VARIATIONAL BAYESIAN FEATURE SELECTION FOR GAUSSIAN MIXTURE MODELS

Fabio Valente, Christian Wellekens

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

ABSTRACT

In this paper we show that feature selection problem can be formulated as a model selection problem. A Bayesian framework for feature selection in unsupervised learning based on Gaussian Mixture Models is applied to speech recognition. In the original formulation (see [1]) a *Minimum Message Length* criterion is used for model selection; we propose a new model selection technique based on *Variational Bayesian Learning* that shows a higher robustness to amount of training data. Results on speech data from the TIMIT database show a high efficiency in determining feature saliency.

1. INTRODUCTION

Feature selection is a fundamental issue in any pattern recognition system. Speech recognition systems, of course, have same need for robust feature selection algorithms. Such a huge number of front-end techniques has been developed that determining the most relevant features out of the total number of possible features is a main point; in fact reducing the number of features reduces computational load and discarding irrelevant features can improve recognition. Furthermore when speech is affected by noise, efficient methods for determining reliable features are the basis for feature-based noise compensation techniques (e.g. missing data theory).

The proposed technique is a statistical model that permits to determine how effective a feature is in discriminating between models. This approach is not new in speech recognition. Previous study in this sense are for example [2] and [3]. In [2] a model based method is used to localize and select segments relevant to speaker recognition; selection is done by comparing likelihood in a given frame with competitors likelihood: if correct model is not dominant, the frame is rejected. In [3] acoustic backing-off is used as an alternative means of implementing missing feature estimation that uses soft decisions instead of hard ones. Acoustic backing-off assumes that a frame probability can be written as weighted sum of a model dependent probability and a model independent probability; the weight is called backing-off parameter and model independent probability is assumed to be uniform on all possible values.

The approach here proposed is somehow similar to the backing off approach, with the difference that a Hierarchical Mixture Model is used to determine in an optimal way the model dependent and independent probabilities as well as their weighting. In other words the model aims at finding the *saliency* of a certain feature, defined as its capacity to discriminate between model dependent and model independent distributions. The feature saliency framework reformulates the feature selection problem as a model selection problem: for this reason an efficient model selection criterion must be used.

The novelty proposed in this paper consists in the use of a new model selection criterion based on Variational Bayesian (VB) learning. Theorically VB model selection accommodate in the same expression the term that must be optimized and the model related penalty. Compared to other approaches like BIC, it does not require any adjusting constant. Furthermore VB learning converges sensibly faster than other learning framework.

The paper is organized as follows: in section 2 the feature saliency model is introduced, section 2.1 describes the the MML framework while section 2.2 describes VB framework. In section 3 we shows experiments on synthetic and speech data and finally we discuss results in section 4.

2. FEATURE SALIENCY IN GMM MODELS

We consider in this section the model proposed in [1] and [4]. A classical gaussian mixture model with diagonal covariance matrix can be written as: $p(y) = \sum_{j=1}^{K} \alpha_j p(y|\theta_j) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} p(y_l|\theta_{jl})$ where y is the observation vector, K is component number, D is feature number, α_j is weight of jth Gaussian component and θ_{jl} are parameters of jth Gaussian component for the lth feature. If each component represents a different cluster, the interest of the lth feature can be seen in its ability to discriminate between clusters. Let us define $u(y_l|\lambda_l)$, the probability of the lth feature regardless the cluster it belongs to. For features irrelevant to discriminate between clusters, we expect to have $u(y_l|\lambda_l) = p(y_l|\theta_{jl})$. To study the capacity of a given feature to discriminate between clusters, a coefficient ρ_l (referred as "feature saliency") is introduced for each feature. ρ_l can be considered as a mixture between distribution $p()$ and $u()$. The GMM model is modified as:

$$p(y) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} (\rho_l p(y_l|\theta_{jl}) + (1-\rho_l) u(y_l|\lambda_l)) \quad (1)$$

