VARIATIONAL BAYESIAN FEATURE SALIENCY FOR AUDIO TYPE CLASSIFICATION

Fabio Valente, Christian Wellekens

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

ABSTRACT

In this paper, an approach based on Variational Bayesian Feature Saliency (VBFS) for robust audio type classification is proposed. VBFS aims at finding the most discriminative features in Gaussian Mixture Models based recognition systems. VBFS is applied to capture inter-type and intra-type feature saliency for different audio type (music, background noise, wide band speech, narrow band speech, etc.) in order to increase model generality that's always poor in non-speech models. We show that inferring saliency for different audio type improves classifications. Experiments are run on Broadcast news 1996-Hub4.

1. INTRODUCTION

An essential part of many speech applications consists in a preliminary phase of audio type classifications in order to separate different data type. A classical example is for instance broadcast news transcription, in which there is a huge variability of audio type: channels can be wide band or narrow band, and speech can be corrupted by music or generic background noise. Benefits of prior acoustic segmentation in applications like automatic transcription and speaker diarization have been proved, see for instance [1],[2],[3].

The idea of audio type classification is to split the audio file in homogeneous parts that can be more easily processed because of their similar acoustic properties (e.g. when adapting models) (see [4]). Hopefully non-speech segments should be removed because they may lead to transcription errors but it is very important not to remove any speech segment that otherwise will be definitively lost.

The most common approach to speech/non-speech discrimination and generally to audio type classification uses gaussian mixture models trained on labeled data. This approach suffers from many drawbacks as outlined in [4]. First of all in broadcast news audio files, the amount of data available for training non-speech models is definitely smaller than the amount of speech. Second, the huge diversity of non-speech segments and the lack of data produce very poor models for background noises and music. In [4], the use of MLLR for adapting models is proposed in order to increase performance.

In this paper we propose the use of a feature saliency (FS) measure to increase model generality. Feature saliency is first introduced in [5] and can be considered a measure of the discriminant power of a given feature. It can be computed in a supervised or unsupervised fashion. FS formulates the feature selection task as a model selection task so a model selection criterion must be used. In the original framework a Minimum Message Length (MML) criterion is used for learning models (see [5]); in [6], we shown that Variational Bayesian (VB) learning can be applied in order to obtain model robust to lack of data. The idea we explore in this paper is that using feature saliency in different acoustic classes can increase their separability. In fact, features that are commonly used are generally designed for speech recognition and their behavior in modeling classes like music and noise can be very different; anyway what we can model is the average discriminative (the 'saliency') capacity of those features that we expect to be very different when input is speech or non-speech. We show that explicitly taking into account this new quantity improves the discrimination.

The paper is organized as follows: in section 2 we review the concept of feature saliency, in section 2.1 we review the variational bayesian framework, in section 3 we describe the audio type classification problem, in section 4 we describe the application of VBFS to the current classification framework and we describe experiments on Broadcast news 1996-Hub4 in section 5.

2. FEATURE SALIENCY IN GAUSSIAN MIXTURE MODELS

The model considered here was first proposed in [5]. A classical gaussian mixture model with diagonal covariance matrix can be written as:

$$p(y) = \sum_{j=1}^{K} \alpha_j p(y|\theta_j) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} p(y_l|\theta_{jl}) \quad (1)$$

where y is the observation vector, K is the component number, D is the feature number, α_j is weight of the $j-th$ Gaussian component and θ_{jl} are parameters of the $j-th$ Gaussian component for the $l-th$ feature. If each component represents a different cluster, the interest of the $l-th$ feature can be seen in its capacity to discriminate between clusters. Let us define $u(y_l|\lambda_l)$, the probability of the $l-th$ 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:

$$p(y) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} (\rho_l p(y_l|\theta_{jl}) + (1-\rho_l) u(y_l|\lambda_l)) \quad (2)$$

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 (see [5]). We would like a completely irrelevant feature to result in $\rho_l = 0$ while a relevant feature to result in $\rho_l = 1$; ρ_l can be estimated using hidden variable approach 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. Anyway if $p(y_l|\theta_{jl})$ is equal to $u(y_l|\lambda_l)$, there are many possible solution for ρ_l; in order to obtain the desired properties, a model selection criterion must be associated with the learning algorithm, in this way if $p() = u()$, the simpler model will be chosen i.e. $\rho_l = 0$. Furthermore, if K is unknown it must be estimated as well with a model selection criterion.

