters can be marginalized and obtain closed form expressions by using conjugate priors. Let us begin with (2) and obtain its closed form expression. For notational simplicity we first assume the trivial case where \mathbf{y} forms a single cluster, denoted by $\mathbf{s}_0 = [1, 1, \dots, 1]$. To do so, we define the prior of μ to be normal, i.e.

$$\mu \sim \mathcal{N}(\mu_0, \frac{1}{\nu} \Sigma) \quad (5)$$

meaning that we center the prior of the mean value on μ_0 , with prior uncertainty given by $\frac{1}{\nu} \Sigma$. In our experiments, we set μ_0 equal to the

mean value of the audio file we examine. By integrating out μ and defining $\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}^{(i)}$ we obtain

$$p(\mathbf{y} | \Sigma) = \left(\frac{\nu}{n + \nu} \right)^{d/2} (2\pi)^{-nd/2} |\Sigma|^{-n/2} \exp \left(-\frac{1}{2} \text{tr}(\Sigma^{-1} P) \right) \quad (6)$$

where

$$P = \sum_{i=1}^n (\mathbf{y}^{(i)} - \bar{\mathbf{y}})(\mathbf{y}^{(i)} - \bar{\mathbf{y}})^T + \frac{n\nu}{n + \nu} (\mu_0 - \bar{\mathbf{y}})(\mu_0 - \bar{\mathbf{y}})^T \quad (7)$$

and

$$m = \frac{1}{n + \nu} (n\bar{\mathbf{y}} + \nu\mu_0) \quad (8)$$

In order to derive $p(\mathbf{y} | K)$ we should attach to Σ the Inverse-Wishart prior, i.e.

$$\Sigma \sim \mathcal{IW}(\Psi, p) \quad (9)$$

meaning that we set the a priori expected value of Σ equal to $\frac{\Psi}{p-d-1}$, while $p > d - 1$ is equal to the number of virtual observations. In our experiments, we set $p \geq d + 3$ to ensure that $\text{var}(\Sigma_{i,j}) < +\infty$ for each (i, j) entry, while keeping the prior as vague as possible. By integrating out Σ and after some matrix algebra, we finally obtain

$$p(\mathbf{y} | \mathbf{s}_0) = \left(\frac{\nu}{n + \nu} \right)^{\frac{d}{2}} \pi^{-\frac{nd}{2}} \frac{|\Psi|^{\frac{p}{2}}}{|\Psi + P|^{\frac{n+p}{2}}} \frac{\Gamma_d(\frac{p+n}{2})}{\Gamma_d(\frac{p}{2})} \quad (10)$$

where $\Gamma_d(x)$ the d -variate Gamma function. Having obtained the expression for the single state case, it is straightforward to show that in the general case where $K \geq 1$, the integral yields

$$p(\mathbf{y} | \mathbf{s}, K) = \prod_{k=1}^K \left(\frac{\nu_k}{n_k + \nu_k} \right)^{\frac{d}{2}} \pi^{-\frac{n_k d}{2}} \frac{|\Psi|^{\frac{p}{2}}}{|\Psi + P_k|^{\frac{n_k+p}{2}}} \frac{\Gamma_d(\frac{p+n_k}{2})}{\Gamma_d(\frac{p}{2})} \quad (11)$$

where P_k as in (7) using only $\mathbf{y}_k = \{\mathbf{y}^{(i)} : s^{(i)} = k\}$. Finally, we derive the closed-form expression of (4) by attaching Dirichlet priors over each line of the augmented transition matrix A^+ , i.e.

$$\alpha_k \sim \text{Dir}(\mathbf{q}_k), k \in [0, K] \quad (12)$$

where $Q^+ = [\mathbf{q}_0^T, \dots, \mathbf{q}_K^T]^T$, and its k th line is a k -dimensional vector that denotes the number of virtual observation (plus one) that depart from the k th state. We set each entry of Q^+ equal to 1 (i.e. zero virtual transition) apart from the diagonal of Q (i.e. the virtual self-transitions), which we set $Q_{kk} = h/K + 1, k \in [1, K], h = 0.1n$. This different treatment is adopted in order to incorporate our prior knowledge regarding the state-persistence of a Markov chain, when each state corresponds a distinct speaker.

By integrating out A^+ we obtain the following expression

$$P(\mathbf{s} | K) = \prod_{k=0}^K \frac{\mathcal{B}(\mathbf{b}_k)}{\mathcal{B}(\mathbf{q}_k)}, \quad \mathcal{B}(\mathbf{q}_k) = \frac{\prod_{l=1}^K \Gamma(q_{kl})}{\Gamma(\sum_{l=1}^K q_{kl})} \quad (13)$$

Finally, $P(\mathbf{s} | K)$ should be multiplied by $\Gamma(K + 1)$, in order to deal with the exchangeability of the prior with respect to the label of the states.