ρ_l is now a model parameter that quantifies how a given feature is relevant for a given cluster with respect to the total distribution. A completely irrelevant feature will result in $\rho_l = 0$ while a relevant feature will result in $\rho_l = 1$; ρ_l can be considered as a hidden variable and optimal value can be inferred using EM algorithm. Basically each component is represented by a GMM with two components in which a component is cluster dependent and the other cluster independent.Thus, feature selection problem becomes a model optimization problem in which the best number of Gaussians must be determined. If data labels are not available, EM algorithm can be used to learn each cluster parameters and feature saliency. In the original formulation learning is done using a EM/ML based algorithm while model selection is done using

an MML approach. We propose the use of Variational Bayesian (VB) learning for jointly learning optimal model and optimal parameters; the VB framework is someway more elegant in the sense that the same objective function can be used for performing model learning and parameter learning and seems to be more robust w.r.t the amount of training data. Furthermore VB learning seems to be able to converge in a smaller number of iterations compared to classical learning algorithms (i.e. ML/EM).

2.1. MML model selection

The MML model selection criterion was used in [1] to learn the optimal number of Gaussians and optimal feature saliency. Let us consider the classical Maximum Likelihood criterion: given an unlabeled training set $Y = (y_1, ..., y_N)$ with $y_i = (y_{i1}, ..., y_{iD})$, estimation of parameters given by $\Theta = \{\alpha_j, \theta_{jl}, \lambda_l, \rho_l\}$ can be expressed as $\hat{\Theta} = argmin\ -log\, p(Y|\Theta)$. If the number of Gaussian components is *a priori* known, a classical EM algorithm can be used to learn $\hat{\Theta}$, considering α_j and ρ_l as hidden variables. If the number of Gaussian components is unknown, a model selection criterion must be used, because GMM suffers from many drawbacks when the component number is inappropriate to the current data set. Using Minimum Message Length (MML) criterion, it is possible to derive (see [1]) a joint optimization criterion for parameters and model i.e.

$$\hat{\Theta} = argmin\{-log\, p(Y|\Theta) + \frac{1}{2}(K + D + KDR + DS)log\, n + \frac{R}{2}\sum_{j=1}^{K}\sum_{l=1}^{D} log(\alpha_j \rho_l) + \frac{S}{2}\sum_{l=1}^{D} log\,(1 - \rho_l)\} \quad (2)$$

where R and S are the number of parameters in θ_{jl} and λ_l, and K is the initial number of Gaussian components. In the proposed model $R = S = 2$. We will refer to this method as MML Feature Saliency (MML-FS). Criterion (2) can be seen as a MAP estimate if the following improper priors on α_j and ρ_l are defined:

$$p(\alpha_1, ..., \alpha_K) \propto \prod_{l=1}^{K} \alpha_j^{-RD/2} \quad (3)$$

$$p(\rho_l) \propto \rho_l^{-RK/2}(1 - \rho_l)^{-S/2} \quad (4)$$

Once priors are defined, posterior distributions can be obtained as follows ([1]):

$$\hat{\alpha_j} = \frac{max(\sum_i w_{ij} - RD/2, 0)}{\sum_j max(\sum_i w_{ij} - RD/2, 0)} \quad (5)$$

$$\hat{\rho_l} = \frac{max(\sum_{ij} u_{ijl} - KR/2, 0)}{max(\sum_{ij} u_{ijl} - KR/2, 0) + max(\sum_{ij} v_{ijl} - S/2, 0)} \quad (6)$$

An important advantage in using formulas (5) and (6) is the capacity of pruning parameters. In fact if the initial model is initialized with a huge number of Gaussians, MML learning should detect the correct number of clusters by pruning extra Gaussian components. Naturally when a component disappears from the model, the number of parameter is reduced and criterion (2) must be explicitly modified to take into account the current number of parameters used by the model. Anyway if the model is initialized with a very huge Gaussian component number, the term $max(\sum_i w_{ij} - RD/2, 0)$ may be zero for all components and all terms may be pruned out at the first iteration. To circumvent the problem a modified version of the EM algorithm that uses unnormalized accumulator has been proposed (CWEM Component Wise EM)(see [4]).

