# Variational Bayesian GMM for Speech Recognition

*Fabio Valente, Christian Wellekens*

Institut Eurecom
Sophia Antipolis, France
{fabio.valente,christian.wellekens}@eurecom.fr

## Abstract

In this paper, we explore the potentialities of Variational Bayesian (VB) learning for speech recognition problems. VB methods deal in a more rigorous way with model selection and are a generalization of MAP learning. VB training for Gaussian Mixture Models is less affected than EM-ML training by overfitting and singular solutions. We compare two types of Variational Bayesian Gaussian Mixture Models (VBGMM) with classical EM-ML GMM in a phoneme recognition task on the TIMIT database. VB learning performs better than EM-ML learning and is less affected by the initial model guess.

## 1. Introduction

This paper aims at investigating the application of variational Bayesian methods to speech recognition dealing with problems related to initial priors choice with speech data. Variational Bayesian (VB) learning is a relatively new learning technique that allows the processing of Bayesian models that cannot be trained using full Bayesian approach because of their complexity. Many classical models like Gaussian mixture models or hidden Markov models can be trained using this new approach. Recently application to speech data shows many advantages (see [6],[5],[4]). VB techniques offers a new framework for doing parameter estimation and model selection in a more rigorous way. Like the MAP, they consider parameter posterior probability distributions but unlike MAP, this is not a *point* estimation, but the whole model probability is evaluated. VB estimation provides information about the model quality while training the model itself i.e. the objective function is a measure itself of how good the model is. Another interesting property of VB learning is the self-pruning i.e. model during the training does not use extra degrees of freedom that get lost. That behavior can be seen as an advantage or not. On the one hand, uncertainty about the model is not taken into account but on the other hand, model selection is done during the learning itself because parameters that are not used disappear. In this paper we investigate the application of VB learning to Gaussian mixture models and test it with some speech data. We compare two different approach to VBGMM with classical maximum likelihood EM algorithm for GMM (that can be seen as a special case of VB learning). Classical GMM-ML algorithms often suffer from overfitting when the component number is not adequate with the data amount. VBGMM are not that affected by initial model choice because the model prunes extra degrees of freedom and because integrating priors leads to a kind of regularization.

## 2. Variational Bayesian learning

Given a set of observed variables $Y$ and some parameters $\theta$, Bayesian learning aims at optimizing the so called *marginal likelihood* $p(Y)$, where parameters $\theta$ have been integrated out. From Bayes rule we have: $p(Y) = p(Y,\theta)/p(\theta|Y)$ and considering the log of both members it is possible to write: $log\, p(Y) = log\, p(Y,\theta) - log\, p(\theta|Y)$. Instead of integrating parameters $\theta$ w.r.t. their true unknown pdf, an approximation called variational posterior, and denoted as $q(\theta|Y)$, is used. Taking expectation w.r.t $q(\theta|Y)$, we obtain:

$$log\, p(Y) = \int q(\theta|Y) log\, p(Y,\theta) d\theta - \int q(\theta|Y) log\, p(\theta|Y) d\theta$$
$$= \int q(\theta|Y) log\, [p(Y,\theta)/q(\theta|Y)] d\theta + D(q(\theta|Y))||p(\theta|Y)) \quad (1)$$

where $D(q(\theta|Y))||p(\theta|V)$ represents the Kullback-Leiber (KL) distance between the variational posteriors and the true posteriors. The term $\int q(\theta|Y) log\, [p(Y,\theta)/q(\theta|Y)] d\theta$ is often indicated as *negative free energy* $F(\theta)$. Because of the KL-distance property $D(a||b) \geq 0$ (with equality if $a = b$), $F(\theta)$ represent a lower bound on $log\, p(Y)$ i.e. $log\, p(Y) \geq F(\theta)$. Variational Bayesian learning aims at maximizing the lower bound $F(\theta)$ that can be rewritten as:

$$F(\theta) = \int q(\theta|Y) log\, p(Y|\theta) d\theta - D(q(\theta|Y)||p(\theta)) \quad (2)$$

The second term in eq. (2) represents the distance between the approximate posterior and the parameter prior and can be interpreted as a penalty term that penalizes more complex models. For this reason $F(\theta)$, can be used to determine the model that best fits to data in the same way the BIC criterion is used.

