Improvements to the equal-parameter BIC for Speaker Diarization

Themos Stafylakis[1,2]*, Xavier Anguera*[3]

[1]Institute for Language and Speech Processing, Athena Research Center, Greece
[2]National Technical University of Athens, Greece
[3]Multimedia Research Group, Telefonica Research, Barcelona, Spain

themosst@ilsp.athena-innovation.gr, xanguera@tid.es

Abstract

This paper discusses a set of modifications regarding the use of the Bayesian Information Criterion (BIC) for the speaker diarization task. We focus on the specific variant of the BIC that deploys models of equal - or roughly equal - statistical complexity under partitions of different number of speakers and we examine three modifications. Firstly, we investigate a way to deal with the permutation-invariance property of the estimators when dealing with mixture models, while the second is derived by attaching a weakly informative prior over the space of speaker-level state sequences. Finally, based on the recently proposed segmental-BIC approach, we examine its effectiveness when mixture of gaussians are used to model the emission probabilities of a speaker. The experiments are carried out using NIST rich transcription evaluation campaign for meeting data and show improvement over the baseline setting.

Index Terms: Bayesian Information Criterion, Speaker Diarization, Clustering Algorithms

1. Introduction

Speaker diarization (SD) is the problem of assessing the state (i.e. speaker)-sequence of a recording (broadcast news show, meeting, a.o.), when no information about the participant speakers nor their cardinality is given a priori. Among the several inferential paradigms that have been proposed in literature, and including both frequentist and bayesian statistical settings, the BIC-based agglomerative hierarchical clustering (AHC) algorithms have become the dominant approach, due to their simplicity and intuitiveness. The BIC is a reference test that has both bayesian and information-theoretical justification, and yields a baseline tool for applying model selection and averaging, [1].

The purpose of this paper is to examine ways to enhance the performance of a penalty-free version of the ΔBIC, proposed in [2]. With this particular approach, one avoids the use of a penalty term by fixing the total number of gaussians used to model each state sequence. Hence, the two penalty terms of ΔBIC, i.e. the BIC value of the two-speaker model (alternative hypothesis $\mathcal{H}_1$) minus the BIC value of the single speaker model (null hypothesis, $\mathcal{H}_0$) cancel out each other, yielding a robust approach to speaker clustering. After an introduction regarding the use of the BIC for SD in Section 2, we proceed with three proposed modification. The first is based on the likelihood's property of being permutation invariant when dealing with mixture models, while the next one is a straightforward derivation of a prior for each candidate state sequence. The prior assumes an HMM as the generative model for the state sequences and yields a closed form expression that is capable of penalizing abrupt transitions and HMMs of large state-cardinality if reasonable hyperparameters are chosen. Both of these modification are analyzed in Section 3. The final proposed modification is examined in Section 4, and is based on the fact that while the overall components of the GMMs remain fixed under the rival models, the overall statistical complexity does not, since the weights of each GMM lie on a simplex. The derived term will be examined also using a different approach - the segmental one - that implies prior densities of fixed-strength for the parameters of each cluster (i.e. for each speaker model) instead of fixing the strength of the parameters of the overall model, [3]. Finally, the experiments are presented in Section 5.

2. Issues regarding the use of the BIC

2.1. Integrating out the GMM-level parameters

In this section we examine some issues with respect to the use of the BIC in DS. Let us denote the observed variables (i.e. the incomplete data) by $\mathbf{y} = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}] \in \mathcal{Y}^n \subseteq \Re^{d \times n}$, where $i = 1, \ldots, n$ stands for the time index and d is the dimensionality of the feature space. We deploy the BIC in order to approximate the conditional (with respect to a state sequence $\mathbf{s} \in \mathcal{S}^n$) - marginal (with respect to the GMM-level parameters $\varphi \in \Phi^K$) log-likelihood of the model give the data $\mathbf{y}$, i.e. $\exp(BIC_{\mathbf{y}|\mathbf{s}}) \approx p(\mathbf{y}|\mathbf{s}) = \int_{\Phi^K} p(\mathbf{y}, \varphi|\mathbf{s}) d\varphi$. The emission (or GMM-level) parameters of the HMM are denoted by $\varphi = (\varphi_k)_{k=1}^K \in \Phi^K$, $\varphi_k = (w_k^m, \mu_k^m, \Sigma_k^m)_{m=1}^{M_k}$ and M_k the order of the GMM that is used to model the emission probabilities of the kth state. The order of the model can implicitly be extracted from $\mathbf{s}$ by the operator $K = \max(\mathbf{s})$. Since $\mathbf{y}$ are conditionally independent given $\mathbf{s}$, the above quantity can be approximated without involving the state-transition probabilities of the HMM. Hence, in order to approximate $p(\mathbf{y}|\mathbf{s})$ we are only integrating out the parameters of the HMM that are modeling the emission probabilities of each state (i.e. each speaker).