2.2. Variational Bayesian Learning

Recently a new Bayesian framework generally referred to as *Variational Bayesian* (VB) framework has been proposed. VB learning is an approximate learning method that allows simultaneous optimization of parameters and model. Given a training set Y, model parameters Θ and hidden variables X, VB assumes that true (and unknown) parameter posterior distributions $p(\Theta|Y)$ and $p(X|Y)$ can be approximated by *variational Bayesian* posterior distributions referred to as $q(\Theta|Y)$ and $q(X|Y)$. VB learning consists in finding $q(\Theta|Y)$ and $q(X|Y)$ that maximize the following upper bound on the marginal likelihood (see [5] for details):

$$F(\Theta, X) = \int q(\Theta|Y)q(X|Y)log\,\frac{p(X, Y|\Theta)}{q(X|Y)}d\Theta dX - D(q(\Theta|Y)||p(\Theta)) \quad (7)$$

where $p(\Theta)$ are parameter prior distributions and $D(.||.)$ indicates KL divergence. Expression (7) is often referred as *negative free energy*. The interest of this approach is that, even if expression (7) is an approximated bound, it contains a term that penalizes more complex models i.e. $D(q(\Theta|Y)||p(\Theta))$. It can be shown that when $N \to \infty$ where N is the number of training samples, it corresponds to the BIC criterion. In other words, the same quantity (7) used as parameters optimization criterion is a model selection criterion as well. To better formalize the model selection problem, let us introduce the model posterior probability $q(m)$. It can be demonstrated that the optimal value can be written as (see [5]):

$$q(m) \propto exp\{F(\Theta, X, m)\}\, p(m) \quad (8)$$

where $p(m)$ is the model priors. In absence of any prior information on model, $p(m)$ is uniform and optimal $q(m)$ will simply depend on the term $F(\Theta, X, m)$ i.e. free energy can be used as model selection criterion.

MAP learning can be seen as a special case of VB learning; in fact if $q(\theta|Y) = \delta(\theta - \theta')$, finding the maximum of equation (7) (where hidden variables are omitted for simplicity) means:

$$max_{Q(\Theta)}F(\Theta) = max_{\Theta'}\int \delta(\Theta - \Theta')log[p(Y|\Theta)p(\Theta)]d\Theta = max_{\Theta'}log[p(Y|\Theta')p(\Theta'))] \quad (9)$$

where the term $\int q(\theta)log\, q(\theta)d\theta$ has been dropped because it is constant. The second line in expression (9) corresponds actually to the MAP criterion; this is because MAP corresponds to a point estimation in the space of parameter distribution while VB integrates out parameters distributions. So if the $q()$ distribution is a delta, integrating parameters gives a point estimation.

VB learning is done using an EM-like algorithm (see [5]) based on the following two steps:

$$q(X) \propto e^{<log\, p(X,Y|\Theta)>_\Theta} \quad (10)$$

$$q(\Theta) \propto e^{<log\, p(X,Y|\Theta)>_X}\, p(\Theta) \quad (11)$$

where $< a >_b$ is the expected value of a w.r.t b. Many models such as GMM or HMM and others can be learned using this EM-like step. If parameter prior distributions are chosen in the conjugate family, variational posterior distributions will have the same form of priors; given $\theta_{jl} = \{\mu_{jl}, \sigma_{jl}\}$,$\lambda_l = \{\mu_l, \sigma_l\}$ let us define the following priors:

$$p(\alpha) = Dir(\lambda_0)\quad p(\rho) = Dir(\tau_0)$$
$$p(\sigma_{jl}) = \Gamma(b_0, c_0)\quad p(\mu_{jl}|\sigma_{jl}) = N(\mu|m_0, \beta_0\sigma_{jl})$$
$$p(\sigma_l) = \Gamma(b_0, c_0)\quad p(\mu_l|\sigma_l) = N(\mu|m_0, \beta_0\sigma_l) \quad (12)$$