3. BIC SETTINGS AS ASYMPTOTICS OF THE CLOSED-FORM EXPRESSIONS

3.1. Global-BIC formula and the tuning parameter

Let us now show how we can examine the several BIC settings from a Bayesian perspective. This includes the global, local and segmental settings along with the tuning parameter. The global-BIC formula is as follows

$$\text{BIC}_{\mathbf{s},K|\mathbf{y}}^G = \sum_{i=1}^n p(\mathbf{y}^{(i)}|\hat{\varphi}_{s(i)}) - \frac{1}{2} \sum_{k=1}^K (\lambda \mathcal{P}_\mu \log n + \mathcal{P}_\Sigma \log n) \quad (14)$$

where \mathcal{P}_μ and \mathcal{P}_Σ denote the number of parameters of μ and Σ , respectively. Let also $\mathcal{P} = \mathcal{P}_\mu + \mathcal{P}_\Sigma$. Considering normal distributions with full covariance matrices, these quantities are equal to $\mathcal{P}_\mu = d$ and $\mathcal{P}_\Sigma = \frac{d(d+1)}{2}$. Note that for reasons that will be clarified below, we apply λ only to \mathcal{P}_μ . From (10), one may derive a coherent interpretation for λ . The key-quantity is the strength of the priors of $\{\mu_k\}_{k=1}^K$, denoted by $\{\nu_k\}_{k=1}^K$. From the equation below

$$\frac{1}{2} \sum_{k=1}^K \lambda \mathcal{P}_\mu \log n = -\log \prod_{k=1}^K \left(\frac{\nu_k}{n_k + \nu_k} \right)^{\frac{d}{2}} \quad (15)$$

it is straightforward to verify that the global-BIC implies $\nu_k = \frac{n_k}{n^{\lambda-1}}$. For $\lambda = 1$ and $n_k \gg 1$ we obtain $\nu_k \approx \frac{n_k}{n^\lambda}$, which is the central property of the global-BIC. The prior of each μ_k shares an amount of a single d -dimensional virtual observation, proportionally to $w_k = \frac{n_k}{n}$. Note though that the prior is clearly sample size dependent, even for $\lambda = 1$. In Directed Acyclic Graphs (DAG) this dependency implies the existence of an edge from the latent variables to the hyperparameters of φ , which is rarely present in the relevant bibliography, [3]. Let us also consider the usual case in diarization where $\lambda > 1$. It is clear that now the priors share jointly a fraction of $n^{1-\lambda}$ approximately of a single virtual observation, proportionally to their sample size. This means that the prior of a fixed cluster of n_k observations becomes more and more noninformative as the overall sample size n increases.

The above interpretation of λ with respect to the implied strength of the prior clarifies why we chose to apply it only to \mathcal{P}_μ . Contrary to μ_k , the prior of Σ_k cannot be as vague as we desire to. This is because the Inverse-Wishart distribution requires a minimum amount of virtual observation (a.k.a. degrees of freedom) in order to be proper, $p > d - 1$. Therefore, for consistency with the interpretation we gave to λ , we should not multiply \mathcal{P}_Σ with it. This property of the prior of Σ_k brings about a limitation of the global-BIC setting. Since the pool of virtual observations is being shared between the priors of $\{\Sigma_k\}_{k=1}^K$ proportionally to $\{n_k\}_{k=1}^K$, we should either a priori fix the overall strength and require a minimal $n_k \geq n_{\min}$, where n_{\min} is determined from the size of the pool, or re-estimate the size of the pool, each time we score a partition \mathbf{s} having one or more n_k below n_{\min} .

The above limitations lead us to alter the standard penalty term of (14) in the following way

$$T_{eq}^s = K \left(\frac{\mathcal{P}_\mu}{2} (\lambda - 1) \log n - \frac{\mathcal{P}_\Sigma}{2} \log p_{\min} \right) + \sum_{k=1}^K \frac{\mathcal{P}}{2} \log K n_k \quad (16)$$

The rationale of T_{eq}^s is to remove the edge by sharing the virtual observations equally between clusters (compared to proportionally to n_k), and consequently to overcome the limitation discussed above. One may verify that the hyperparameters implied by T_{eq}^s are equal

to $\nu_k \approx \frac{n^{1-\lambda}}{K}$ and $p = p_{\min} \frac{K_{\max}}{K}$. Notice that the overall strength (i.e. $\sum_{k=1}^K \nu_k + Kp$) is preserved, given n .