*Maximum a Posteriori* can be seen as special cases of VB learning. In fact, if $q(\theta|Y) = \delta(\theta - \theta^{'})$, finding the maximum of equation (2) means:

$$max_{Q(\theta)} F(\theta) = max_{\theta'} \int \delta(\theta - \theta^{'}) log[p(Y|\theta)p(\theta)] d\theta$$
$$= max_{\theta'} log[p(Y|\theta^{'})p(\theta^{'}))] \quad (3)$$

where the term $\int q(\theta) log\, q(\theta) d\theta$ has been dropped because it is constant. Expression (3) corresponds to the classical MAP criterion. It is important to notice that the VB approach carries information about the uncertainty on parameters $\theta$ while MAP does not. In fact in MAP parameter learning is done punctually ($max\, log\, [p(Y|\theta^{'})p(\theta^{'}))]$) while in VB, parameters are integrated out, even if they are integrated w.r.t. variational posterior ($max\, \int q(\theta|Y)\, log[p(Y|\theta)p(\theta)]d\theta$). Furthermore VB allows model comparison: free energy value gives information on the model quality, while MAP only gives best parameters for an imposed model. The price to pay is that the free energy is only a lower bound and not an exact value.

## 3. Variational Bayesian learning with hidden variables

Variational Bayesian learning can be extended to the incomplete data case. In many machine learning problems, algorithms must take care of hidden variables $X$ as well as of parameters $\theta$ (see [1]). In the hidden variables case, the variational posterior becomes $q(X, \theta|Y)$ and a further simplification is assumed considering it factorizable as $q(X, \theta|Y) = q(X|Y)q(\theta|Y)$. Then the free energy to maximize is:

$$F(\theta, X) = \int d\theta q(X)q(\theta)log[p(Y, X, \theta)/q(X)q(\theta)]$$
$$= < log \frac{p(Y, X|\theta)}{q(X)} >_{X,\theta} -D[q(\theta)||p(\theta)] \qquad (4)$$

where $< \; . \; >_z$ means average w.r.t. $z$. Note that $q$ is always understood to be conditioned on $Y$. It can be shown that when $N \rightarrow \infty$ the penalty term reduce to $(|\theta_0|/2)log\, N$ where $\theta_0$ is the number of parameters i.e. the free energy becomes the Bayesian Information Criterion (BIC). To find the optimum $q(\theta)$ and $q(X)$ an EM-like algorithm is proposed in [1] based on the following steps:

$$q(X) \propto e^{<log\, p(Y,X|\Theta)>_\theta} \qquad (5)$$
$$q(\theta) \propto e^{<log\, p(Y,X|\theta)>_X} p(\theta) \qquad (6)$$

Iteratively applying eq.(5) and eq.(6) it is possible to estimate variational posteriors for parameters and hidden variables. If $p(\theta)$ belongs to a conjugate family, posterior distribution $q(\theta)$ will have the same form as $p(\theta)$.

A different method for optimizing variational parameters can be found in [2]. Latent variables and model parameters are considered as unobserved stochastic variables and are treated in the same way. If we define $\omega$ to designate both $\theta$ and $X$ and assume the factorization $q(\omega) = \prod_i q(\omega_i)$, an iterative set of re-estimation formulae can be found minimizing the KL distance between the variational distribution $q(\omega_i)$ and prior distributions giving:

$$q(\omega_i) = \frac{exp < log\, p(Y, \omega) >_{k \neq i}}{\int exp < log\, p(Y, \omega) >_{k \neq i} d\omega_i}. \qquad (7)$$

An interesting property of VB learning is that extra degrees of freedom are not used i.e. the model prunes itself. There are two possible philosophies about the correctness of the model self pruning: on the one hand it is not satisfactory because prediction will not take into account uncertainty that models with extra parameters can provide (see [7]), on the other hand it can be used to find the optimal model while learning the model itself, initializing it with a lot of parameters and letting the model prune parameters that are not used.

## 4. Variational Bayesian GMM

GMM can be trained using VB learning. GMM makes the hypothesis that a given data set has pdf of the form $\sum_i \pi_i N(\mu_i, \Sigma_i)$, where $N(\mu_i, \Sigma_i)$ is Gaussian with mean $\mu_i$ and covariance matrix $\Sigma_i$. Many models for priors and latent variables have been proposed. In this paper, we will consider models described in [1] and [3].

