VARIATIONAL BAYESIAN SPEAKER DIARIZATION OF MEETING RECORDINGS

Fabio Valente, Petr Motlicek and Deepu Vijayasenan

fabio.valente@idiap.ch, petr.motlicek@idiap.ch, dvijaya@idiap.ch
IDIAP Research Institute, Martigny CH-1920, Switzerland.

ABSTRACT

This paper investigates the use of the Variational Bayesian (VB) framework for speaker diarization of meetings data extending previous related works on Broadcast News audio. VB learning aims at maximizing a bound, known as Free Energy, on the model marginal likelihood and allows joint model learning and model selection according to the same objective function. While the BIC is valid only in the asymptotic limit, the Free Energy is always a valid bound. The paper proposes the use of Free Energy as objective function in speaker diarization. It can be used to select *dynamically* without any supervision or tuning, elements that typically affect the diarization performance i.e. the inferred number of speakers, the size of the GMM and the initialization. The proposed approach is compared with a conventional state-of-the-art system on the RT06 evaluation data for meeting recordings diarization and shows an improvement of 8.4% relative in terms of speaker error.

Index Terms— Variational Bayesian Methods, Speaker Diarization, Meetings Data

1. INTRODUCTION

Most of the current diarization systems are based on agglomerative hierarchical clustering of audio segments (for a review see [1]). The system measures the similarity between clusters and iteratively merges the closest pairs of segments until a stopping criterion is met. A common criterion choice is based on the Bayesian Information Criterion (BIC) [2] which has been applied to speaker clustering problems for the first time in [3]. The BIC is obtained as an approximation of the marginal log-likelihood $log\, p(Y|m)$ of the data Y given a model m and is valid only in the large data limit. The estimation of $\log\, p(Y|m)$ is untractable for complex models that contain hidden variables. In those cases, the BIC is used because of its simplicity. Large number of studies have been devoted to making the BIC effective in case of limited amounts of data (see for instance [4], [5]).

This work investigates a type of approximation referred as Variational Bayesian (VB) methods. VB methods [6],[7],[8] aims at directly maximizing a bound F_m on the marginal log-likelihood $\log\, p(Y|m)$. This bound is also known as Variational Free Energy F_m and allows *joint model learning and model selection* using the same objective function. While the BIC is obtained in the large data limit, the bound is always valid. Speaker diarization involves simultaneous clustering and model selection; this work shows that the Variational Free Energy F_m can be used as objective function for achieving both goals. The parameters that typically affect the diarization error like the number of speakers, the number of Gaussian components per speaker and the initialization can be obtained maximizing the Free Energy. The procedure is *unsupervised* i.e. does not need tuning on a separate data set. The paper extends previous related works on Variational Bayesian clustering of speakers carried in the framework of Broadcast News (BN) audio [9], [10] where the Free Energy has been used to select only the actual number of speakers. BN data are clean recordings and the clustering is typically evaluated according to speaker/cluster purity measures. On the other hand, this work focus on conversational meetings recorded with far-field microphones. The results are scored in terms of Diarization Error Rate (DER).

The remainder of the paper is organized as follows: section 2 describes the basic of Bayesian model selection and the BIC, section 3 describes the Variational Bayesian framework. Section 4 introduce the diarization system based on the VB framework and section 5 describes experiments and comparisons on the RT06 evaluation data.

2. BAYESIAN MODEL SELECTION

This section describes the theoretical Bayesian model selection and the most common approximation to it, the Bayesian Information Criterion (BIC). Let us consider a data set Y, a set of statistical models $M = \{m_j\}$ and a probability estimate $p(Y|m_j)$ for each model m_j in M. The model $\bar{m}$ that better "explains" data Y is the model that maximizes the posterior probability $p(m_j|Y)$ given by:

$$\bar{m} = argmax_{m_j}\, p(m_j|Y) = argmax_{m_j}\, p(Y|m_j)p(m_j)/p(Y) \quad (1)$$

