USING VARIATIONAL BAYES FREE ENERGY FOR UNSUPERVISED VOICE ACTIVITY DETECTION

David Cournapeau, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
david@ar.media.kyoto-u.ac.jp

ABSTRACT

This paper addresses the problem of Voice Active Detection (VAD) in noisy environments. We introduce Variational Bayes approach to EM for classification to replace the heuristic state machines. The Variational Bayes approach provides an explicit approximation of the evidence called Free Energy. Free Energy is used to assess the reliability of the classification model, and can be periodically updated with a small number of samples. We apply this scheme to the detection of invalid classification caused in noise-only portions for more reliable VAD, avoiding some of the heuristics conventionally used in many VAD algorithms. An experimental evaluation is conducted on the CENSREC-1-C database for VAD evaluation, and the proposed method gives a significant improvement.

Index Terms— Voice Activity Detection, online EM, Variational Bayes, Free Energy

1. INTRODUCTION

Voice Activity Detection (VAD), which automatically detects speech from audio signals, plays an important role in many speech applications. VAD is often used as a pre-processing step for ASR, speaker recognition and speech coding.

Most VAD algorithms consist in two parts: the first one performs the feature extraction, and the second one the classification. For the classification, supervised classifiers based on techniques such as SVM [1], GMM [2] and HMM [3] have been used. We explore another approach for unsupervised, real-time classification. It is often realized with a state machine system with a threshold based on SNR estimation. But as noted in [2], conventional state-machine systems often rely on heuristics for noise floor estimation. The goal of this study is to propose a simple statistical model for online classification, providing a more robust, less heuristic classification scheme.

We assume that a feature for VAD, such as energy, spectrum or High Order Statistics (HOS, as we proposed in [4]), is distributed as a binary mixture of Gaussian, whose state is estimated using online EM [5][6]. Each Gaussian is then assumed to be representative of one class (speech or non-speech). Thus, the statistical model gives a concurrent speech/noise level estimation, without the requirement of noise floor estimation. This method gives satisfactory results [4], but conceptually suffers from some deficiencies: when speech is not present for some time (or not present at all, e.g. at the beginning of the signal), the statistical model is forced to look for two components, which may not be representative of two classes. In order to enhance the online EM classification, in this paper, we incorporate assessment of the reliability of the model, using a Bayesian approach to EM for model comparison.

The organization of the paper is as follows: the online EM method as well as its limitations is reviewed in Section 2. In Section 3, we show how the evidence of the observation in a Bayesian context can be used to overcome these limitations. Free Energy, a practical estimation of the evidence in Variational Bayes approximation, is reviewed and its behavior on simple examples is presented in Section 4. An evaluation on CENSREC-1-C, a framework for noise robust VAD evaluation, is then presented in Section 5.

2. ONLINE EM FOR CLASSIFICATION: ADVANTAGES AND LIMITATIONS

When we assume unsupervised classification without training data, the classification often relies on thresholding the feature, whose value is estimated and updated from the background noise level. Instead, we adopt a simple model where each class (speech/non-speech) is represented by one Gaussian, and use an online EM algorithm to estimate the parameters of the binary mixture [4]. By estimating the mixture online, we realize a concurrent speech/noise level estimation. Once some speech samples are available to the algorithm, the model parameters start changing and adapting to the signal, and the resulting probability density function (pdf) can be used for the following classification.

Nevertheless, this scheme suffers from some deficiencies. First, at the beginning of the signal, because there is only noise or speech, the training of the Bayesian classifier

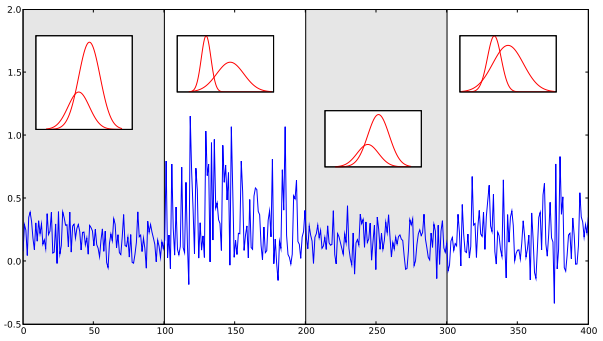

Fig. 1. Example generative model: mostly overlapping (uni-modal) vs. well separated mixtures (multi-modal)

is highly unreliable; this problem can be somewhat alleviated by using some heuristics (as used in many works, assuming that the first second of the signal is noise only), but we present a more theoretically sound solution. Also, when there is no speech for a long time, the means of the mixture components will become close to each other, and as such, again, the classifier will be unreliable. Both problems are related to the fact that, when the Gaussian distributions of the mixture are mostly overlapping, the mixture does not properly represent two-class model as designed.