The BIC is based on the Laplace method for approximating integrals, i.e. the idea of considering $p(\mathbf{y}, \varphi|\mathbf{s})$ as being approximately normal around its MAP (or ML $\hat{\varphi}$ for large n) mode, and furthermore discarding terms that do not scale with the sample size, [6]. Doing so, and assuming a prior $\pi(\varphi)$ of fixed strength (i.e. that carries a fixed amount of virtual observations independently of K) we end-up with the familiar expression

$$BIC_{\mathbf{y}|\mathbf{s}} = \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\hat{\varphi}_{s^{(i)}}) - \frac{1}{2}\mathcal{P}_{in}^{\mathbf{s}} \log n \qquad (1)$$

where $\mathcal{P}_{in}^{\mathbf{s}}$ the amount of free GMM-level parameters of the model. By fixing $\mathcal{P}_{in}^{\mathbf{s}}$, the last term on the *r.h.s.* of (1) becomes independent of $\mathbf{s}$ and K, allowing as to proceed only with the

log-likelihood terms. This particular setting of BIC was introduced in [2] and we will be referring to it as the equal-parameter BIC.

2.2. Density Estimation vs. Cluster Analysis

As briefly explained above, the use of the BIC in SD is related to model based cluster analysis rather that model selection for density estimation. This means that we are trying to assess the MAP state sequence rather than the implied number of speakers directly. An estimation based on the $K = \max(\mathbf{s})$ operation does not coincide with a direct maximization of $\pi(K|\mathbf{y}) \propto \pi(K)p(\mathbf{y}|K)$. The cluster analysis setting we adopt is also in compatibility with the loss function usually used, the Diarization Error Rate, i.e. a loss function over the state-sequence space and not over the order of the model. However, in the model selection setting, placing an informative prior over the order of the model $\pi(K)$ is rather useless, since the observations will overwrite it even for moderate n, [1]. Nevertheless, in the chosen setting of the BIC, the corresponding prior $\pi(\mathbf{s}, K) = \pi(\mathbf{s}|K)\pi(K)$ scales with n, since is defined over the state-sequence space. Therefore, it would be interesting to derive an expression for $\pi(\mathbf{s}|K)$ and consider $\pi(K)$ as uniform.
We should also note that the statistical quantity $\pi(s)$ offers complementary information to the BIC. Notice that by considering that the state dynamics of an HMM are completely governed by the $K \times K$ state transition probability matrix $A = \{a_{kl}\}, k = 1, \ldots, K, l = 1, \ldots, K$ and the initial probabilities $\{\alpha_l\}_{l=1}^K$, where $\sum_{l=1}^K a_{kl} = 1$ and $\sum_{l=1}^K \alpha_l = 1$. Combining this result with a natural choice of parameter prior that is separable $\pi(\varphi, A, \alpha) = \pi(\varphi)\pi(A, \alpha)$, we may verify that marginalizing $\pi(\mathbf{s}, A, \alpha)$ with respect to (A, α) has no interaction with the BIC values, since the former approximates $\log p(\mathbf{y}|\mathbf{s})$ which is independent of (A, α).

