USING ONLINE MODEL COMPARISON IN THE VARIATIONAL BAYES FRAMEWORK FOR ONLINE UNSUPERVISED VOICE ACTIVITY DETECTION

David Cournapeau[1], *Shinji Watanabe*[2], *Atsushi Nakamura*[2], *Tatsuya Kawahara*[1]

[1] School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2] NTT Communication Science Laboratories, 2-4, Hikaridai, Seika Cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper presents the use of online Variational Bayes method for online Voice Activity Detection (VAD) in an unsupervised context. In conventional VAD, the final step often relies on state machines whose parameters are heuristically tuned. The goal of this study is to propose a solid statistical scheme for VAD using online model comparison which is provided from the Variational Bayes framework. In this scheme, two models are estimated online in parallel: one for the noise-only situation , and the other for the noise-plus-signal situation The VAD decision is done automatically depending on the selected model. An experimental evaluation on the CENSREC-1-C database shows a significant improvement by the proposed method compared to conventional statistical VAD methods.

Index Terms— Sequential Estimation, Robustness, Voice Activity Detection, Variational Bayes

1. INTRODUCTION

Voice Activity Detection (VAD) is the task of segmenting speech boundaries from audio signals, and is important for many speech applications, e.g. as a front-end for Automatic Speech Recognition (ASR) [1]. Especially, noise-robust VAD becomes important for ASR in noisy environments, since the number of insertion errors becomes large otherwise [2].

VAD methods are concerned with two sub-tasks. First, a set of features is selected to be robust to different kinds of noise. A classification method is then designed to segment speech sections from the feature vectors. Recently, several methods based on supervised training of classifiers, such as GMM, SVM and linked HMM, have been investigated. The best performance is obtained when the training data and test data have similar distribution. If there is a mismatch between training and unseen data, however, significant degradation is often observed. It is often the case that the system needs to operate in any environment without training the model. In this study, we focus on an approach of unsupervised, online classification, without requiring training data. Such classifiers often rely on a state machine with one or several thresholds updated from SNR estimation. As noted in [3], those state machines often rely on some heuristics for the noise floor estimation, and usually have several modes to adapt the classification to different SNRs. The goal of this study is to develop a statistical scheme for online classification, with a reliability measure for an alternative to manually tuned mode transitions.

For simplicity, we assume a scalar feature relevant for VAD, such as energy and High Order Statistics (HOS, [4]), is available. Each class (speech and non-speech) is assumed to follow a normal distribution whose parameters (mean and variance) change online. When both speech and noise are present in the signal, the classification problem is reduced to the online estimation of the parameters of a binary mixture model. However, the assumption of a binary mixture model is not correct in the case of noise-only sections, and the online estimated classifier cannot be used reliably. Therefore, we introduce an online model comparison scheme which can switch to unimodal model for those sections. Specifically, we use the Variational Bayes (VB) framework which provides an explicit approximation of the log-marginalized likelihood called the free energy, which can be used for comparing models [5]. Preliminary results using free energy in a mini-batch setting was already presented by the authors in [6], but model comparison and classification were provided by two separate methods – mini-batch free energy and online EM respectively. In this work, we present instead a method where both classifier parameters estimation and model comparison are undertaken by the same statistical model in a purely online fashion. Online extension of the Variational Bayes based on the stochastic approximation of the free energy [7] is used for online model comparison, to take into account possible changes in the acoustical environment. This method also provides the online parameter estimation of the mixture model, via posterior distributions of model parameters.

The organization of the paper is as follows. Section 2 introduces an online Expectation Maximization (EM) algorithm for unsupervised, online classification in the context of mixture models. Section 3 reviews the VB-EM framework for explicit computation of the free energy. Based on the stated equivalence between the VB-EM procedure and direct minimization of the parametrized free energy, we present the online extension of the VB-EM using a stochastic approximation of the parametrized free energy. Its application to the VAD task as well as an evaluation on CENSREC-1-C, a framework for noise robust VAD evaluation, is then presented in Section 4.