where Dir is a Dirichlet distribution Γ is a Gamma distribution and N is a Normal distribution. Coming back to our feature selection model, it can be easily learned using VB learning, we will refer to it as VB-FS (Variational Bayesian Feature Saliency). Iteratively applying equations (10) and (11), optimal variational Bayesian posterior distributions can be learned . A closed form for the free energy can be derived and used for selecting the best model (for detailed reestimation formula see [6]). VB-FS, as well as MML-FS, can be initialized with a huge number of Gaussian components and the training will learn the best component number. This time, as the penalty term is contained in the criterion expression, no artificial adjustment of penalty term is needed because when a component disappears, its parameter distribution has the same expression of prior distribution and the term $D(q(\Theta|Y)||p(\Theta))$ will be zero for pruned components.

3. EXPERIMENTS

3.1. Synthetic data

To test GMM/feature saliency methods, we generated 1000 vectors of dimension 5 using a 3 component GMM with the following mean vectors: $m1 = [0, 0, 0, 0, 1]$,$m2 = [-1, 0, -1, -1, 1]$,$m3 = [1, 0, 1, -1, 1]$ and diagonal covariance matrixes. GMM weighs are respectively 0.3 0.4 and 0.3. We can notice that feature one and three can discriminate between three Gaussians, feature 4 cannot discriminate between $m2$ and $m3$, and features two and five cannot discriminate at all. In an ideal experiments we should have $\phi_1 = \phi_3 = 1$,$\phi_2 = \phi_5 = 0$ and $0 \leq \phi_4 \leq 1$. We run the MML-FS and VB-FS algorithms: both algorithms recovered correctly component numbers and features saliency. Anyway VB learning converges after a few iterations (about 30) while MML needs more than 400 iterations. Actually both techniques are affected by local minima problem. It is known by experimental evidence that VB converges faster than classical ML.

3.2. Speech data

In this section we study the application of MML-FS and VB-FS algorithms to speech data. Experiments are run on the TIMIT database. We tried to determine feature saliency for a feature set of dimension 75 constituted by: 12 MFCC+Δ+$\Delta\Delta$, 12 PLP+Δ+$\Delta\Delta$, Energy+Δ+$\Delta\Delta$. Another point we are interested in, is the quantity of available training data; in fact Bayesian approaches are generally more robust to lack of data. For this reason we run experiments with three training set of different size: 2k,20k,and 200k acoustic vectors. Furthermore MML-FS and VB-FS are tested with clean speech and with speech contaminated with noise; the noise is f16 cockpit noise from the NOISEX database. Both algorithms were initialized with 100 Gaussian components. Following values where used for parameters priors $\lambda_0 = \tau_0 = 1$, $b_0 = D$, $c_0 = 1$ and $\beta_0 = 1$, $m_0 = \bar{y}$. Table 1 shows in brackets the final dimension for clean and noisy speech (SNR=10db). MML learning is more sensitive to different amount of training data. On the other hand, VB learning seems to be more robust in determining feature saliency. This is probably due to the regularization introduced by parameter prior distributions. Furthermore when only sparse data are available MML learning prunes very hard all features i.e. feature saliency is zero almost for all features while again VB learning benefits from prior regularization resulting in less hard pruning. Figures 2 and 3 show feature saliency for the 200k observation set; feature saliency is actually different for the two techniques. MML learning gives more importance to first coefficients of MFCC,PLP and their deltas, but does not prune any feature as weak (i.e. very low saliency). A main point of the discussion is how the evaluate the quality of the clustering and the quality of features because learning was actually unsupervised. We will use recognition rate performed using strongest features and entropy measure.