3. Aliases of the likelihood of a GMM and a prior over the state-sequences

3.1. Dealing with the aliases

Suppose now we are integrating out φ from $p(\mathbf{y}, \varphi|\mathbf{s})$. Since $\mathbf{y}$ are conditionally independent given $\mathbf{s}$, we may focus on a single state of the HMM (say the kth) and let $\mathbf{y}_k = \{\mathbf{y}^{(i)} : s^{(i)} = k\}$ be the observation under the k state given $\mathbf{s}$. The issue regarding the use of the Laplace approximation is that the likelihood of its parameters φ_k has $M_k!$ identical peaks (or aliases), since the ordering of the components is arbitrary, [4]. The problem is also present in the Markov Chain Monte Carlo inferential paradigm, commonly termed as label switching, [5]. We say that the likelihood is invariant under permutations of the labels, yielding a non-identifiable model. This in turn means that the Laplace approximation of the model's evidence cannot be accurate. A straightforward solution is to multiply the evidence of the model by the amount of such aliases that the likelihood exhibits. This approximation is exact only if the peaks of the likelihood are well separated (easily attained for large n) and the prior is permutation-invariant. The implied prior of BIC is the (inherently data-dependent) unit-information priors ([6]) i.e. a gaussian distribution centered at the ML estimate $\hat{\varphi}$ with precision equal to the information carried in a single observation of $\mathbf{y}$. Hence, despite the fact that the implied prior is not permutation-invariant, its strength (only one virtual observation) allows to consider it as vague enough and examine this approach. Hence, we add the term $\log(M_k!) = \log\Gamma(M_k + 1)$ to the log-evidence of a GMM with M_k components.
Thus, if the usual agglomerative hierarchical clustering algorithm is considered, the term

$$C_1 = \log\Gamma(M_k + M_l + 1) - \log\Gamma(M_k + 1) - \log\Gamma(M_l + 1) \quad (2)$$

should be subtracted from the ΔBIC value, defined as the log-evidence of the model under $\mathcal{H}_1$ minus the log-evidence of the model under $\mathcal{H}_0$, for the (k, l) pair of states. These terms, although of low magnitude on average when compared to the difference of the log-evidences, can be beneficial when the ΔBIC values are close to zero.

3.2. An informative prior over the state-sequence

For a more compact notation, we index the non-emitting state by $k = 0$ and the transition matrix becomes $(K + 1) \times K$, i.e. $\alpha_l = a_{0l}$ and we augment $\mathbf{s}$ by $s^{(0)} = 0$. To derive $\pi(\mathbf{s}|K)$, where $K = \max(\mathbf{s})$, the parameters A should be integrated out, i.e.

$$\pi(\mathbf{s}|K, Q) = \int_{\mathcal{A}^{K+1}} p(\mathbf{s}|A, K)\pi(A|K, Q)dA \quad (3)$$

assuming a prior density $\pi(A|K, Q) = \prod_{k=0}^K \pi(\mathbf{a}_k|\mathbf{q}_k)$. As we explain below, Q are the hyperparameters of the prior on A. The density $\pi(\mathbf{a}_k|\mathbf{q}_k)$ denotes the prior on the kth line of A, (i.e. $\mathbf{a}_k = [a_{k1}, a_{k2}, \ldots, a_{kK}]$), over the $(K-1)$-simplex Δ^{K-1} and $\mathcal{A}^{K+1} = \underbrace{\Delta^{K-1} \times \Delta^{K-1} \ldots \Delta^{K-1}}_{K+1}$. A common choice for the prior on $\mathbf{a}_k$ is the Dirichlet distribution, $\mathbf{a}_k \sim \mathcal{D}ir(\mathbf{q}_k)$. The set of hyperparameters $Q = \{\mathbf{q}_k\}_{k=0}^K$ of the Dirichlet is a $(K+1) \times K$ matrix, where $q_{kl} - 1$ is equal to the number of virtual transitions from the kth state to lth state we add to those appear in $\mathbf{s}$. Typically, we may place $q_{kl} = 1$ if $k \neq l$ and $q_{kl} \gg 1$ if $k = l$, in order to bias the results towards self-transition, [8]. Furthermore, we may adopt a compact parametrization $q_{kl} = 1 + \delta(k, l)(\tilde{b}/K - 2)$, i.e. assume a fixed zero number of virtual transitions between different states and $\tilde{b} \gg 1$ a fixed amount of virtual self-transitions, which we may set to $\tilde{b} = 0.1n$.
Since the Dirichlet distribution is the natural conjugate to the multinomial distribution of $\mathbf{a}_k$, a closed-form expression of $\pi(\mathbf{s}|K)$ can be easily derived, that is as follows

$$\pi_{\mathbf{s}}(\mathbf{s}|K, Q) = \prod_{k=0}^K \frac{\mathcal{B}(\mathbf{b}_k)}{\mathcal{B}(\mathbf{q}_k)} \quad (4)$$