2. ONLINE EM FOR UNSUPERVISED CLASSIFICATION

When no training data is assumed, classification often relies on thresholding the feature, where the threshold is adapted online (e.g. from energy levels [3]). If we use a statistical framework instead, presence/absence of speech can be regarded as the realization of a binary random variable h, and the feature values as the realizations of a random variable (or vector for multi-dimensional features) x. The model for the observations is thus a simple binary mixture of Gaussian distributions, whose parameters can be estimated using the Expectation-Maximization (EM) algorithm [8] applied to latent models. As each iteration of the EM algorithm requires the whole dataset, the conventional EM cannot be used when online classification is needed. Online extensions have been proposed in the statistical literature to alleviate this problem. In this section, we will briefly review the principles of this online extension, as well as its limitations for VAD, which motivated the Bayesian extension presented in the later sections.

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

2.1. EM Algorithm

EM algorithm is a widely used method to estimate parameters in models where the Maximum Likelihood Estimation (MLE) would be hard to compute explicitly. Given N IID observations $x \triangleq x_1, \dots, x_N$, the log-likelihood L of a model parametrized by θ is defined as:

$$L(\theta) \triangleq \ln p(x;\theta) = \sum_{n=1}^{N} \ln p(x_n;\theta) \tag{1}$$

When the model contains latent (unobserved) data, maximizing L is often intractable. The principle of EM applied to the MLE framework is to build an auxiliary function $Q(\theta)$ which is easier to maximize than the observed log-likelihood $L(\theta)$, while its maximization will give a local optimum of the observed log-likelihood. In the conventional EM framework, Q is defined as the expected log-likelihood of the complete data (x, h), with latent data $h \triangleq h_1, \dots, h_N$, conditioned on the observation x only:

$$\begin{aligned} Q_{\theta_i}(\theta) &\triangleq E[\ln p(x,h;\theta)|x;\theta_i] && (2)\\ \theta_{i+1} &\triangleq \arg\max_{\theta} Q_{\theta_i}(\theta) && (3) \end{aligned}$$

Iteratively running Eq. (2) and (3) gives a sequence $\{\theta_i\}$ which converges to a local maximum of L.

2.2. Online EM

If the complete data (x, h) follow a density in the (Natural) Exponential Family[1] (EF, [9]):

$$\ln p(x,h;\theta) \triangleq \langle s(x,h), \theta\rangle + s_0(x,h) \; - \psi(\theta) \tag{4}$$

where s, s_0 and $\psi(\theta)$ are functions of appropriate dimension which define the density and $\langle .,. \rangle$ the scalar product, Eqs. (2) and (3) may be rewritten with the following concise form:

$$\theta_{i+1} = f(\overline{s}(x;\theta_i)) \tag{5}$$

with f defined as:

$$f(s) \triangleq \arg\max_{\theta} \left[\langle s, \theta\rangle - \psi(\theta)\right]$$

and $\overline{s}(x;\theta_i)$ the averaged sufficient statistics under the parameter θ_i

$$\overline{s}(x;\theta_i) \triangleq \frac{1}{N}\sum_{n=1}^{N} \overline{s}(x_n;\theta_i) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[s(x_n,h_n)|x_n;\theta_i] \tag{6}$$