### 4.1. Variational Bayesian GMM I

Let's consider the model proposed in [1]. Given set of $N$ observation vectors $Y = \{y_1, ..., y_N\}$ and latent variables $S = \{s_1, ..., s_n\}$ that denotes the hidden component that generated $y_n$ where $s_n \in [1, m]$ and m is the number of Gaussians. The GMM is:

$$p(y_n|\theta) = \sum_{s=1}^{m} p(y_n|s_n = s, \theta)p(s_n = s|\theta) \qquad (8)$$

where $p(y_n|s_n = s, \theta) = N(\mu_s, \Gamma_s)$ and $p(s_n = s|\theta) = \pi_s$. Let's now define conjugate priors on the parameters $\theta = \{\pi_s, \mu_s, \Gamma_s\}$: weight coefficients are jointly Dirichlet, $p(\{\pi_s\}) = D(\lambda_0)$, means conditioned on precisions are Normal, $p(\mu_s|\Gamma_s) = N(\rho^0, \beta^0\Gamma_s)$, and precisions are Wishart, $p(\Gamma_s) = W(\nu_0, \Phi_0)$ (notice that $W(a, B)$ is sometimes defined as $W(1/a, B^{-1})$). Parameter variational posteriors factorize into $q(\theta) = q(\{\pi_s\}) \prod_s q(\mu_s, \Gamma_s)$ under severe independence hypothesis. Posteriors for hidden variables $S$ factorize into $q(S) = \prod_n q(s_n)$. Using (5) (E-step) we can compute $q(s_n) = \gamma_s^n$. Using (6) (M-step) it is possible to compute new parameters and new posterior parameters. For GMM parameters $\{\pi_s, \mu_s, \Sigma_s\}$ we obtain:

$$\bar{\pi}_s = \frac{1}{N}\sum_{n=1}^{N} \gamma_s^n, \quad \bar{\mu}_s = \frac{1}{\bar{N}_s}\sum_{n=1}^{N} \gamma_s^n y_n, \quad \bar{\Sigma}_s = \frac{1}{\bar{N}_s}\sum_{n=1}^{N} \gamma_s^n C_s^n \qquad (9)$$

where $C_s^n = (y_n - \bar{\mu}_s)(y_n - \bar{\mu}_s)^T$ and $\bar{N}_s = N\bar{\pi}_s$. Parameter posteriors will have the same form as priors because they belong to a conjugate priors family: $q(\{\pi_s\}) = D(\{\lambda_s\})$, $q(\mu_s|\Gamma_s) = N(\rho^s, \beta^s\Gamma_s)$, $q(\Gamma_s) = W(\nu_s, \Phi_s)$ where:

$$\lambda_s = \bar{N}_s + \lambda^0, \qquad \rho_s = (\bar{N}_s\mu_s + \beta^0\rho^0)/(\bar{N}_s + \beta^0)$$
$$\beta_s = \bar{N}_s + \beta^0 \qquad \nu_s = \bar{N}_s + \nu^0$$
$$\Phi_s = \bar{N}_s\bar{\Sigma}_s + \bar{N}_s\beta^0(\bar{\mu}_s - \rho^0)(\bar{\mu}_s - \rho^0)^T/(\bar{N}_s + \beta^0) + \Phi^0 \qquad (10)$$

Prediction of unseen data can be made integrating out parameters in eq. (8); it is possible to show that the resulting distribution is a mixture of Student-t distribution (see [1]):

$$p(y|Y) = \int p(y|\theta)\, q(\theta|Y)d\theta = \sum_{s=1}^{m} \bar{\pi}_s t_{\omega_s}(y|\rho_s, \Lambda_s) \quad (11)$$

where component s has $\omega_s = \nu_s + 1 - d$ d.o.f., mean $\rho_s$, covariance $\Lambda_s = ((\beta_s + 1)/\beta_s\omega_s)\Phi_s$ and weight coefficient $\bar{\pi}_s = \lambda_s/\sum_{s'} \lambda_{s'}$. Eq. (11) reduces to a GMM when $N \rightarrow \infty$. A close form for the free energy $F$ can be obtained and used for choosing the model that best fits to data. A serious numerical problem with classical ML-EM algorithm occurs when very few vectors are assigned to a Gaussian component so that its covariance matrix becomes almost singular. In VBGMM I formulation, this problem does not exists thanks to initial priors $\Phi_0$: when only a very few vectors will be assigned to a Gaussian components, its coefficient $\pi_s$ will converge to zero, reducing the number of Gaussians and avoiding any singularity.