where $b_{kl} = q_{kl} + \#\{k \to l|\mathbf{s}\}$ and $\#\{k \to l|\mathbf{s}\}$ is the total number of transitions from state k to state l that occur in $\mathbf{s}$. Moreover, $\mathcal{B}(\cdot)$ is the normalizing constant of the Dirichlet distribution, which is defined in terms of the Gamma function $\Gamma(\cdot)$ as

$$\mathcal{B}(\mathbf{q}_k) = \frac{\prod_{l=1}^K \Gamma(q_{kl})}{\Gamma(\sum_{l=1}^K q_{kl})}. \quad (5)$$

3.3. Permutation invariance of the state sequence

Note again though that the derived prior in (4) is invariant under permutations of the speaker labels. An intuitive way to show it is to consider that while the prior allows a transition $k \to l$ on the ith frame, where $k, l \leq K$ and l is the label we give to a new speaker entry, we arbitrarily label $s^{(i)}$ by

$l = \max(\mathbf{s}_{1,\ldots,i-1}) + 1$. However, the prior allows $s^{(i)}$ to take any label $l \in [\max(\mathbf{s}_{1,\ldots,i-1}) + 1, K]$ with equal probability. To cope with this invariance we should multiply the prior in (4) by $T_{\mathbf{s}} = \Gamma(K+1)$, so that the probability of the first visit to the lth state is multiplied by $K - l$ and the prior mass of $\mathbf{s}$ to contain all the $T_{\mathbf{s}}$ state sequences that have identical baseform labeling. In the ΔBIC between two state-sequences $\mathbf{s}_1$ and $\mathbf{s}_2$ (of K_1 and $K_2 = K_1 + 1$ number of speakers, respectively) the following penalty term embodies the difference between their log-priors

$$C_2 = \log \frac{\pi_{\mathbf{s}}(\mathbf{s}_1|K_1, \tilde{b})}{\pi_{\mathbf{s}}(\mathbf{s}_2|K_2, \tilde{b})} - \log K_2. \quad (6)$$

4. Weights, simplex and the priors' strength

4.1. The issue with the weights

The final modification has to do with the fact that the weights of a Mth order GMM lie on a $(M-1)$-simplex (since $\sum_{m=1}^{M} w_m = 1$), hence they correspond to $M-1$ free parameters. Therefore, when calculating the pairwise ΔBIC values by integrating out φ given two rival state sequences that differ on a single merging of two states into one, the alternative $\mathcal{H}_1$ (i.e. two speakers) is modeled with one less parameter than the parameters of the model under the null hypothesis (i.e. one speaker). A possible solution would be to add to ΔBIC the term $T_{\tilde{n}} = \frac{1}{2}\log \tilde{n}$, i.e. to penalize the null hypothesis for having one more parameter.

4.2. Global, local and segmental settings

In the expression of $T_{\tilde{n}}$, $\tilde{n}$ can either be $\tilde{n} = n$ for the global-BIC setting, or $\tilde{n} = n_k + n_l$ if the local setting is assumed, where n_k, n_l are the number of observations under the kth and lth clusters. Whichever of the two settings is assumed, this term is of very low magnitude compared to the logarithmic Generalized Likelihood Ratio (GLR). If however the segmental approach is examined, where we implicitly fix the strength of the parameter priors of clusters rather that the summed strength, a much stronger penalty term would be raised. Due to its good performance in the single-gaussian formulation of the BIC, we derive some formulae for it, keeping the same algorithmic strategy.

By adopting the segmental approach, we penalize φ_k only by its effective sample size, namely $n_k = \sum_{i=1}^{n} \delta(s^{(i)}, k)$. Therefore, the penalty term for the BIC becomes $P_s = \sum_{k=1}^{K} \frac{\mathcal{P}_k}{2} \log n_k$, where $\mathcal{P}_k = M_k[d(1+\lambda)+1] - 1$ if diagonal covariance matrices are used. Therefore the penalty term for the ΔBIC can be obtained to be

$$C_3 = \frac{\mathcal{P}_k}{2}\log n_k + \frac{\mathcal{P}_l}{2}\log n_l - \frac{\mathcal{P}_{k\cup l}}{2}\log(n_k + n_l) \quad (7)$$