As updating θ_{i+1} requires all the data (Eq. (6)), the EM algorithm cannot be used for online unsupervised classification where the parameters need to be updated after each observation x_n. The online extension of the EM algorithm as developed in [10], is based on the formulation of the EM algorithm as in Eq. (5). Cappé et al. [10] proposed the following online EM algorithm

$$\begin{aligned} \hat{s}_{n+1} &= \hat{s}_n + \gamma_{n+1}\big(\overline{s}(x_n;\hat{\theta}_n) - \hat{s}_n\big) && (7)\\ \hat{\theta}_{n+1} &\triangleq f(\hat{s}_{n+1}) && (8) \end{aligned}$$

where γ_n is a learning parameter. The iteration index i is replaced by the frame index n, and the parameter $\hat{\theta}_n$ is updated after every observation x_n. As this online derivation is close to the original EM formulation, it can be easily implemented for models such as Gaussian Mixture Models (GMM) as the explicit formula for the average sufficient statistics $\overline{s}$ is exactly the same as in the case of GMM estimation with the conventional EM algorithm. The main constraint of the method is making the relationship between sufficient statistics and $\hat{\theta}$ explicit (i.e. f must be explicited) [10].

2.3. Application to Voice Activity Detection and Limitations

When online EM is applied directly to the estimation of a binary mixture of Gaussian distributions, it can be used for concurrent noise/speech level estimation, where each class (speech and noise) is assumed to be normally distributed, and where each Gaussian parameters are updated frame per frame. When the components of the mixture are mostly overlapping, the mixture does not properly represent a two-class model, and the decision value obtained from the Bayesian classifier is not reliable. This may be the case at the beginning of the online EM algorithm (where little data has been seen by the classifier), or when no speech has been present for a significant amount of time. The main contribution of this work is to use online model comparison in the Bayesian framework to alleviate those issues.

3. VARIATIONAL BAYES APPROACH

3.1. Using Free Energy for Model Comparison

For a latent model $p(x, h|\theta, m)$ of parameter θ and structure m[2], Bayesian estimators are built from the posterior over hidden and parameter variables:

$$p(\theta, h|x, m) = \frac{p(x, h|\theta, m)p_0(\theta|m)}{p(x|m)} \tag{9}$$

where $p_0(\theta|m)$ is the prior, and $p(x|m)$ only depends on the model and the observations:

$$p(x|m) = \int p(x, h|\theta, m)p_0(\theta|m)dhd\theta \tag{10}$$

Although the quantity $p(x|m)$ can be ignored when computing the posterior (since it depends neither on θ or h), it is useful when considering model comparison based on the following:

$$p(m|x) = p(x|m)p_0(m)/p(x) \tag{11}$$

To make computation tractable, we use the Variational Bayes framework (VB [5]) which restricts the posterior $q(\theta, h) \triangleq p(\theta, h|x, m)$ to a simpler functional form, making integrals involved in Bayesian computation tractable for a large class of models. Gaussian mixtures are a particular case. For any function $\tilde{q}(h, \theta)$ over the hidden data h and parameter θ, the Kullback-Leibler divergence between $\tilde{q}(h, \theta)$ and the true posterior $q(h, \theta)$ can be computed using Eq. (9) and (10) as follows:

$$\begin{aligned} KL(\tilde{q}||q) &\triangleq \int \tilde{q}(\theta, h) \ln \frac{\tilde{q}(\theta, h)}{p(\theta, h|x, m)} d\theta dh \\ &\triangleq \ln p(x|m) - F_m(\tilde{q}) \geq 0 && (12) \end{aligned}$$

where the Free energy F_m is defined as:

$$F_m(\tilde{q}) \triangleq \int \tilde{q}(\theta, h) \ln \frac{p(x, h|\theta, m)p_0(\theta|m)}{\tilde{q}(\theta, h)} d\theta dh \tag{13}$$

[1] x is also said to follow a density in the Exponential Hidden Family (EHF)

[2] For a mixture of Gaussians, m may represent the number of Gaussians in the task addressed in this work.

and the inequality (12) is derived from the Kullback-Leibler divergence definition, and a consequence of the Jensen's inequality applied to the concave function $(\ln)$. Inequality (12) shows that F_m is a lower bound of the log-marginalized likelihood for any $\tilde{q}$. Thus, maximizing the negative free energy $-F_m$ with respect to the approximate distributions $\tilde{q}$ will give an approximation of the log-marginalized likelihood $p(x|m)$. As Bayesian model comparison is based on evaluating the log-marginalized likelihood for different models, F_m may be used in place of the log-marginalized likelihood for model comparison if it is a good approximate of the log-marginalized likelihood.