where $p(m_j)$ is the prior probability of the model m_j and $p(Y)$ is the probability of the data (independent from the model). If $p(m_j)$ is uniform, the model that maximizes $p(m_j|Y)$ is the model that maximizes $p(Y|m_j)$. Let us denote with Θ_j the set of parameters associated with model m_j. It is possible to write:

$$p(Y|m_j) = \int p(Y, \Theta_j|m_j)d\Theta_j = \int p(Y|\Theta_j, m_j)p(\Theta_j|m_j)d\Theta_j. \quad (2)$$

where $p(\Theta_j|m_j)$ is the prior distribution of the parameters Θ_j. Expression (2) is referred as *marginal likelihood* and is a key quantity in Bayesian model selection.

The relation between the marginal likelihood and the model complexity has been established in [11] where it is shown that the integral (2) embeds a term (referred as Occam's factor) that penalizes more complex models with respect to simpler models. The conventional Bayesian Information Criterion (BIC) [2] is obtained approximating the marginal likelihood in the large data limit :

$$lim_{N\to\infty}\, \log\left(p(Y|m_j)\right) = \log\left(p(Y|\Theta_{ML}, m_j)\right) - \frac{d}{2}\log\, N = BIC(m_j) \quad (3)$$

where d is the number of free parameters,Θ_{ML} is the Maximum Likelihood estimation of the model parameters and N is the available amount of data i.e. the cardinality of $|Y|$. In case of limited amounts of data, the BIC criterion is not effective. A simple solution is tuning the penalty factor $\frac{d}{2}\log\, N$ with an heuristic constant as proposed in many speaker clustering systems (see e.g. [4]).

The BIC is used because of its simplicity and because the exact form of the marginal likelihood (2) is not available for complex models. In fact, if the model contains hidden variables X, the maximization of $\int p(Y, X, \Theta_j | m_j) dX d\Theta$ can became easily untractable given the dimension of the joint space (X, Θ). *Variational Bayesian* methods aim at bounding the integral using a simpler (approximated) posterior distribution for (X, Θ). Those approximations are described in the next section.

3. VARIATIONAL BAYESIAN METHODS

Let us consider a model m with a parameter set Θ and an hidden variable set X. The marginal log-likelihood can be written as:

$$\log\, p(Y|m) = \log \int\, p(Y, X, \Theta|m)\, d\Theta\, dX \tag{4}$$

Variational Bayesian approximations (e.g. [7],[8]) assume that the unknown and untractable distribution $p(\Theta, X|Y, m)$ can be approximated with another (simpler) distribution $q(\Theta, X)$ referred as Variational Bayesian posterior distribution. An upper bound on the marginal log-likelihood (4) can be obtained multiplying and dividing the argument of the integral by $q(\Theta, X)$ and applying the Jensen inequality as follows:

$$\begin{aligned} \log\, p(Y|m) &= \log \int \frac{q(\Theta, X) p(Y, \Theta, X|m)}{q(\Theta, X)} d\Theta\, dX \\ &\geq \int q(\Theta, X) \log \frac{p(Y, \Theta, X|m)}{q(\Theta, X)} d\Theta\, dX \end{aligned} \tag{5}$$

The key point of this approximation is the definition of the distribution $q(\Theta, X)$ simple enough to allow tractability but close to $p(\Theta, X|m)$ to obtain a reasonable approximation. In [7], the use of the mean-field approximation is proposed i.e. $q(\Theta, X) = q(\Theta)q(X)$ i.e. the approximated distributions over parameters and hidden variables are considered independent. Thus bound, (5) can be rewritten as:

$$\begin{aligned} F_m(q(X), q(\Theta)) &= \int q(X) q(\Theta) \log p(Y, X|\Theta, m) dX d\Theta \\ &- \int q(X) \log q(X) dX - \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|m)} d\Theta \end{aligned} \tag{6}$$

Expression (6) is generally referred as Variational Free Energy $F_m(.)$ and is composed of three terms:

Term 1: $\int q(X) q(\Theta) \log p(Y|X, \Theta, m) dX d\Theta$ is the expected log-likelihood computed w.r.t. the approximated posterior distributions $q(X)$ and $q(\Theta)$.