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 framework ([5]) the model selection criterion used is the Minimum Message Length criterion. An EM algorithm is derived to learn model parameters $\Theta = \{\alpha_j, \theta_{jl}, \lambda_l, \rho_l\}$. A problem with this approach is that the MML selection criterion is not robust w.r.t. lack of data. To overcome this problem, we proposed the use of Variational Bayesian learning for model training and selection. In [6], we compared VB and MML feature saliency for feature pruning in speech recognition, coming to conclusion that when poor training data are available VB outperforms MML. In next section we briefly describe the Variational Bayesian method.

2.1. Variational Bayesian Learning

Variational Bayesian learning is a very effective framework for doing model selection and parameter learning at the same time. It consists in approximating the bayesian integral (that's intractable) with tractable integral.

Let us consider a data set $Y = \{y_1, ..., y_n\}$ and a model m. Model selection can be done optimizing the so called marginal likelihood:

$$p(Y|m) = \int d\theta \, p(\theta|m)p(Y|\theta, m) \tag{3}$$

where $p(\theta|m)$ is parameter probability given the model and $p(Y|\theta, m)$ is data likelihood given model and parameters. Expression (3) can be computed in an exact way using numerical methods (e.g. Monte-Carlo methods) but when parameter space is huge, the task can be computationally prohibitive. Variational Bayesian learning consists in approximating (3) with a lower bound that makes inference possible using an *Expectation-Maximization*-like (EM) algorithm. Let us introduce an approximated parameter density (the variational posterior) $q(\theta|Y)$ and let us consider the log marginal-likelihood $log \int d\theta \, p(\theta|m)p(Y|\theta, m)$. Considering Jensen inequality it is possible to write:

$$\begin{aligned} log\, p(Y|m) &= log \int d\theta q(\theta|Y) \frac{p(\theta|m)p(Y, \theta|m)}{q(\theta|Y)} \\ &\geq \int d\theta q(\theta|Y) log \frac{p(Y, \theta|m)}{q(\theta|Y)} = F(\theta) \end{aligned} \tag{4}$$

$F(\theta)$ is called *Free Energy* and it is a strict lower bound on the log marginal-likelihood. Variational Bayesian learning aims at optimizing $F(\theta)$ w.r.t. variational posterior distribution $q(\Theta|Y)$. It is possible to rewrite expression (4) as:

$$F(\theta, Y) = \int d\theta \, q(\theta|Y) log\, p(Y|\theta, m) - D(q(\theta|Y)||p(\theta|m)) \tag{5}$$

In our model we have to take care of many hidden variables as well as parameters. A solution to the hidden variables case is proposed in [7]. Let us define X the hidden variable set, and let us introduce the joint variational posterior distribution $q(X, \theta|Y)$; if the mean field simplification is assumed i.e. $q(X, \theta|Y) = q(X|Y)q(\theta|Y)$, In this case the expression for the free energy becomes:

$$\begin{aligned} F(\theta, X, Y) &= \int d\theta \, dX q(X) q(\theta) log[p(Y, X, \theta|m)/q(X)q(\theta)] \\ &= < log \frac{p(Y, X|\theta, m)}{q(X)} >_{X,\theta} - D[q(\theta)||p(\theta|m)]. \end{aligned} \tag{6}$$

It is possible to find an EM-like algorithm to find variational posterior parameters distributions. As described in [7], the algorithm iteratively updates variational posterior distributions over parameters and over hidden variables following those rules:

$$q(X) \propto e^{<log\, p(Y,X|\Theta)>_\theta} \tag{7}$$

$$q(\theta) \propto e^{<log\, p(Y,X|\theta)>_X} p(\theta|m) \tag{8}$$

Because free energy approximates the bayesian integral, it is intuitively a measure of the model quality. It can be easily proved using Jensen inequality that if a variational posterior distribution over models $q(m)$ is defined, then it is proportional to free energy.

