社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（36）
2007／12／20

変分ベイズ法の自由エネルギーを用いた雑音に頑健なオンライン発話区間検出

クナーポ　ダビド†　河原　達也†

† 京都大学 情報学研究科
〒606-8501 京都市左京区吉田二本松町

あらまし　本稿では、雑音に頑健な発話区間検出の問題を扱っている。従来の発話区間検出ではヒューリスティックな状態機械が一般的に用いられているが、この代替として、識別器のオンライン EM 学習に変分ベイズ法のアプローチを導入することを提案する。変分ベイズ法では、モデルのデータに対する証拠の近似として自由エネルギーが得られる。この自由エネルギーは、識別器の信頼性の尺度として用いることができ、しかも少数のサンプルで更新することができる。本研究ではこれを利用して、雑音のみからなる区間を検出し、従来法のようにヒューリスティックに依存しない形で、より信頼性の高い発話区間検出の実現を図る。CENSREC-1-C データベースを用いた評価実験の結果、提案する手法により有意な改善を得ることができた。
キーワード　発話区間検出, オンライン EM, 変分ベイズ法, 自由エネルギー

Using Variational Bayes Free Energy for Noise Robust Online Voice Activity Detection

David COURNAPEAU† and Tatsuya KAWAHARA†

† School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract　The problem of Voice Activity Detection (VAD) is addressed. This paper proposes to use Variational Bayes Expectation Maximization for classification as a replacement of heuristic-based state machines for online classification. Because the Variational Bayes framework provides an explicit approximation of the evidence of the model, and can be updated with a small number of samples. It can be used to assess the reliability of the classification model by comparing different alternative models. This model comparison is then used for the detection of invalid classification in noise-only portions for more reliable VAD. The method is evaluated on the CENSREC-1-C database for VAD evaluation, and the proposed method gives a significant improvement compared to a previously presented method.
Key words　Voice Activity Detection, Online EM, Variational Bayes, Free Energy

1. Introduction

We are interested in solving the problem of Voice Activity Detection (VAD), which consists in automatically detecting speech segments from audio signals. VAD plays an important role in many speech applications, and is often used as a pre-processing step for ASR, speaker recognition and speech coding.

Traditionally, VAD algorithms consist in mainly two stages: the first one which extracts a suitable feature, and the second stage for the classification. This work will focus on the classification part. Supervised classifiers based on techniques such as SVM [1], GMM [2] and HMM [3] have already been proposed for VAD; we instead explore a statistically-based approach for unsupervised, real-time classification. Unsupervised VAD algorithms are often realized with a state machine system with a threshold based on SNR estimation. But as noted in [4], 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-based classification scheme, without an explicit stage for noise floor estimation.

We assume that a feature for VAD, such as energy, spectrum or High Order Statistics (HOS, as we suggested in [5], and as suggested previously in [6] and [7]), is distributed as a binary mixture of Gaussian, whose state is estimated using online EM, [8], [9]. 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. This method gives satisfactory results [5], 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 are reviewed in Section 2. In Section 3., we show how the evidence of the observation in a Bayesian context can be used to overcome those 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.

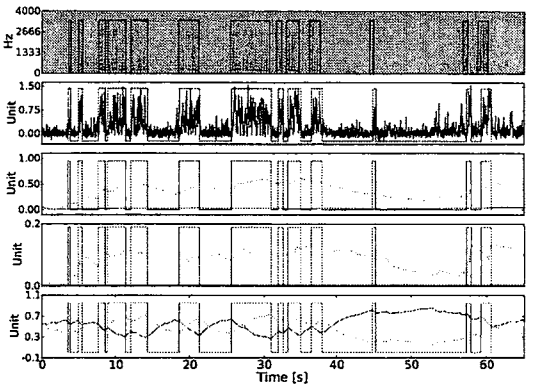

Fig. 1 Spectrogram of audio segment (1^{st}), a one dimensional feature (2^{nd}), means (3^{rd}), variances (4^{th}), and weights (5^{th}) of the components estimated by online EM (dashed red for speech, plain green for noise)

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 [5]. 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.