3.2. Variational Bayes EM (VB-EM)

The maximization of $-F_m$ is done using the tools of calculus of variations, which is a branch of mathematics concerned with functionals, that is functions of functions. For practical computation, the VB method is often restricted to densities within the EHF, as in Section 2, that is $p(x, h|\theta, m)$ will be given by Eq. (4). In a Bayesian context, the EHF also has the advantage to always have at least one prior conjugate to the likelihood, that is the resulting posterior has the same functional form as the prior [9]:

$$\ln p_0(\theta|\tau_0, \alpha_0) \quad \propto \quad \langle \theta, \alpha_0 \rangle - \tau_0 \psi(\theta) \tag{14}$$

where τ_0, α_0 are the hyper-parameters. The scalar τ_0 can be interpreted as the pseudo count of the prior, e.g. for N observations, a weak prior will be such as the ratio $\tau_0/(\tau_0 + N) \ll 1$. The vector α has the same dimension as θ, and is the prior on the possible values for θ.

The Variational Bayes framework optimizes the negative free energy with respect to $\tilde{q}$, under the assumption $\tilde{q}(\theta, h) \approx q_\theta(\theta) q_h(h)$ [5]. In this context, maximization of $-F_m$ is reduced to a set of two coupled equations involving q_θ and q_h. This is solved iteratively, to give the VB-EM algorithm. At iteration i:

$$q_\theta^{i+1}(\theta) \quad = \quad p(\theta|\tau_{i+1}, \alpha_{i+1}) \tag{15}$$

$$q_h^{i+1}(h) \quad = \quad \prod_{n=1}^{N} p(h_n|x_n; \bar{\theta}_i) \tag{16}$$

with:

$$\overline{\theta}_{i+1} = \overline{\theta}(\tau_i, \alpha_i) \quad \triangleq \quad \int \theta q_\theta^i(\theta; \tau_i, \alpha_i) d\theta \tag{17}$$

$$\tau_{i+1} \quad = \quad \tau_0 + N \tag{18}$$

$$\alpha_{i+1} \quad = \quad \alpha_0 + \sum_{n=1}^{N} \bar{s}(x_n; \overline{\theta}_{i+1}) \tag{19}$$

q_h and q_θ are used for parameter estimation, as well as for model comparison, by replacing $\tilde{q}(\theta, h)$ by q_h and q_θ in the free energy definition (Eq. (13)). Updates of posteriors q_h and q_θ are reduced to updates of averaged parameter $\overline{\theta}_{i+1}$ and hyper-parameters $(\tau_{i+1}, \alpha_{i+1})$, and they keep the same parametrized forms for each iteration [5].

3.3. Online VB-EM

As in the conventional EM algorithm, the VB-EM algorithm requires the whole dataset at once. An online extension of VB-EM similar to the online extension of the EM algorithm reviewed in Section 2 may be derived, once the VB-EM algorithm is formulated in a form similar to Eq. (5). Here, instead of an explicit relation on the estimation of parameter θ between successive iterations (Eqs. (7)-(8)), we require a similar relationship on the hyper-parameters (τ_i, α_i). As explicitly formulated in [7], one iteration of the VB-EM procedure can be rewritten as:

$$(\tau_{i+1}, \alpha_{i+1}) = g(\bar{s}(x; \overline{\theta}(\tau_i, \alpha_i)) \tag{20}$$

where g is linked to a parametrized free energy F_m^p, defined as the free energy F_m where $\tilde{q}(\theta, h)$ is replaced by q_h and q_θ as defined in Eqs. (15) and (16):

$$g(\bar{s}(x; \overline{\theta}(\tau_i, \alpha_i))) \triangleq \arg\max_{\tau, \alpha} F_m^p(\tau, \alpha, \tau_i, \alpha_i) \tag{21}$$