3. USING MODEL COMPARISON TO ASSESS MODEL RELIABILITY

3.1. Revisiting the model: When does a binary mixture really model two classes?

Intuitively, the statistical model used in online EM can be simply described as a binary mixture, whose state changes in time. If we generate data from a model which is 'locally' distributed as a binary mixture of Gaussian, and whose state can change abruptly (as in HMM), we obtain a behavior similar to Fig. 1. In this figure, the data were generated from four different mixtures (alternating the background to illustrate the change of the mixture state). We can observe that when the components are mostly overlapping, the feature distribution looks like noise; only the second section shows there are two different underlying classes. To answer the question whether a given mixture models one or two classes in an objective manner, we propose to use Bayesian inference for model comparison, that is, whether a model with one component or a model with two or more components is more likely to describe the observed data.

3.2. Using Bayesian inference for model comparison

In Bayesian inference, parameters are assumed to be random variables, and estimators are based on posterior probabilities. One advantage of this approach is that the model itself can be regarded as a random variable, and thus can be inferred using the data (see [7] chapter 28). For a given Gaussian mixture model m_j of j components, the joint pdf for the observation O, the latent data H, and the parameters θ is given by the pdf $p(O, \theta, H|m_j)$; Bayesian estimators are then based on the posterior $p(\theta, H|O, m_j)$:

$$p(\theta, H|O, m_j) \propto p(O|\theta, H, m_j) \cdot p_0(\theta, H|m_j) \quad (1)$$

where $p_0(\theta, H|m_j)$ is the prior of the parameters and hidden variables given the model m_j. But because the model m_j itself is also a random variable, we can also estimate the model posterior given the data:

$$p(m_j|O) \quad \propto \quad p(O|m_j) \cdot p(m_j) \quad (2)$$

The marginalized likelihood $p(O|m_j)$, also called the evidence, is obtained by marginalizing over both the parameters θ and the latent variables H:

$$\begin{aligned} p(O|m_j) &= \int p(O, \theta, H|m_j) d\theta dH \\ &= \int p(O|\theta, H, m_j) \cdot p_0(\theta, H|m_j) d\theta dH \quad (3) \end{aligned}$$

To summarize, one of the advantages of Bayesian inference is that a second level of inference is possible, namely, once a prior on the model $p(m_j)$ is given, scoring different models can be done using the evidence (3) through eq. (2). So if we can evaluate the integral (3) for different models, we can compare them, and thus detect cases where the data are better explained by one component than multiple components. The problem is that such integrals are intractable for all but trivial models. We will show in next Section how the Variational Bayes framework, with a few approximation, can approximate the log-evidence through a functional called Free Energy, and provides an explicit measure for model comparison.

4. VARIATIONAL FREE ENERGY FOR BAYESIAN INFERENCE

4.1. Variational Bayesian approach to mixture models

A popular way to estimate integrals such as eq. (3) is Markov Chain Monte Carlo (MCMC). We adopt in this work another approach, Variational Bayes (VB [8][9]), which restricts the posterior $q(\theta, H) \triangleq p(\theta, H|O, m)$ to a simpler functional form, making the integral (3) tractable for a large class of models, of which Gaussian mixtures are a particular case. The later approach has an advantage of being less computationally intensive when applicable [8].

4.1.1. Variational Bayes principles

The main idea of Variational Bayes is to restrict the posterior $q(\theta, H)$ to a factorized form. More precisely, if:

- The prior is conjugate to the likelihood, and
- The true posterior $q(\theta, H)$ is approximated by the factorized distribution: $q(\theta, H) \approx \tilde{q}(\theta, H) \triangleq q_\theta(\theta) \cdot q_H(H)$,

then the integration in eq. (3) can be done analytically. The VB method then maximizes a cost function called Free Energy with respect to the free pdf $q(\theta)$ and $q(H)$, described in next Sub-section.