Let us consider an entropy measure based on data entropy; let us define as in [1]:

$$w_{ij} \propto \hat{\alpha}_j \; p(y_i|\Theta_j) \;\; v_{ij} \propto \hat{\alpha}_j \; p(y_i|\Theta_U) \tag{13}$$

w_{ij} measures the probability of the observation y_i to belong to cluster j and not to the common data distribution whose probability is given by v_{ij}. In the case of only relevant features, it should be $w_{ij} = 1$; a simple way to measure how good the feature subset is, consists in entropy $H(w_{ij}) = 1/n \; \sum_i \sum_j w_{ij} \, log(w_{ij})$. Figure (1) shows entropy obtained progressively removing features sorted by saliency for VB and MML learned models. When weak features are removed, there is small entropy variation; on the contrary when strong features are removed, entropy increases fast.

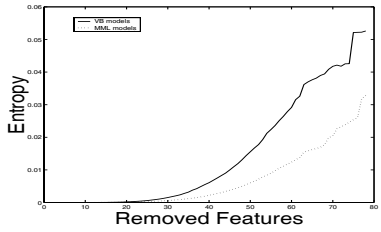

Fig. 1. Entropy obtained progressively removing features for VB (solid line) and MML (dashed line) learned models

Let us now consider phoneme recognition experiments. Table 1 shows context independent recognition for the classical 39 phoneme set obtained used the first 24 features with highest feature saliency in clean and noisy conditions. Phonemes are represented with a three state left to right HMM. Subsets obtained with VB learning performs better than subset obtained with MML learning. When 2k acoustic vectors are used, MML prunes almost all features, making recognition impossible. Anyway a point must be outlined: the more robust features do not constitute the optimal subset because the model proposed do not take into account information redundancy between features but only their robustness w.r.t. inferred clusters.

data amount (a)	2k	20K	200K
MML learning	N/A (2)	59.3% (68)	60.1% (100)
VB learning	57.2% (5)	60.6% (50)	61.8% (75)
data amount (b)	2k	20K	200K
MML learning	N/A (3)	50.8% (32)	51.2% (100)
VB learning	50.4% (7)	50.8% (70)	52.4% (75)

Table 1. Recognition rate for the first 24 features with highest saliency; in brackets final cluster dimension for (a) clean speech (b) noisy speech (SNR=10dB)

3.3. Bayesian decision

The previously defined framework allows to take global a decision on the quality of the feature. Let us consider an acoustic vector

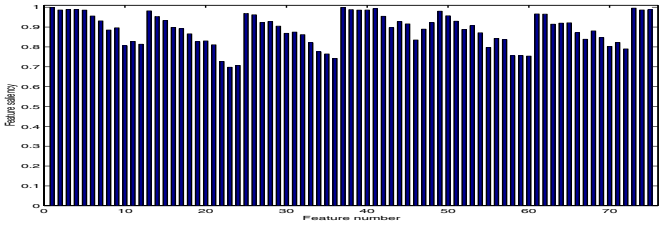

Fig. 2. Feature saliency with MML algorithm (initial components 100; inferred components 100) with 200k training vectors

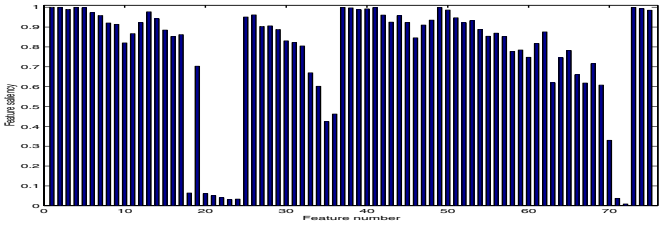

Fig. 3. Feature saliency with VB algorithm (initial components 100; inferred components 75) with 200k training vectors

$y = \{y_1, ..., y_L\}$; it is possible to take a Bayesian decision locally on the reliability of each feature in a frame. Let us designate with j the Gaussian component that holds the highest likelihood on the observation y (i.e. the cluster that best model the observation). Decision on feature l can be taken on the following criterion:

$$\rho_l \, p(y_l|\theta_{jl}) \lessgtr (1-\rho_l) \, u(y_l|\lambda_l) \quad for \;\; l = 1, ..., L \tag{14}$$