Term 2: $-\int q(X) \log q(X)$ is the entropy of the approximated posterior distribution over hidden variables $q(X)$.

Term 3: $\int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|m)} = KL(q(\Theta)||p(\Theta|m)) \geq 0$ is the Kullback-Leibner divergence between the Variational posterior parameter distributions $q(\Theta)$ and the prior parameter distributions $p(\Theta|m)$. Being the KL divergence positive, this term acts as penalty (i.e. the Occam factor), reducing the value of the Free Energy for models that have more parameters. Contrarily to the BIC, where the model complexity is proportional to the number of free parameters, *Term 3* explicitly considers the divergence between posterior and prior distributions. The BIC is only an asymptotic approximation while the Variational bound is *always* valid.

The Variational posterior distributions $q(\Theta)$ and $q(X)$ that maximize the Free Energy $F_m(q(\Theta), q(X))$ can be obtained using an inference method similar to the conventional Expectation-Maximization (EM). A local maximum of the Free Energy can be obtained iterating across the following set of equations :

$$\textit{E-like}\text{-step:}\quad q(X) = \frac{1}{Z_X} e^{<\log\, p(Y, X|\Theta)>_{q(\Theta)}} \tag{7}$$

$$\textit{M-like}\text{-step:}\quad q(\Theta) = \frac{1}{Z_\Theta} e^{<\log\, p(Y, X|\Theta)>_{q(X)}} p(\Theta|m) \tag{8}$$

where Z_X and Z_Θ are normalization constant and $< a >_b$ designates the expected value of a w.r.t. b. Equations (7-8) are often referred as Variational Bayesian Expectation Maximization (VBEM, see [8]). The Free Energy F_m can be used to perform model selection as it is related to the complexity of the model (for details see [7]).

Coming back to the initial question of section 2, the model selection given by $\bar{m} = argmax_{m_j}\, p(m_j|Y)$ needs the estimation of the untractable quantity $p(Y|m_j) = \int p(Y, \Theta, X|m_j)$. The decision can be replaced by $\bar{m} = argmax_{m_j}\, F_{m_j}(q(\Theta), q(X))$ where F_{m_j} is tractable and can be estimated using VBEM. F_{m_j} can be used at the same time as objective function for clustering and for model selection criterion.

4. VARIATIONAL BAYESIAN SPEAKER DIARIZATION

This section describes the application of the VB framework to diarization of an audio recording with an associated acoustic feature stream O (i.e. MFCC features). Let us divide the stream into segments of equal length D^1 i.e. $O = \{O_t\}$ with $t = 1, ..., T$. We use $D = 300$ms. O_t is composed of D consecutive frames i.e. $O_t = \{O_{t1}, ...O_{tD}\}$.

Most of the current systems are based on an agglomerative hierarchical clustering of speech segments and the Bayesian Information Criterion (BIC). We investigate the use of Variational Bayesian framework for learning a set of clustering models $\{m_j\}$ and selecting the most probable model $\bar{m}$ that explains the data. The learning and the selection is done according to the same objective function, the Variational Bayesian Free Energy F_m. Let us define the probabilistic model $p(Y|X, \theta)$ and the prior distributions $p(\theta|m)$. We assume that speech segments $\{O_t\}$ are independent and can be generated from one of the S available speakers. Each speaker is modeled using a Gaussian Mixture Model. The likelihood can be written as:

$$\begin{aligned} p(O|\Theta) &= \prod_{t=1}^{T} p(O_t|\Theta) \quad p(O_t|\Theta) = \sum_{j=1}^{S} \alpha_j p(O_t|\Theta_j) \qquad (9) \\ p(O_t|\Theta_j) &= \prod_{p=1}^{D} p(O_{tp}|\Theta_j) \;\; p(O_{tp}|\Theta_j) = \sum_{i=1}^{N} \beta_{ij} \mathcal{N}(O_t|\mu_{ij}, \Gamma_{ij}) \end{aligned}$$