4.1.2. Free Energy as approximation of evidence

To derive Free Energy, we start from the Kullback-Leibler (KL) divergence between the approximate posterior $\tilde{q}$ and the true one q, from which we derive the log-evidence $\ln p(O|m)$:

$$\begin{aligned} KL(\tilde{q}||q) &\triangleq \int \tilde{q}(\theta, H) \ln \frac{\tilde{q}(\theta, H)}{q(\theta, H)} d\theta dH \\ &\triangleq \ln p(O|m) - F_m(q_\theta, q_H) \geq 0 \end{aligned} \quad (4)$$

where the inequality is by definition of the Kullback-Leibler (direct consequence of the Jensen's inequality), and Free Energy F_m is defined by:

$$F_m \triangleq \int \tilde{q}(\theta, H) \ln \frac{p(O, \theta, H|m)}{\tilde{q}(\theta, H)} d\theta dH \quad (5)$$

So maximizing F_m with respect to the approximate distributions q_θ and q_H minimizes the KL divergence, and approaches the true log-evidence. To maximize F_m, we use the calculus of variations, which is a branch of mathematics concerned with functionals, that is functions of functions (see [10] for a primer). By taking a partial derivative of F_m with respect to q_H and then to q_θ, we obtain the following formulae:

$$q_H(H) \propto \exp\left\{ \int \ln p(O, H|\theta) q_\theta(\theta) d\theta \right\} \quad (6)$$

$$q_\theta(\theta) \propto p_0(\theta) \cdot \exp \int \ln p(O, H|\theta) q_H(H) dH \quad (7)$$

As both eq. (6) and (7) are coupled, we iterate these equations until convergence (measured by F_m); the algorithm is thus similar to EM algorithm [8]. As mentioned in Section 3.2, the log-evidence can be used for model comparison; here, we can use F_m instead, since it is an approximation of the log-evidence. Compared to other measures for model comparison such as the Schwartz Information Criterion, Free Energy does not rely on a large number of samples' approximation. This is particularly useful for our application, as our goal is to compare models when only a few samples are available.

4.2. Examples

We implemented the above algorithm for a Gaussian mixture, first applied it to the artificial data as shown in Section 3.1.

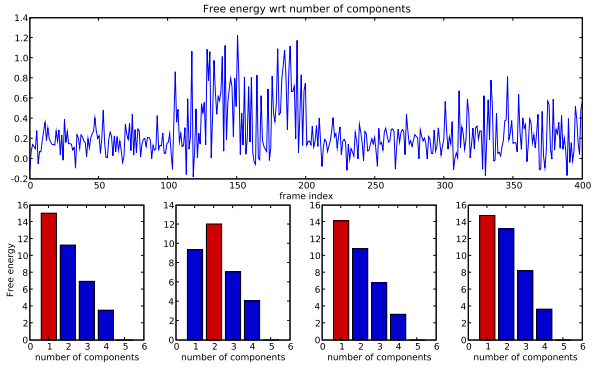

Fig. 2. Results of Free Energy on synthetic data (values translated so that the minimal value is 0).

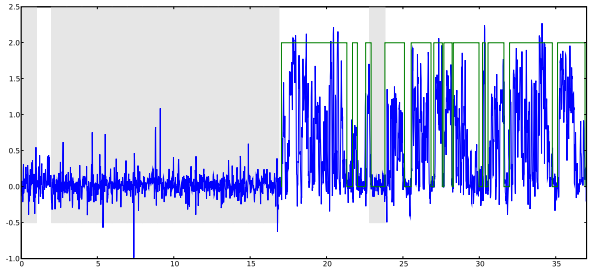

Fig. 3. Real speech example: we compute Free Energy every second, and sections where Free Energy is maximal for one component model are grayed.

We performed the Variational Bayes Expectation Maximization (VB-EM) for each section of 100 samples, with models of one to fives components (we are mostly interested in comparing models with one and two components, but we display here more models to show the global behavior of Free Energy). In Fig. 2, we display the final values of Free Energy for each model and each section. We can observe that on this particular signal, the most probable model (assuming each model equiprobable, i.e. we adopt a flat prior for the model $p(m_j)$) is always the one with one component, except in the second section, where the two components are well separated.