### 4.2. Variational Bayesian GMM II

Let's now consider the model proposed in [3]. The latent variables $s_{in}$ where i=1,..,M is defined in order to have $s_{in} = 1$ if i=j where j is the component that generates $y_n$ and 0 in all other cases. It follows that it is possible to write the likelihood of a data set $Y$ as:

$$P(Y|\mu, \Sigma, s) = \prod_{n=1}^{N}\prod_{i=1}^{M} N(y_n|\mu_i, \Sigma_i)^{s_{in}} \qquad (12)$$

Given weight coefficients $\pi_i$, latent variables will have the following discrete distribution: $P(s|\pi) = \prod_{i=1}^{M} \prod_{n=1}^{N} \pi_i^{s_{in}}$. Contrarily to VB GMM I priors on weight coefficients will not be defined; the learning procedure consists in iteratively optimizing the lower bound on $p(Y, \theta|\pi)$ w.r.t $\pi$ and $\theta$. Let's define conjugate priors on means and covariance matrices:

$$P(\mu) = \prod_{i=1}^{M} N(\mu_i|0, \beta I), \quad P(\Sigma) = \prod_{i=1}^{M} W(\Sigma_i|\nu, V) \tag{13}$$

Defining $\theta = \{\mu, \Sigma, s\}$, assuming $q(\theta) = q(\mu, \Sigma, s) = q_\mu(\mu) q_\Sigma(\Sigma), q(s)$ and applying (7) to $p(Y, \theta|\pi)$, variational posterior distributions are (for details see [3]):

$$q_s(s) = \prod_{n=1}^{N} \prod_{i=1}^{M} p_{in}^{s_{in}} \quad q_\mu(\mu) = \prod_{i=1}^{M} N(\mu_i|m_\mu^{(i)}, T_\mu^{(i)})$$
$$q_\Sigma(\Sigma) = \prod_{i=1}^{M} W(\Sigma_i|\nu_\Sigma^{(i)}, V_\Sigma^{(i)}) \tag{14}$$

where $p_{in}, m_\mu^{(i)}, T_\mu^{(i)}, \nu_\Sigma^{(i)}, V_\Sigma^{(i)}$ are posterior distribution parameters. Once the lower bound on $P(Y, \theta|\pi)$ has been maximized w.r.t $q(\theta)$, weight coefficients can be updated. Setting derivatives of negative free energy to zero we obtain:

$$\pi_i = \frac{1}{N} \sum_{n=1}^{N} s_{in} \tag{15}$$

Even in this case, free energy can be computed in closed form. In [3] the previously discussed property of model self pruning was used successfully to recover the correct model dimension in synthetic data problems. An initial guess model with a high number of Gaussian is used, and during the training degrees of freedom that are not used are dropped, converging at the end of the training to the optimal number of Gaussians.

### 4.3. Variational Bayesian learning and priors

In those approaches, both authors have used non-informative priors. When relatively small amount of synthetic data are used both techniques seem to be robust to initial prior choice. Anyway, as it was observed in [4] when large amount of speech data is used, a certain sensitivity to initial priors is observed, and optimal initial priors depends on the amount of data. In VBGMM, priors have a consistent influence in how the model prunes itself: the final number of Gaussians that survive to the training will depends on the values of $\Phi_0$ for VBGMM I and $V$ for VBGMM II (see [5]). Because speech recognition training uses important amount of data, estimation of optimal initial priors is an extremely computationally expensive tasks and for this reason they are often empirically chosen.

## 5. Experiments

To test the efficiency of VBGMM I and VBGMM II in determining the original data set dimension we carried out experiments on a synthetic data set constituted by 5000 vectors randomly generated following a 3 components GMM pdf. We tried to recover the correct data set dimension using the ML-BIC criterion and the VBGMM I free energy. Figure 1 shows results for BIC criterion, the maximum is achieved for $m = 3$ and then other models are progressively penalized. Figure 2 shows the negative free energy obtained with $\Phi_0 = \xi I$ with $\xi = 1$: the function has an important peak at m=3 that gives the right dimension without any ambiguity. Figure 3 shows the negative free energy for $\xi = 100$; paradoxally all models with $m \geq 3$ have the same variational free energy; looking at models with more than 3 Gaussians it is possible to notice that at the end of the learning only 3 components have survived while all other extra components have zero weight; in other words even if the model was initially a n-Gaussian model with $n > 3$, the final model has only 3 components i.e. the original data set dimension. On the same test set we run VBGMM II with a different number of initial models and with a value of $V = \xi I$ with $\xi$ moving from 10 to 100 and in all cases we recovered the correct number of Gaussians, their means and their covariance matrices.