where S is the number of speakers, N is the number of Gaussian components per speaker, α_j is the probability of speaker j and Θ_j are parameters for the model of the jth speaker, $\Theta_j = \{\beta_{ij}, \mu_{ij}, \Gamma_{ij}\}$ are respectively weights, means and covariance matrix of the ith Gaussian component of the jth speaker. $\mathcal{N}()$ denotes the Gaussian distribution.

The complete parameter set for model (9) is given by $\Theta = \{\alpha_j, \beta_{ij}, \mu_{ij}, \Gamma_{ij}\}$. Prior distributions are chosen in the conjugate family over parameters Θ i.e.:

$$\begin{gathered} p(\Theta) = p(\{\alpha_j\}) \prod_{j=1}^{S} p(\{\beta_{ij}\}) \prod_{i=1}^{N} p(\mu_{ij}|\Gamma_{ij}) p(\Gamma_{ij}) \\ p(\{\alpha_j\}) = \mathbf{Dir}(\lambda_{\alpha_0}) \quad p(\{\beta_{ij}\}) = \mathbf{Dir}(\lambda_{\beta_0}) \\ p(\mu_{ij}|\Gamma_{ij}) = N(\rho_0, \xi_0 \Gamma_{ij}) \;\; p(\Gamma_{ij}) = W(\nu_0, \Phi_0) \end{gathered} \tag{10}$$

where $Dir(.)$ is a Dirichlet distribution, $\mathcal{N}(.)$ is a Normal distribution, $W(.)$ is a Wishart distribution. $\Lambda_0 = \{\lambda_{\alpha_0}, \lambda_{\beta_0}, \rho_0, \xi_0, \nu_0, \Phi_0\}$

[1] Those segments can be obtained by uniform segmentation or by speaker change detection. We limit here the investigation to the uniform segmentation

are hyperparameters associated with the prior distributions and they are fixed as follows: $\rho_0 = \bar{O}$, $\Phi_0 = (\tau_0 + p)\,Cov(O)$ where $\bar{O}$ and $Cov(O)$ are mean and covariance matrix of the acoustic vectors as estimated on the entire recording; $\nu_0 = \tau_0 + p$, $\{\lambda_{\alpha_0} = \lambda_{\beta_0} = \xi_0\} = \tau_0$ where p is the dimension of the acoustic vector[2].

Posterior Variational distributions have the same parametric form of the priors with updated hyperparameters denoted with $\Lambda = \{\lambda_{\alpha_i}, \lambda_{\beta_{ij}}, \rho_{ij}, \xi_{ij}, \nu_{ij}, \Phi_{ij}\}$. Posterior hyperparameters are initialized equal to priors apart from $\{\rho_{ij}\}$ which are randomly initialized.

The VBEM algorithm can be applied in a straightforward way to the model (9) under prior distributions (10) and the Free Energy can be computed in close form (detailed formula are reported in [9],[10]). After convergence this will provide a clustering of segments O_t into speakers and an approximation of the model complexity. The diarization output is obtained mapping the segments O_t to the most probable speaker in the mixture model.

The clustering output, thus the diarization error, depends on: 1) the number of speakers S 2) the number of Gaussian components per speaker N 3) the value of the prior τ_0 4) the initialization $\{\rho_{ij}\}$. Let us denote this set of elements with $I = \{S, N, \tau_0, \{\rho_{ij}\}\}$.

We propose a method that searches in an exhaustive way the elements I and selects the diarization output that corresponds to the maximum Free Energy. It can be summarized in the following steps:

1 Extraction of the MFCC features from the audio stream; speech/non-speech detection and rejection of the non-speech frames.

2 Uniform segmentation of the speech into segments of length D. In the following we use $D = 300ms$ i.e. 300 speech frames.