where $\mathcal{P}_{k\cup l} = (M_k + M_l)[d(1+\lambda)+1] - 1$ number of the GMM-level free parameters of the single-speaker model regarding the two states. Also, λ is the usual tuning parameter and as in [9] (theoretically $\lambda = 1$), and we choose to apply it only to the set of parameters that encode mean values, due to reasons related to the implied priors and discussed in [9]. Since the strength of the prior $\pi(\varphi)$ increases with the number of speakers given n and the total number of gaussians, is natural to anticipate that the proposed BIC version penalizes to model's complexity more softly that the Global-BIC setting. Therefore, is a systematic way to bias the results towards more pure clusters, with a risk of overfitting the data.

Finally, we also try to evaluate a stricter penalty term that involves a square root of the sample sizes that has been proposed in [9], and showed improved performance for the single gaussian BIC setting. The corresponding penalty term remains as in (7) but with $\mathcal{P}_k^{'} = M_k[d(1+\lambda\sqrt{n_k})+1] - 1$ and $\mathcal{P}_{k\cup l}^{'} = (M_k + M_l)[d(1+\lambda\sqrt{n_k+n_l})+1] - 1$.

5. Experiments

5.1. Set-up and algorithm

In this section we test the penalties proposed in the previous sections for the task of speaker diarization. As test data we use meeting room recordings from the NIST RT evaluation campaigns [1], in particular the recordings corresponding to the latest two evaluations to date (RT07s and RT09s). From the different acoustic conditions considered in the NIST evaluations, we evaluated our algorithms using the multiple distant microphones (MDM), which we first combined into a single channel for each recording via delay&sum beamforming [10] using the BeamformIt toolkit[2]. Next, we applied a speech activity detection algorithm as described in [11] and finally performed agglomerative bottom-up clustering using a system closely based on [2]. Results computed without adding any penalty terms are taken as baseline for performance comparisons.

In order to compare the results with the ground-truth we use the diarization error rate (DER), which is the standard metric used in the NIST RT evaluations. The DER measures the percentage of time that the system hypothesizes the wrong label (including the different speakers and silence regions). When evaluating the results we include the overlap regions, penalizing for regions where more than one speaker is speaking but only one (or the wrong multiple speakers) have been hypothesized. As the diarization system used for the experiments does not do anything to deal with overlap, this only has the effect of raising the overall DER errors, but we think it is a more realistic and comparable result.

5.2. Experimental results

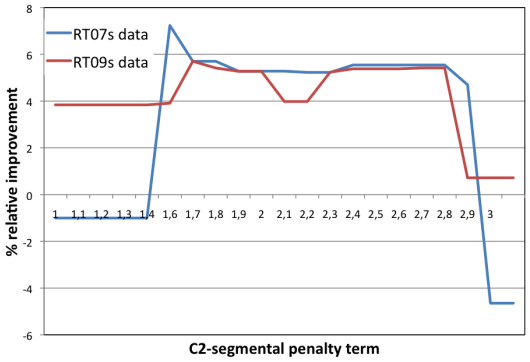
Among the three penalty terms proposed in the previous sections, we proposed two alternatives for the C_3 term: one derived from segmental-BIC, which we will call C_3-segm., and one which adds the additional square root term to the previous one which we will call C_3-sqrt. Each of these penalties accepts a penalty adjusting parameter λ that determines how much should the term affect the final output. Figure 1 shows the percentage improvement over the baseline by using all proposed penalty terms with either C_3-segmental or C_3-sqrt, computed for several values of λ. In 1a the C_3-segmental obtains relative improvements around 5% both for RT07 ans RT09 for λ values ranging from 1.6 to 2.8. The sudden jumps observed in the curves are due to the nature of the DER metric, averaged over only 8 meetings for RT07 and 7 meetings for RT09. Although each meeting contains from 20 to 30 minutes of evaluated speech, small changes in the algorithm can cause errors in the estimation of the final number of speakers, which results in big DER jumps for a particular meeting and therefore noticeable jumps in the average DER. Similarly, 1b shows results for the C_3-sqrt algorithm. In this case RT09 data obtains important improvements of around 10% relative from λ values between 0.15 and 0.225 while RT07 does not have a clear maximum point in the range