To run experiments on speech data we used confusing phonemes set like in [5] from the TIMIT database. Acoustic vectors consist in 12 MFCC obtained with the HTK system. Confusing phonemes set consists of 6 stop consonants: p,t,k,b,d,g. Training data for each phoneme is limited to 10000 acoustic vectors for reducing computational charge. Our experiments aim at comparing the classical ML-EM algorithm for GMM with VBGMM I and VBGMM II previously described in a speech recognition task studying the influence that initial priors can have on final models. First we run recognition experiments on the 6 stop consonants using ML EM changing the initial number of Gaussians for the model from 1 to 10; results are given in table 1. Then we tested the VBGMM approach giving an initial model of 10 Gaussian components to observe how VB learning behaves. If VBGMM I is used, recognition can be done using (a) GMM parameters i.e. parameters (9) or using (b) inferred distribution i.e. equation (11), or even simply using (c) the part of the free energy relative to data $Y$ i.e. the first term of equation (4). We studied as well dependency on initial prior $\xi$ where $\Phi_0 = \xi I$ and the consequent dimension inferred by the VB learning. Results for VBGMM I are in table 2, with the following initial priors $\lambda_0 = 1$, $\beta^0 = 1$, $\rho^0 = \bar{y}$, $\nu_0 = 1$ where $\bar{y}$ is the average value of input data. The second part of table 2 shows the inferred number of Gaussian per phoneme after VB training.

| GMM | 1 | 2 | 6 | 10 |
|---|---|---|---|---|
| recognition rate | 51.0% | 55.2% | 62.4% | 62.6% |

Table 1: Recognition rate function of Gaussian components

| $\xi$ | 1 | 10 | 100 | 1000 | 2000 | 10000 |
|---|---|---|---|---|---|---|
| (a) | 64.2% | 63.4% | 64.0% | **66.3**% | 64.6% | 56.1% |
| (b) | 64.1% | 63.5% | 64.2% | 65.5% | 63.4% | 56.1% |
| (c) | 56.5% | 55.1% | 56.0% | 56.1% | 55.2% | 51.9% |
| /p | 10 | 10 | 10 | 10 | 8 | 3 |
| /t | 10 | 10 | 10 | 10 | 7 | 2 |
| /k | 10 | 10 | 10 | 10 | 10 | 4 |
| /b | 10 | 10 | 10 | 10 | 6 | 4 |
| /d | 9 | 9 | 9 | 9 | 7 | 2 |
| /g | 10 | 10 | 10 | 10 | 7 | 2 |

Table 2: Recognition rate and phoneme dimension function of priors. (a) GMM parameters, (b) Student t parameter (c) Free energy

Results show that recognition rate is sensitive to prior choice. VBGMM I always outperforms ML-EM GMM for 10 Gaussian as initial models. Only when very large initial priors are used, performance degrades; it means that for some

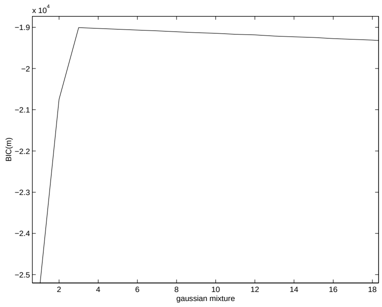

Figure 1: BIC(m) function of Gaussian mixtures

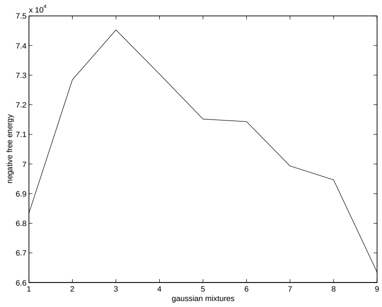

Figure 2: Negative free energy function of Gaussian mixtures for $\xi = 1$ (VBGMM I)

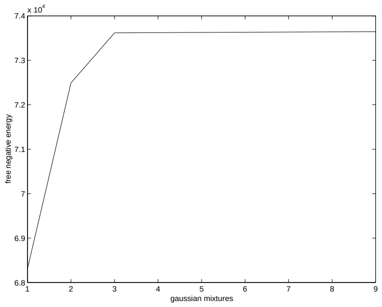

Figure 3: Negative free energy function of Gaussian mixtures for $\xi = 100$ (VBGMM I)