3 For each $I^* = \{S^*, N^*, \tau_0^*, \{\rho_{ij}^*\}\}$ in $I = \{S, N, \tau_0, \{\rho_{ij}\}\}$

    3-1 Initialization of the model (9) with $S = S^*$,$N = N^*$,$\tau_0 = \tau_0^*$,$\{\rho_{ij}\} = \{\rho_{ij}^*\}$.

    3-2 Perform Variational Bayesian Expectation Maximization (VBEM) method and estimate the Free Energy $F(I^*)$.

4 Selection of the clustering corresponding to the maximum Free Energy $F(I)$.

5 Estimation of the speaker assignment thus the diarization output.

In words, this method generates a large number of diarization outputs and selects the one which holds the highest Free Energy. The search in the space of elements $I = \{S, N, \tau_0, \{\rho_{ij}\}\}$ is unsupervised and based only on the objective function F_m

5. EXPERIMENTS

The data used for the experiment consists of meeting recordings obtained using an array of far-field microphones also referred as Multiple Distant Microphones (MDM). Studies are carried on the NIST RT06 evaluation data for Meeting Recognition Diarization task. A sum-and-delay beamforming is applied to the MDM signals using the *BeamformIt* toolkit [12][3]. Such pre-processing produces a single enhanced audio signal out of those recorded with the far-field microphones. 19 MFCC features are then extracted from the beam-formed signal.

Being interested in comparing the clustering algorithms, the same speech/non-speech segmentation will be used across all experiments only the speaker error is reported in the following.

5.1. Number of speakers/clusters

In this first experiment, we investigate the use of the Free Energy to select the actual number of speakers (clusters) in the audio file. The value of S (i.e. the number of speakers in the model) changes in between 1 and 10. Other parameters are arbitrary set to $N = 10$ (10 components per GMM), $\tau_0 = 1$ and ρ_{ij} are randomly initialized. The sensitivity to those parameters will be studied in section 5.2.

Table 1 reports speaker error obtained using the VB Free Energy selection. The table also reports oracle results obtained manually selecting the lowest and the highest speaker error as a function of the number of speakers and speaker error obtained randomly sampling a large number of times the generated solutions function of the number of speakers. In order to be effective, the selection method should be as close as possible to the best solution and never worst then the random selection.

File	Spkr Error (VB)	Best	Worst	Random
CMU_20050912-0900	**19.1**	**19.1**	38.6	29.7
CMU_20050914-0900	**10.8**	**9.2**	30.7	20.1
EDI_20050216-1051	27.6	20.7	57.6	46.6
EDI_20050218-0900	**26.5**	**26.5**	47.3	33.0
NIST_20051024-0930	21.8	10.0	33.3	20.1
NIST_20051102-1323	**6.7**	**6.7**	51.5	18.5
TNO_20041103-1130	**24.2**	**22.7**	47.1	34.0
VT_20050623-1400	20.3	14.2	42.7	24.0
VT_20051027-1400	**19.2**	**19.2**	43.9	34.0
ALL	16.20	12.7	43.40	28.50

Table 1. RT06 Speaker Error obtained using Free Energy, oracle best/worst selection and random selection.

In 7 of the 9 meetings, the selection is equal or close to the lowest possible speaker error. In 3 meetings, the proposed approach fails to select the lowest speaker error. However the VB selection is better then random selection in 9 meeetings.

In order to illustrate how the model selection operates, figure 1 (left) plots Term1+Term2 and Term3 (Penalty) for a particular meeting (VT_200510271400) which contains 4 speakers. Both Term1+Term2 and Term3 increase as the number of speakers (thus of parameters) increase. Figure 1 (center) plots the Free Energy (Term1+Term2-Term3) which shows a maximum for $S = 4$ speakers. The same figure also plots the speaker error which shows a minimum in correspondence of four speakers. On the other hand, figure 1 (right) plots the quantity Term1+Term2-$\lambda\, d\, \log \frac{N}{2}$ which corresponds to replacing the penalty term (Term 3) with the conventional BIC penalty. The theoretical value $\lambda = 1$ estimates a number of speakers equal to 8. The correct number of speakers is inferred tuning $\lambda = 2$.