$$q(m) \propto exp\{F(\theta, X, m)\} \, p(m). \tag{9}$$

In order to learn parameters of model (2), parameters prior distributions must be set. The choice of distributions belonging to conjugate prior family comes very useful because posterior distributions will have the same form as prior distributions. We set the following prior distributions:

$$\begin{aligned} p(\alpha) = Dir(\lambda_0) \quad p(\rho) = Dir(\tau_0) \\ p(\sigma_{jl}) = \Gamma(b_0, c_0) \quad p(\mu_{jl}|\sigma_{jl}) = N(\mu|m_0, \beta_0\sigma_{jl}) \\ p(\sigma_l) = \Gamma(b_0, c_0) \quad p(\mu_l|\sigma_l) = N(\mu|m_0, \beta_0\sigma_l) \end{aligned} \tag{10}$$

where $\lambda_0, \tau_0, b_0, c_0, m_0, \beta_0$ are hyperparameters, Dir,N and W designate respectively a Dirichelt, Normal and Wishart distribution. Details about reestimation formulas and a close form for the free energy (6) can be found in [6].

The main issue on using Variational Bayesian Learning for this problem is the lack of data for some classes that are generally modeled like background noise. The bayesian framework makes the estimation more robust w.r.t. spare training data, providing better models and pruning extra freedom degrees.

3. AUDIO TYPE CLASSIFICATION SYSTEM

In order to test VBFS, an audio type classification system designed for Broadcast News (BN) audio type is used. Experiments are run on the Hub-4 BN96. The aim here is to classify each frame in a given category. In a classical BN application the first discrimination consists in speech/non-speech separation. Once non-speech parts are discarded, the discrimination consists generally in wide-band/narrow-band speech.

Generally the recognizer is based on gaussian mixture models.

Speech in broadcast news data is labeled as belonging to six classes (tagged as F0-FX), noise is labeled as belonging to three classes: music, background speech and other background. Building a GMM for each labeled classes becomes unpractical because lack of data in some classes. It has been shown (see for instance [2]) that discrimination improves if the following models are considered: a model for speech corrupted with music (the F3 labeled speech), a model for narrow-band speech, a model for the other speech labels that are actually wide-band speech under different

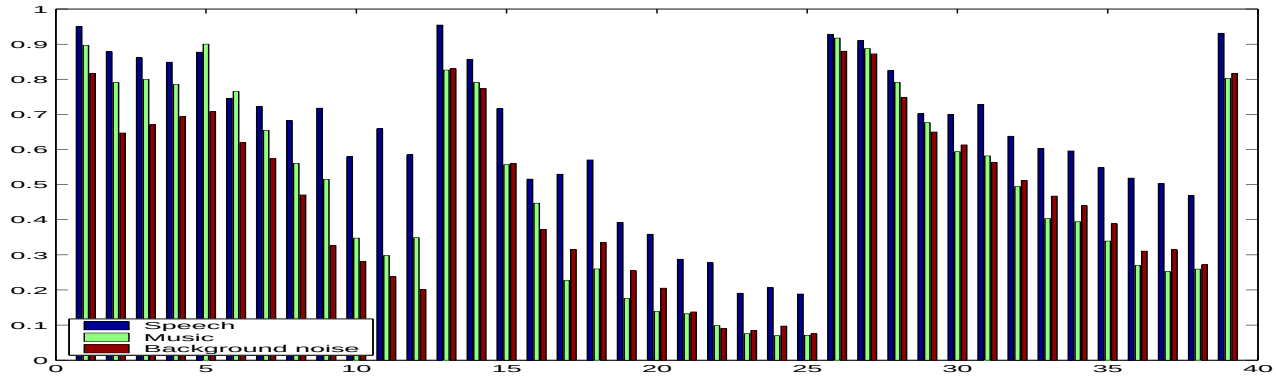

Fig. 1. Feature saliency for Speech and non-speech; feature order is 12 MFCC+energy+Δ+$\Delta\Delta$

	Background	Music	Speech+music	Speech wide	Speech narrow
Components	134	143	37	301	124
Pair KL	1E+07	1.6E+07	1E+08	5.9E+05	5.3E+06

Table 1. Final components number and pair-wise KL divergence for the considered audio type models

conditions, and finally a model for music and a model for other non-speech data (background speech and other background noises). Here arises the need for robust models because the huge variability in non-speech data makes difficult the model generalization.