3.2. Local and segmental settings

We now examine two other settings of BIC, the local and the segmental. The local-BIC setting cannot fit into the analytical framework we presented above, since it is only a Δ BIC formula and not a Bayesian strategy capable of scoring the evidence of overall partitions. However, it exhibits an interesting property that will give rise to the segmental approach. Assume two clusters and consider the gain in log-evidence we obtain by merging them. Denoting by \mathcal{H}_0 and \mathcal{H}_1 the unique and the two-cluster hypotheses respectively, we obtain

$$\Delta \text{BIC} = l(\hat{\varphi}_a, \hat{\varphi}_b|\mathcal{H}_1) - l(\hat{\varphi}|\mathcal{H}_0) - \frac{1}{2} (\lambda \mathcal{P}_\mu \log \tilde{n} + \mathcal{P}_\Sigma \log \tilde{n}) \quad (17)$$

where $\tilde{n} = n$ and $\tilde{n} = n_a + n_b$ for the global and local BIC respectively. The local-BIC simply focuses on the pair of clusters being examined and ignores the rest of the audio document. This approach, although reasonable for the segmentation task, has no global objective function to optimize and therefore poses severe algorithmic restrictions for the clustering task. Note also that it suffers from the same problem with the global-BIC, although we may overcome it by using the refined penalty term, given in (16), for $K = \{1, 2\}$. We denote this approach by local-BIC_{eq}. Despite its limitations though, its superiority against the global-BIC in many benchmark tests led to the so-called segmental-BIC approach. The central idea of the latter is to mimic a basic principle of the local-BIC, while retaining the capacity of the criterion to score overall partitions. This principle is to have a Δ BIC formula that depends only on statistical quantities of the pair of clusters being examined, i.e. to eliminate the dependency on n that the global Δ BIC formula exhibits. In the analysis of the global-BIC presented above, we showed that this property of the global-BIC is due to the strength of its implied prior, i.e. the conservation of the size of the pool of virtual observation, given n . However, a more principled approach to eliminate such a dependency compared to the local-BIC is to allow the size of this pool to grow with K . The segmental-BIC (plain variant) is as follows

$$\text{BIC}_{\mathbf{s},K|\mathbf{y}}^S = \sum_{i=1}^n p(\mathbf{y}^{(i)}|\hat{\varphi}_{s(i)}) - \sum_{k=1}^K \left(\lambda \frac{\mathcal{P}_\mu}{2} \log n_k + \frac{\mathcal{P}_\Sigma}{2} \log \frac{n_k}{p_{\min}} \right) \quad (18)$$

while its square-root variant has the same formula, by placing $\lambda = n_k^{1/2} \lambda'$. For both formulae, it is straightforward to obtain their Δ BIC expressions and verify their independence from n .

Let us now fit the segmental-BIC approximation to the closed-form expressions. By equating the *r.h.s* of (15) with the corresponding penalty term, we may derive that the strength of the implied prior of

μ_k is $\nu_k = \frac{n_k^{1-\lambda}}{1-n_k^{1-\lambda}} \approx n_k^{1-\lambda}$, for $n_k \gg 1$. Using $\lambda = 1$ we simply obtain one virtual observation per cluster, i.e. a linear increase of the pool with K . Similarly, the pool of virtual observations for $\{\Sigma_k\}_{k=1}^K$ increases linearly with K , since each Σ_k shares p_{\min} such observations to form its prior, independently of n_k .

Finally, for the square-root variant the strength is $\nu_k \approx n_k^{1-\lambda'} \sqrt{n_k}$, i.e. it asymptotically tends to zero in a much higher rate. This strength results in a much stricter penalty term, capable of dealing with the high variability between segments of a single speaker.

4. EXPERIMENTAL RESULTS

The experiments are based on the ESTER Speaker Diarization (SD) benchmark, [2]. The corpus consists of 32 shows from various France Radio Channels. The shows are divided to development (14 shows, about 8 hours total duration, denoted by DEV) and test set (18 shows, about 10 hours total duration, denoted by TEST). The algorithm we use is the step-by-step approach described in [2]. All the criteria are provided with the same segmentation file in order to focus on Hierarchical Clustering. A GMM classifier has been applied to discard the non-speech segments while 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, [2].