5.2. Optimization of other parameters

This section investigates the joint optimization of all parameters $I = \{S, N, \tau_0, \{\rho_{ij}\}\}$. The three following experiments are run and the Free Energy is used to select the best system:

1- Simultaneous selection of the best prior τ_0 and speakers S. The value of S (i.e. the number of speakers in the model) ranges in $\{1, ..., 10\}$ as before and the value of τ_0 ranges in $\{1, 10, 100, 1E3\}$. The number of Gaussian components is arbitrarily fixed to $N = 10$ and ρ_{ij} are randomly initialized. Results are reported in table 2 first line.

2-Simultaneous selection of the number of Gaussian components N, τ_0 and S. The value of S (i.e. the number of speakers in the model) ranges in $\{1, ..., 10\}$ and the value of τ_0 ranges in $\{1, 10, 100, 1E3\}$, the number of gaussian component ranges in

[2] From the definition of the Wishart distribution ν_0 must be larger or equal to the dimension of the vector p.

[3] The bug-corrected version of the Beamforming 2.0 is used for experiments. This provides improved results w.r.t. previous versions.

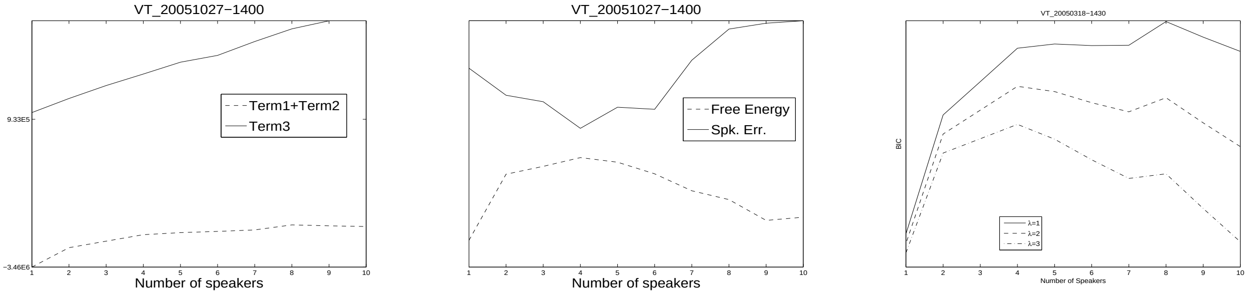

Fig. 1. Left figure plots Term1+Term2 and Term3 as a function of the number of speakers for meeting VT_20051027-1400. Center figure plots the Free Energy (Term1+Term2-Term3) and the speaker error for meeting VT_20051027-1400; the maximum of the Free Energy corresponds to the minimum speaker error in correspondence of 4 speakers. Right picture plots Term1+Term2$-\lambda\, d\, \frac{N}{2}$: the actual number of speaker is obtained tuning the BIC penalty with $\lambda = 2$.

$N = \{5, 10, 15, 20\}$. ρ_{ij} are random initialized. Results are reported in table 2 second line.

3-Simultaneous selection of the initialization ρ_{ij} and N, τ_0 and S. The value of S (i.e. the number of speakers in the model) ranges in $\{1, ..., 10\}$ and the value of τ_0 ranges in $\{1, 10, 100, 1E3\}$, the number of Gaussian component ranges in $N = \{5, 10, 15, 20\}$. The values ρ_{ij} are set according to 10 different random initializations. Results are reported in table 2 third line.

The joint optimization of all the elements in I reduces the speaker error from 16.2% to 12.7%. The Free Energy based model selection outperforms the random model selection in all cases.