Decoding is performed using a Viterbi based decoder with all models in parallel. To avoid spare solution, segments resulting smaller than 20 frames after the decoding are aggregated in order to obtain more compact segments.

4. VBFS AUDIO TYPE CLASSIFICATION

The previously described feature saliency framework can be applied to audio type classification in two possible ways in order to capture the intra-type (i.e. the saliency for a given audio type) and the inter-type feature saliency (i.e. the saliency for all audio types). Let us consider separately the two cases.

When the modeling aims at determining the intra-type saliency, model (2) is simply provided with an audio type data and an unspervised EM-like algorithm described in [8] is used to learn parameters and saliency. In this way a feature saliency specific to a given audio type is determined i.e. the capacity of given features to separate the data in different clusters. We expect a high feature saliency for speech data (because features were specifically designed to process speech) and low feature saliency for non-speech (because features are not designed to model different noise type or music type). The recognizer will replace the classical GMM with a model like (2) for each audio type. Intuitively the VBFS should be more general in the sense that, even when the noise (or the music) in the test will be very different from the one in the training set, we expect to retain the information about the discriminant power of each feature when the non-speech occurs. Formally speaking, when the algorithm aims at modeling intra-type feature saliency, it looks for $\Theta_m = \{\alpha_{mj}, \theta_{mjl}, \lambda_{ml}, \rho_{ml}\}$ so that $\Theta_m = argmax\{F(\Theta_m, Y_m)\}$ where Θ_m are parameters of model m and Y_m are data labeled as belonging to model m.

When the modeling aims at determining inter-type saliency, a single model (2) for the entire audio file is learned i.e. we find a global saliency for the feature set. Anyway labels with different audio type are provided resulting in a semi-supervised learning. In this case the saliency simply becomes a sort of feature weighting in order to determine the most discriminant features for separation in between models. From this side this approach should be more efficient in the sense that it directly aims at the discriminative task, but on the other may suffer because of poor generalization properties. Formally speaking, when the algorithm aims at modeling intra-type feature saliency, it looks for $\Theta_m = \{\alpha_{mj}, \theta_{mjl}, \lambda_{ml}\}$ and $\Gamma = \{\rho_l\}$ so that $(\{\Theta_m\}, \Gamma) = argmax\{F(\{\Theta_m\}, \Gamma, \{Y_m\})\}$; now a single set of feature saliency for all the training set is found.

4.1. Practicalities

Feature saliency parameters can be found using an EM like algorithm; details about update formula can be found in [8]. Anyway an implementation issue related with this model is that hidden variable number increases when FS is considered. A simple GMM with N components must handle N hidden variables, a GMM with feature saliency must handle approximatively $N + N \times D$ hidden variables, where D is the feature number. This results in a very slow convergence of the batch EM algorithm. In order to increase the convergence speed of the algorithm, learning is implemented in an incremental fashion. The choice for the incremental algorithm follows from considerations in [9].

5. EXPERIMENTS

In this section we describe experiments we run on Broadcast news 1996-Hub4. As proposed in [2], we used the development set for training session and evaluation set for decoding session. As outlined in [4], training data for music and background noise are definitely poor and offers poor generalization. For this purpose we compare the classical GMM based classifier with the previously proposed inter-type and intra-type feature saliency based GMM.

All GMM were initialized with 512 components. Feature set consists in 12 MFCC+energy and their delta and delta-delta coef-

	GMM		inter-type FS		intra-type FS	
	Speech	Non-Speech	Speech	Non-Speech	Speech	Non-Speech
Background		67.6%		76.7%		77.2%
Music		71.3%		76.9%		78.7%
Speech + Music	83.8%		87.2%		89.2%	
Speech Wide	97.7%		98.6%		98.7%	
Speech Narrow	95.3%		95.5%		96.5%	

Table 2. Speech/Non-Speech discrimination using classical GMM, intra-type FS and inter-type FS

ficients.

Figure (1) shows intra-type feature saliency for speech, music and background noise. It is interesting to notice that saliency of speech is always bigger than saliency of non-speech as we expected because MFCC features are actually designed to model speech. For MFCC and derivatives of first orders saliency of music is higher than saliency of noise; when delta-delta coefficients are considered the difference is less relevant.