reasonable initial prior choice, VB learning outperforms traditional ML learning. For an appropriate prior value ($\xi = 1000$) VBGMM I with 10 Gaussians outperforms largely EM-ML model. Looking at the final phonemes dimension, it is possible to observe that when initial priors are big, the model prunes itself to a lower number of components, when initial priors are small all components survive to training. Even if all components survive to VB learning, priors lead a sort of regularization on it, avoiding this way overfitting problems. For inferring decision, methods (a) and (b) look almost equivalent; bound (c) seems not robust enough for doing recognition. Let's now consider analog experiments for VBGMM II with the following initial priors $\nu = 1, V = \xi I$. This time the recognition can be done using (a) the inferred parameters or (b) the first term of equation (4). Results are shown in table 3 as well as the inferred number of Gaussian per phoneme. VB GMM II again performs better than EM-ML on 10 Gaussian models. Method (b) is not as precise as method (a) for doing recognition. Again when too large initial priors are used, performances are affected. The final number of Gaussian per phoneme is lower than the one inferred with VBGMM I; this is probably due to the fact that there is no distribution imposed on weights. We can conclude that VBGMM II prunes models harder than VBGMM I.

| $\xi$ | 1 | 10 | 100 | 1000 | 2000 | 10000 |
|---|---|---|---|---|---|---|
| (a) | 64.7% | 65.4% | 64.0% | 64.7% | 64.9% | 57.19% |
| (b) | 63.6% | 63.8% | 63.1% | 62.9% | 62.9% | 56.1% |
| /p | 10 | 10 | 10 | 7 | 5 | 2 |
| /t | 9 | 9 | 8 | 7 | 6 | 2 |
| /k | 9 | 9 | 9 | 7 | 7 | 3 |
| /b | 10 | 10 | 9 | 7 | 4 | 2 |
| /d | 9 | 7 | 7 | 6 | 5 | 2 |
| /g | 7 | 6 | 6 | 5 | 5 | 2 |

Table 3: Recognition rate and phoneme dimension function of priors. (a) GMM parameters,(b) Free energy

Another important remark concerns the number of iterations needed by those methods to converge. EM-ML and VBGMM II converge almost with the same speed while VBGMM I seems to be always faster than the two other techniques.

## 6. Discussion and conclusions

Results shows that VB learning suffers less from overfitting problems than traditional EM ML learning. Experiments on speech data seem to confirm results previously achieved on synthetic data i.e. VB methods are able to deal with problems related to model selection better than classical techniques. Even if preliminary results are interesting many open questions are left. On the one hand those methods take advantage of regularization carried out by initial priors (that permits to avoid many problems e.g. singular solutions) and from informations that come from the explicit computation (even if in an approximated way) of integral w.r.t parameter distributions. On the other hand a certain sensitivity to initial prior choice seems to be an undesirable characteristic of those approaches; anyway there are reasonable ranges for initial priors in which recognition rates are still competitive. Estimation of optimal initial priors is a prohibitive task because of considerable amount of training data in speech recognition. Furthermore initial priors strongly affect the way in which the model prunes itself: for VBGMM I model pruning has the same results for $\xi$ in the range $[1, 1000]$ while VBGMM II makes hard pruning depending on initial prior value. Another open question is about the way self-pruning is done; dropping out degree of freedom avoid singularities and provides an efficient way for doing model selection and parameters learning at the same time but there are actually no guaranty (but experimental evidence) that the model pruning is done in the correct way.

## 7. References

[1] Attias, H., "A Variational Bayesian framework for graphical models", Advances in Neural Information Processing Systems 12, MIT Press,Cambridge, 2000.

[2] Bishop C.M., Winn J. "Structured variational distribution in VIBES",Proceedings Artificial Intelligence and Statistics, Key West, Florida, 2003 .

[3] Corduneau A., Bishop C.M. "Variational Bayesian model selection for mixture distributions",In T. Richardson and T. Jaakkola (Eds.), Proc. 8th Int. Conf. on Artificial Intelligence and Statistics, pp. 27-34. Morgan Kaufmann.

[4] Watanabe S. et al. "Application of the Variational Bayesian approach to speech recognition" NIPS'02. MIT Press.

[5] O.-W. Kwon, T.-W. Lee, K. Chan, "Application of variational Bayesian PCA for speech feature extraction," Proc. ICASSP 2002, Orlando, FL, pp. I-825–I-828, May 2002.

[6] Somervuo P., "Speech modeling using Variational Bayesian mixture of gaussians",Proc ICSLP 2002.

[7] MacKay D.J.C. "Local Minima, symmetry breaking and model pruning in variational free energy minimization"