[1] http://www.itl.nist.gov/iad/mig/tests/rt

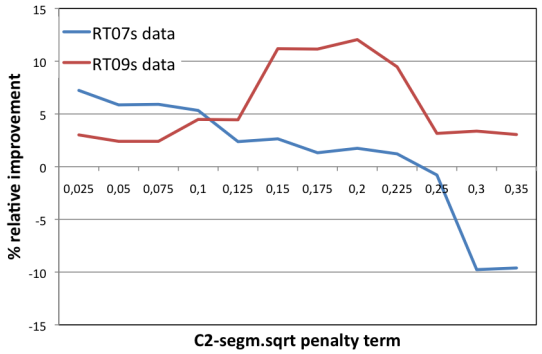
[2] http://www.xavieranguera.com/beamformit

Table 1: *Results.*

System eval	RT07s		RT09s		Average	
	DER	Variation	DER	Variation	DER	Variation
Baseline	18.94	—	27.88	—	23.38	—
C_1 alone	18.85	+0.50%	27.88	0%	23.33	+0.19%
C_2 alone	18.85	0%	27.88	0.50%	23.33	+0.19%
C_3-sqrt alone ($\lambda = 0.15$)	18.45	+2.28%	24.76	**+11.19%**	21.58	+7.68%
C_3-segm. alone ($\lambda = 2$)	17.92	+5.38%	26.77	+3.98%	22.32	+4.55%
C_1+C_2+C_3-sqrt ($\lambda = 0.15$)	18.44	+2.63%	24.76	**+11.19%**	21.58	**+7.70%**
C_1+C_2+C_3-segm. ($\lambda = 2$)	17.94	+5.27%	26.41	+5.27%	22.15%	**+5.27%**



(a)



(b)

Figure 1: Relative percentage improvement on DER of the overall system *w.r.t.* the λ parameter using (a) C_3-segmental and (b) C_3-sqrt.

we analyzed. For the following results we selected $\lambda = 2$ for the C_3-segmental penalty and $\lambda = 0.15$ for the C_3-sqrt penalty. Fig. 1 shows the algorithm performance in terms of DER for the baseline system (no penalty terms applied) and for different penalty term configurations. The top section in the table shows the performances achieved by applying each of the proposed penalties independently. We can observe how C_1 and C_2 are almost not affecting the output when applied on their own. This is due to the fact that their values are usually one order of magnitude smaller than the obtained BIC value. On the other hand, C_3-sqrt and C_3-segm. obtain on their own important improvements, 11.9% for RT09s when using C_3-sqrt. On the bottom rows we show the performances achieved when all three penalties are used together. Note that although C_1 and C_2 did not achieve any significant improvement on their own, when applying all penalties together these help improve the final results compared to the individual penalties alone.

6. Conclusion

In this report, we derived some possible improvement to the equal-parameter BIC, [2]. Each of the three terms is based on a different aspect of the approximation made by the BIC, namely the invariance of the posterior under permutation, the prior of the state-sequence, and the fact that the total number of free parameters are only only roughly equal. The above modifications alone showed some minor improvement over the baseline formulation, that especially when combined with an alternative BIC-setting (the segmental-BIC) are capable of enhancing the performance with respect to two NIST datasets of meeting data.

7. References

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

[2] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, november 2003, pp. 411–416.

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

[4] M. J. Beal, "Variational algorithms for approximate bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[5] M. Stephens, "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society, Series B*, vol. 62, pp. 795–809, 2000.

[6] 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.

[7] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *Comput. J.*, vol. 41, pp. 578–588, 1998.

[8] N. Chopin and F. Pelgrin, "Bayesian inference and state number determination for Hidden Markov Models: an application to the information content of the yield curve about inflation," *Journal of Econometrics*, vol. 123, no. 2, pp. 327–344, December 2004.

[9] T. Stafylakis, G. Tzimiropoulos, V. Katsouros, and G. Carayannis, "A new penalty term for the BIC with respect to speaker diarization," in *Proceedings of ICASSP*, 2010, pp. 4978–4981.

[10] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE TASLP*, vol. 15, no. 7, pp. 2011–2021, September 2007.

[11] C. Fredouille and N. Evans, "The influence of speech activity detection and overlap on the speaker diarizationfor meeting room recordings," in *Proc. Interspeech'07*, September 2007.