A first effect of the Variational Bayesian learning is that extra degrees of freedom are pruned out i.e. the final number of gaussian components is generally smaller than the initial one and depends on data quantity and on current data distribution. This is a direct consequence of the fact VB makes model selection and parameter learning at the same time. First line of table (1) shows the final gaussian components for the five models we consider: obviously there is a strong relation between number of components and data available to train the model.

Another interesting point is quantifying how different VBFS distributions are from classical GMM distributions. A natural measure would be the KL divergence. KL divergence cannot be computed in close form when distributions are mixtures. For this reason, a pairwise KL divergence that constiutes an upper bound to real KL divergence is used (see [10]): if $a = \sum_i \alpha_i q_i$ and $b = \sum_i \beta_i p_i$ the pairwise KL divergence (PKL) is defined as:

$$PKL(a||b) = \sum_i \sum_j \alpha_i KL(p_i||q_j) + \alpha_i log \frac{\alpha_i}{\beta_j} \quad (11)$$

The PKL is symmetrized resulting in the Symmetric PKL divergence:

$$SPKL(a||b) = PKL(a||b) + PKL(b||a) \quad (12)$$

Second line of table (1) shows the SPKL divergence between the traditional GMM distribution and the feature saliency distribution for the same audio type. Divergences for wide and narrow speech are smaller compared to other audio type i.e. the saliency distribution is near to the traditional distribution when saliency is high.

Following the idea that the most important parameter is the loss of speech (see [4]), because once speech is discarded, it cannot be recovered, table (3) reports values of loss and missed non-speech for classical GMM system and for the two inter-type and intra-type feature saliency systems. The most performing technique seems to be the intra-type feature saliency.

Table (2) reports the classification error rate for each of the five models (Music, Background, Speech+music, Speech wide-band and Speech narrow-band) using again the three different techniques. Wide band speech and narrow band speech do not have any important improvements using the features saliency framework; this is easy to explain looking at figure (1). In fact for speech, feature saliency is very high, that means that there is no significant difference with the classical GMM model without any saliency (in other words if $\rho_l \rightarrow 1$ there is no more $u()$ and the model reduces to a traditional GMM). On the other hand in the case of music or background noise the saliency is definitely lower that means a bigger difference from the classical GMM model. Looking at hte error rate, the gain is considerable for background noise, music and speech+music that actually corresponds to those models where the saliency is less significant and models differs from the traditional GMM models.

	GMM	inter-type FS	intra-type FS
Speech lost	2.6 %	2.2%	1.9%
Missed non speech	1.9%	1.7%	1.6%

Table 3. Speech lost and missed non speech using the three different techniques

6. CONCLUSION

In this paper we investigated the use of Variational Bayesian feature saliency in an audio type classification framework in two different fashions (inter-type and intra-type). Results show that the two methods perform better than classical GMM on the Broadcast news 1996-Hub4 database. Gain is particularly interesting on those audio types that are difficult to model because of high variability and that are processed using standard speech recognition features.

7. REFERENCES

[1] Woodland P.C., "The development of the htk broadcast news transcription system:an overview," *Speech Communication*, vol. 37, 2002.

[2] Gauvin J.L., Lamel L., and Adda G., "The limsi broadcast news transcription system," *Speech Communication*, vol. 37, 2002.

[3] Meignier S., Moraru D., Fredouille C., Besacier L., and Bonastre J.F., "Benefi ts of prior acoustic segmentation for automatic speaker segmentation," *ICASSP*, 2004.

[4] Hain T. and Woodland P.C., "Segmentation and classifi cation of broadcast news audio," *ICSLP*, 1998.

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

[6] Valente F. and Wellekens C., "Variational bayeisan feature selection for gaussian mixture models," *ICASSP*, 2004.

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

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

[9] Neal R. and Hinton G., "A view of the em algorithm that justifi es incremental, sparse and other variants," *Learning in graphical models*, 1999.

[10] Do Minh N., "Fast approximation of kullback-leibler distance for dependence trees and hidden markov models," *IEEE Signal Processing Letters*, vol. 10, April 2003.