It is so possible to detect in a frame unreliable features; it is then possible to reconstruct features or simply discard them. Here we run recognition with the full 75 feature set and then discarding features that have high rejection rate (in our application features with more than 0.9 of rejection rate). Table 2 shows the PER for the full set and for the strongest features in the Bayesian sense; in brackets the final feature number. Using VB learned models, only 67 features hold a rejection rate lower than 0.9; discarding weak features does not affect recognition. MML learned models cannot prune any feature. In noisy conditions only 5 features are pruned for VB-FS while again MML-FS does not prune any feature. Eventually a decoder that take into account feature saliency frame by frame (here a global decision was used) can be used to further improve recognition.

	Full set	VB driven	MML driven
PER clean	66.2% (75)	66.8% (67)	66.2% (75)
PER (SNR=10db)	58.1% (75)	58.6% (70)	58.1% (75)

Table 2. PER with feature subset obtained with Bayesian decision; in brackets final features number.

4. CONCLUSION AND DISCUSSION

After running experiments on synthetic and speech data we can conclude that using VB learning has many advantages compared to MML learning. First of all, VB converges faster than MML, saving computational time. Then VB seems to determine features in a more robust way. It comes from the fact that the recognition rate coming from VB selected features is always higher than the recognition rate coming from MML selected features. This is probably due to the fact that the clustering is rather different for the two approaches. Furthermore VB learning benefits from regularization effects coming from prior distributions. This can be seen when very little training data are used: MML based method prunes almost all components, while VB based method achieves a "regularized" solution thanks to priors. Many considerations must be done about this approach to feature selection. An important advantage of this approach is that, contrarily to many feature selection algorithms, it does not need labels. In speech recognition applications, labels are not always available and furthermore the choice of the class used to discriminate is a critical point that can affect performance (phonemes,HMM states,single Gaussians). The absence of labels of course makes the model selection criterion a crucial point in algorithm performance.

On the other hand this approach suffers from many drawbacks. First of all, features are considered independent in the proposed model, so the information extracted is only the relevance of a certain feature w.r.t. the identified clusters, but redundancy between features is not considered. For this reason it is not possible to say that N strongest features represent N best feature subset because mutual information between features is not considered. Another important point is that single features are modeled using single Gaussians i.e. $p(y_l|\theta_{jl}) \;\; u(y_l|\lambda_l)$ are single Gaussians. It would be interesting to use a more complicated model like GMM to represent $p()$ and $u()$. A third drawback comes on the automatic clustering algorithm. It is actually difficult to obtain correct speech data clustering even with highly robust model selection algorithms. It has been shown that introducing duration constraints can increase the quality of clustering (see for example [7]). It may be interesting to incorporate duration constraints in the GMM based feature selection algorithm in order to obtain a better clustering. The current approach has been applied to feature selection but a very huge range of possible applications can be imagined. For example, exactly the same framework can be applied to filterbank output for determining unreliable information (e.g. because of noise) in missing feature approach. The same model can be used to model HMM emission probabilities with the same advantage of learning feature saliency and model parameters at the same time.

5. REFERENCES

[1] Figueiredo M. A. Law M. H., Jain A. K., "Feature selection in mixture based clustering," *NIPS*, 2002.

[2] Fredouille C. Besacier L., Bonastre J.F., "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, 2000.

[3] Boves L. de Veth J., Cranen B., "Acoustic backing-off as an implementation of missing feature theory," *Speech Communication*, vol. 34, 2001.

[4] Jain A. Law M., Figueiredo M., "Feature saliency in unsupervised learning," Tech. Rep., Michigan state university, 2002.

[5] Attias H., "A variational bayesian framework for graphical models," *Advances in Neural Information Processing Systems*, vol. 12, 2000.

[6] Valente F.; Wellekens C., "Variational bayesian feature selection," Tech. Rep. RR-03-087, Institut Eurecom, 2003.

[7] Ajmera J.; Bourlard H.; Lapidot I.; McCowan I., "Unknown multiple speaker clustering using hmm," *ICSLP*, 2002.