To compare the criteria, we first use the average cluster purity (*acp*) vs. average speaker purity (*asp*) trade-off. The *acp-asp* curves on the DEV set are illustrated in Fig. 1. The second validation scheme is the Overall Speaker Diarization Error Rate (DER, %). We used the DEV set to estimate the optimal (in terms of DER) value of λ for each criterion, and then evaluate it on the TEST set. The performance of Local-BIC_{eq} was very similar to Local-BIC_{eq} with closed-form expression, and therefore is not depicted. The results are shown in Table 1. The hyperparameters are set to $\mu_0 = \bar{\mathbf{y}}$ and $\Psi = \frac{p}{n} \sum_{i=1}^n (\mathbf{y}^{(i)} - \bar{\mathbf{y}})(\mathbf{y}^{(i)} - \bar{\mathbf{y}})^T$. As expected, the results exhibited little sensitivity to this set of hyperparameters. Fixed-hyperparameters (i.e. independent of \mathbf{y}) may be considered as well. The results show that an increase in performance is feasible, compared to the BIC approximation.

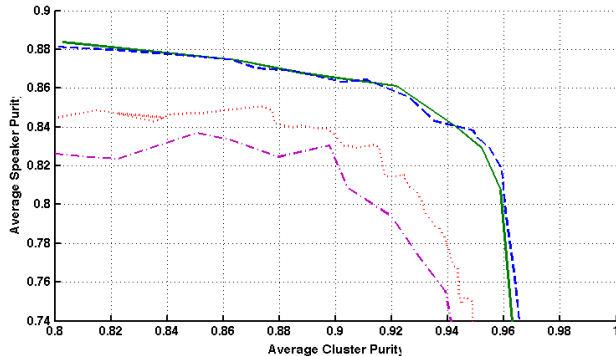


Fig. 1. *ACP vs. ASP on the ESTER development data. Magenta-Dashed-Dotted line: local-BIC, Red-Dotted line: Local-BIC_{eq}, Solid-Green line: Segmental-sqrt-BIC, Blue-Dashed line: Segmental-sqrt-closed-form expression*

Table 1. Overall Speaker Diarization Error Rate (%) on ESTER

Criterion	DEV	TEST	DEV	TEST
	BIC approx.		Closed-forms	
Local-BIC	15.76	16.28	-	-
Local-BIC _{eq}	15.49	16.44	15.21	16.14
Segmental-BIC	16.43	18.37	16.32	18.07
Segmental-sqrt-BIC	12.78	13.27	13.26	13.12
False Alarm Rate	0.3	0.6	0.3	0.6
Missed Speech Rate	0.9	1.2	0.9	1.2

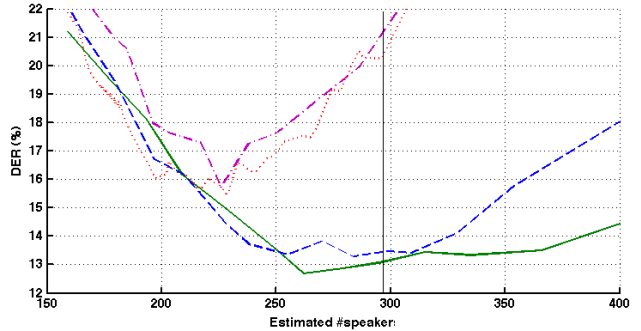


Fig. 2. *Estimated number of speakers vs. DER on the ESTER development data. The vertical line depicts the true number of speakers. Colors and line-styles: same as Fig.1*

5. CONCLUSIONS AND FUTURE WORK

In this paper, we derived the closed-form expression of the integrated likelihood, and compared their performance to the one attained by the BIC approximation. These expression allows us to give a Bayesian explanation to several BIC setting, along with the tuning parameter. We demonstrated that the key-quantity distinguishing these approaches is the strength of the parameters' prior. The experiments showed that an increase in performance can be attained by using these expressions, compared to the BIC approximation.

As a future work, we are planning to apply similar ideas to GMMs, by considering the complete-data likelihood, as well as examining the above expressions to newer speaker-recognition features, such as the *i*-vectors.

6. REFERENCES

- [1] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, 1978.
- [2] X. Zhu, C. Barras, S. Meignier, and J. Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization," in *Proceedings of Interspeech*, September 2005, pp. 2441 – 2444.
- [3] 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.
- [4] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "The Sticky HDP-HMM: Bayesian nonparametric Hidden Markov Models with Persistent States," 2009.
- [5] F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, September 2005.
- [6] T. Stafylakis, V. Katsouros, and G. Carayannis, "The Segmental Bayesian Information Criterion and its applications to Speaker Diarization," *IEEE Selected topics in Signal Processing*, pp. 857 – 866, October 2010.
- [7] G. Celeux and J.-B. Durand, "Selecting hidden Markov model state number with cross-validated likelihood," *Computational Statistics*, vol. 23, pp. 541–564, 2008.