We also computed the VB-EM on a real speech signal, shown in Fig. 3, using the HOS feature which brought significantly better performances than the energy feature [4]. We divided the signal in sections of one second (which correspond to approximately 60 samples in our setting, for a window size of 30ms with 50% overlap), and compared models with one and two components only. The sections where the model with one component being the most probable are grayed.

This provides a simple enhancement of the online EM-based algorithm; every new second, we compute Free Energy, and discard the section if it is best explained by the one-component model, judging that the section contains only

Table 1. VAD performance on CENSREC-1-C database

Proposed method	FAR	FRR	GER
high SNR	5.3 %	7.9 %	6.1 %
low SNR	7.8 %	5.4 %	6.8 %
Without model/data selection	FAR	FRR	GER
high SNR	8.7 %	8.0 %	8.5 %
low SNR	9.5 %	9.6 %	9.5 %

noise. We then perform the classification as conventional for other sections. The computational cost of the VB-EM method is of the same order of complexity as the online EM method.

5. EVALUATION IN VAD PERFORMANCE

As an experimental evaluation, we tested the proposed method on a public database, CENSREC-1 [11]. This database consists of noisy continuous digit utterances in Japanese. The recordings were realized in two kinds of noisy environments (street and restaurant), and high (SNR $>$ 10 dB) and low (-5 $\leq$ SNR $\leq$ 10 dB) SNRs. For each of these conditions, close and remote recordings were available [11]; in this study, we used the close recordings as the HOS feature is more suited to the close talking speech. 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), False Rejection Rate (FRR: ratio of speech frames detected as noise divided by the number of speech frames), and Global Error Rate (GER: weighted mean of FAR and FRR, the weights being the relative ratio of speech and noise frames). The results by using online EM without model/data selection based on Free Energy are also given in Table 1. An overall improvement is observed with the proposed method: both FAR and FRR are reduced; the GER is reduced by 2.4 points for high SNR, and 2.7 points for low SNR.

6. CONCLUSION

A new scheme to improve the reliability of classification based on online EM has been proposed. It uses Free Energy, an approximation of log-evidence in the Variational Bayes framework, to assess the classifier online. Since Free Energy is not derived from large numbers' approximation, it can be used successfully with a relatively small number of samples. The method is intended to replace the state machines, and thus can be applied to other problems than VAD, providing a simple statistical solution without relying on heuristics.

7. ACKNOWLEDGMENTS

The authors would like to thank W. Penny, who kindly provided his Matlab implementation of Variational Bayes, which was used to verify our own implementation.

8. REFERENCES

[1] Dong Enqing, Liu Guizhong, Zhou Yatong, and Zhang Xiaodi, "Applying Support Vector Machine to Voice Activity Detection," in *6th International Conference on Signal Processing Procedings (ICSP'02)*, 2002.

[2] Jashmin K. Shah, Ananth N. Iyer, Brett Y. Smolenski, and Robert E. Yantorno, "Robust voiced - unvoiced classification usgin novel features and gaussian mixture model," in *IEEE ICASSP'04*, 2004.

[3] Sumit Basu, *Conversational Scene Analysis*, Ph.D. thesis, MIT, 2002.

[4] D. Cournapeau and T. Kawahara, "Evaluation of real-time voice activity detection based on high order statistics," in *Proceedings of Interspeech07*, 2007.

[5] Masa-aki Sato, "Convergence of on-line EM algorithm," in *7th International Conference on Neural Information Processing*, 2000, vol. 1.

[6] Olivier Cappé, Maurice Charbit, and Eric Moulines, "Recursive em algorithm with applications to doa estimation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006.

[7] David J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[8] Matthew J. Beal and Zoubin Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, vol. 7, 2002.

[9] U. Noppeney, W. D. Penny, C. J. Price, G. Flandin, and K. J. Friston, "Identification of degenerate neuronal systems based on intersubject variability," *Neuroimage*, vol. 30, pp. 885–890, 2006.

[10] I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, Dover, 2000.

[11] Norihide Kitaoka et al., "CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment," Tech. Rep., IPSJ SIG technical report, 2006.