The online extension of the VB method is thus in principle similar to the online extension of the EM applied to the MLE. For every observation x_n, a stochastic approximation $\hat{F}_{n,m}^p$ of F_m^p is considered, and a series of approximated hyper-parameters $\{(\hat{\tau}_n, \hat{\alpha}_n)\}$ is recursively estimated to maximize F_m^p. At sample $n+1$, this is written as [7]:

$$\hat{s}_{n+1} \quad \triangleq \quad \hat{s}_n + \gamma_{n+1} \left[\bar{s}(x_{n+1}; \overline{\theta}(\hat{\tau}_n, \hat{\alpha}_n) - \hat{s}_n \right] \tag{22}$$

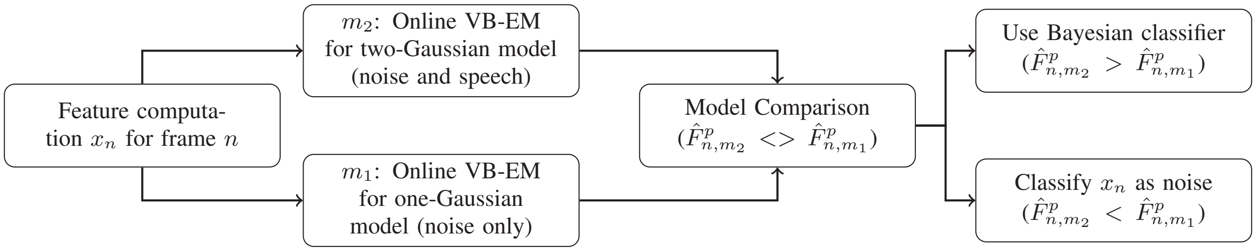
$$(\hat{\tau}_{n+1}, \hat{\alpha}_{n+1}) \quad \triangleq \quad g(\hat{s}_{n+1}) \tag{23}$$

Those online updates of hyper-parameters are used to compute $\hat{F}_{n,m}^p$ itself, thus both model comparison and model parameters are computed in an online manner.

4. APPLICATION TO VOICE ACTIVITY DETECTION AND EVALUATION

The online VB-EM is applied to VAD in a straightforward manner; using a one- dimension feature (enhanced High Order Statistics [6]), we conduct the online VB-EM for models with one and two components at the same time as well as estimating $\hat{F}_{n,m_1}^p$ (online free energy for single-Gaussian mixture) and $\hat{F}_{n,m_2}^p$ (online free energy for two-Gaussian mixture). The model with one component corresponds to the noise-only case, and the model with two components the noise-and-speech case. Both models are estimated concurrently, and we update the classifier assuming a model with two components, but when $\hat{F}_{n,m_1}^p > \hat{F}_{n,m_2}^p$, we assume the signal contains only noise for the corresponding samples, and vice versa. This is summarized in Figure 1. Since online VB-EM and free energy is computed on simple models, the computational cost is minimal (Real Time Factor of 0.03 in our implementation on a standard workstation).

We evaluate this method on the CENSREC-1-C database [11], which consists of noisy continuous digit utterances in Japanese. The recordings were done in two kinds of noisy environments (street and restaurant), with low and high SNRs. We use the remote data (where the speaker is approximately 50 cm away from the microphone). The results are given by frame error rates: False Alarm Rate (FAR: ratio of noise frames detected as speech divided by the number of noise frames) and False Rejection Rate (FRR: ratio of speech frames detected as noise divided by the number of speech frames). The results are compared against the statistical VAD described in [12]. The method is based on spectral features and HMM-based hangover; the HMM emission probabilities are computed from a-priori and a-posteriori noise statistics, which are estimated from the beginning of the signal, and updated on a frame-per-frame basis respectively. For each method, the decision level was set up so that FAR and FRR were approximately equal (equal error rates) on the whole dataset. The method is also compared against a VAD method based on online



m_2: Online VB-EM for two-Gaussian model (noise and speech)
Feature computation x_n for frame n
m_1: Online VB-EM for one-Gaussian model (noise only)
Model Comparison ($\hat{F}^p_{n,m_2} <> \hat{F}^p_{n,m_1}$)
Use Bayesian classifier ($\hat{F}^p_{n,m_2} > \hat{F}^p_{n,m_1}$)
Classify x_n as noise ($\hat{F}^p_{n,m_2} < \hat{F}^p_{n,m_1}$)