Parameter	Spkr Error (VB)	Best	Worst	Random
τ_0	16.0	12.0	43.4	25.5
N	14.7	11.7	44.4	25.4
ρ_{ij}	12.7	10.7	44.4	24.9

Table 2. Speaker Error for optimization of parameters τ_0,N,ρ_{ij} obtained using VB selection, oracle best/worst and random selection.

The VB system is compared with a state-of-the-art diarization system based on a modified BIC criterion [13] that performs hierarchical agglomerative clustering. This system performs iteratively clustering and realignment until the stopping criterion is met [5]. The baseline is tuned on a development data set and is initialized with 16 speakers modeled by a 5 component Gaussian Mixture Model. The baseline achieves a speaker error equal to 13.6% as compared to the 12.7% obtained by the VB system thus the proposed approach outperform the baseline by 8.4% relative speaker error.

6. CONCLUSION

This paper investigates the use of the Variational Free Energy as objective function for speaker diarization. The conventional BIC criterion is valid only in large data limit and typically needs tuning to be effective. The Free Energy is an always valid, tractable bound on the marginal log-likelihood of the model.

Experiments on the RT06 data reveal that F_m is an effective criterion for selecting in between the number of actual speakers in the audio file, the number of Gaussian components per speaker, the initialization and the prior τ_0. Comparison with a state-of-the-art system based on a modified BIC and agglomerative clustering reveals a speaker error reduction of 8.4% relative.

We emphasize that the proposed approach operates in completely unsupervised fashion and all the related parameters are obtained as maximization of an objective function. Those findings extends previous related work on Broadcast news data [9],[10] where Free Energy was used only to select the actual number of speakers and shows the effectiveness also on meetings data. In future work we plan to experiment with more informative priors like, the one proposed in [14].

7. ACKNOLEDGMENTS

This work was supported by the European Union under the integrated projects AMIDA, Augmented Multi-party Interaction with Distance Access, contract number IST-033812, by the Swiss National Center of Competence in Research on Interactive Multimodal Information Management IM2 and by the Indo-swiss project KERSEQ. The author would like to thank colleagues at IDIAP and ICSI for their help with the meeting setup.

8. REFERENCES

[1] Tranter S.E. and Reynolds D.A., "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14(5), 2006.

[2] Schwartz G., "Estimation of the dimension of a model," *Annals of Statistics, 6, 1978.*

[3] Chen S. and Gopalakrishnan P., "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proceedings of the DARPA Workshop, 1998.*

[4] Tritschler A. and Gopinath R., "Improved speaker segmentation and segments clustering using the bayesian information criterion," *Proceedings of Eurospeech 99.*

[5] Ajmera J. and C. Wooters, "A robust speaker clustering algorithm," *IEEE Automatic Speech Recognition Understanding Workshop, 2003, pp. 411-416.*

[6] MacKay D.J.C., "Developments in probabilistic modelling with neural networks – ensemble learning," *Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands, 1995.*

[7] Attias H., "A variational bayesian framework for graphical models," *Advances in Neural Information Processing Systems, 12 ,pp. 209–215, 2000.*

[8] Ghahramani Z. and Beal M.J., "Propagation algorithms for variational bayesian learning," *Advances in Neural Information Processing Systems, 13 , MIT Press, 2001.*

[9] Valente F. and Wellekens C.J., "Variational bayesian speaker clustering," *Odyssey'2004, The speaker and language recognition workshop, 2004, Toledo, Spain.*

[10] Valente F., "Variational bayesian methods for audio indexing.," *PhD Thesis, Universite' de Nice-Sophia Antipolis, 2005.*

[11] MacKay D. J. C., "Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks," *Network:Comput. Neural Syst. 6, 469–505, 1995.*

[12] http://www.xavieranguera.com/beamformit/, ," .

[13] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *Signal Processing Letters, IEEE*, vol. 11, no. 8, pp. 649–651, 2004.

[14] Reynolds D., Kenny P., and Castaldo F., "A study of new approaches to Speaker Diarization," *Proceedings of Interspeech 2009.*