Fig. 1. Proposed scheme based on online VB-EM

Table 1. Results of VAD per SNR.

Proposed method	High SNR	Low SNR	Overall
FAR	17.2 %	21.4 %	19.2 %
FRR	8.6 %	29.6 %	18.9 %
Without model comparison	High SNR	Low SNR	Overall
FAR	15.2 %	26.8 %	21.0 %
FRR	13.1 %	30.9 %	22.0 %
Statistical VAD	High SNR	Low SNR	Overall
FAR	19.9 %	31.1 %	25.5 %
FRR	16.0 %	33.3 %	24.7 %

Table 2. Results of VAD per noise type.

Proposed method	Restaurant	Street	Average
FAR	24.6 %	14.9 %	19.7 %
FRR	17.6 %	20.6 %	19.1 %
Statistical VAD	Restaurant	Street	Average
FAR	49.1 %	1.6 %	25.4 %
FRR	14.3 %	33.8 %	24.1 %

VB-EM, but without using online free energy: in that case, we assumed a two-Gaussian model throughout the whole signal, and used the decision level given by the corresponding Bayesian classifier.

The results in Table 1 show that the proposed method (top) outperformed the statistical VAD (bottom). Both FAR and FRR are reduced in both low and high SNR cases. Compared with the online VB-EM without model comparison (middle), although FAR in high SNR case was degraded, the other cases were improved, and the effect of incorporating online free energy is confirmed. In Table 2, we compare the proposed method and the statistical VAD method per noise type. Although the proposed method is worse than the statistical VAD in some cases, it is better on average and is also more consistent across noise types. The proposed method is also less biased toward street noise compared to the statistical VAD, as the noise-specific error rates are closer to the average error rates over both noise types. The result suggests that the proposed online VAD method can operate in different noise conditions. We thus confirm the effectiveness of the proposed method.

5. CONCLUSION

A new scheme to improve the robustness of online classification for VAD has been proposed. It uses online free energy, an online approximation of log-evidence in the Variational Bayes framework, to assess the classifier reliability online. The method is intended to replace the state machines with a statistical solution, and may be applied to other problems where different mixture models must be compared online, such as speaker diarization. This will be investigated in future works.

6. REFERENCES

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM Spine evaluation system," in *ICASSP*, 2002.

[3] I. Shafran and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," in *Proc. ICASSP*, 2003, pp. 432–435.

[4] K. Li, M. S. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using high order statistics," *Speech and Audio Processing, IEEE Trans. on*, vol. 13, no. 5, pp. 965–974, September 2005.

[5] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, vol. 7, pp. 453–464, 2002.

[6] D. Cournapeau and T. Kawahara, "Using variational bayes free energy for unsupervised voice activity detection," in *Proc. ICASSP*, 2008.

[7] M. Sato, "Online model selection based on the variational Bayes," *Neural Computation*, vol. 13, pp. 1649–1681, 2001.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 39, no. 1, pp. 1–38, 1977.

[9] D. R. Cox, *Principles of Statistical Inference*, Cambridge university press, 2006.

[10] O. Cappé and E. Moulines, "Online EM algorithm for latent data models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 593–613, 2009.

[11] N. K. et al., "CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment (in Japanese)," Tech. Rep., IPSJ SIG SLP, 2006.

[12] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett*, vol. 6, pp. 1–3, 1999.