An example of this method on a relatively clean speech signal is shown on Fig. 1. Once some speech samples are available to the algorithm, the mixture's parameters start changing and adapting to the signal, and the resulting pdf can be used for classification. Nevertheless, this figure also suggests some apparent problems. First, at the beginning of the signal, because there is only noise, the decision value given by the Bayesian classifier is highly unreliable; this problem can be somewhat alleviated by using some heuristics (similar to the ones used in standard machine, like assuming the first second of the signal is noise only), but we would like a more theoretically sound solution. Also, when there is no speech for a long time, the means of each component of the mixture will become really close to each other, and as such, again, the Bayesian classifier will be unreliable. This latter problem is particularly significant in the kind of tasks we are

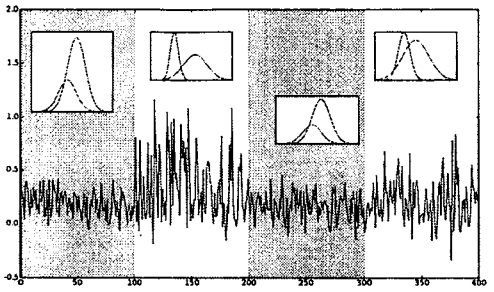

Fig. 2 Example generative model: mostly overlapping (unimodal) vs. well separated mixtures (multi-modal)

interested, where it is possible that no speech is inputted for a long times.

Both problems are related to the fact that when the Gaussian of the mixture are mostly overlapping, the mixture does not really represent a two classes model; the components can then either represent the non gaussianity of the data, or this can just be a consequence of forcing a model of two components where one Gaussian would actually be more representative of the data.

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. 2. 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 looks like 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 ([10] 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:

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

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

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

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, [11], [12]), 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 ([11]).

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 $\ln p(O|m)$:

$$KL(\tilde{q}||q) \triangleq \int \tilde{q}(\theta, H) \ln \frac{\tilde{q}(\theta, H)}{q(\theta, H)} d\theta dH$$
$$\triangleq \ln p(O|m) - F_m(q_\theta, q_H) \geqq 0 \qquad (4)$$

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:

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

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 [13] for a primer). By taking a partial derivative of F_m with respect to q_H and then to q_θ, we get the following formulae:

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

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

As both equation are coupled, we iterate those equations until convergence (measured by F_m); the algorithm can then be seen as a generalization of EM algorithm for MAP estimation, where the distribution over θ would be a Dirac [11], that is MAP EM does a point estimate of θ.

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. In the limit of large number of samples, it can be proved that F_m and the Schwartz Information Criterion (SIC, also confusingly called Bayesian Information Criterion) converges to the same value for a given set of observation and a given same model, thus making the SIC a special case of the free energy, since the SIC uses the large number of samples approximation [11]. The fact that the bound provided by the Free energy is not based on a large number approximation is particularly useful for our application, as we would like to be able to compare models when only a few samples are available.

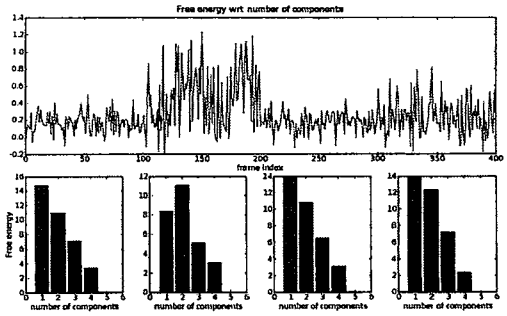

Fig. 3 Results of Free Energy on synthetic data (values translated such as the minimal value is 0).

4. 2 Examples

We implemented the above algorithm for a Gaussian mixture, first applied it to a toy model as in Section 3. 1. 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. 3, 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 adopted 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. 4, using the HOS feature we proposed in [5]. 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. For mostly silent parts, it can be observed that the Free energy is maximized for a model with only one component.

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 noise. We then do the classification as conventional for other sections.

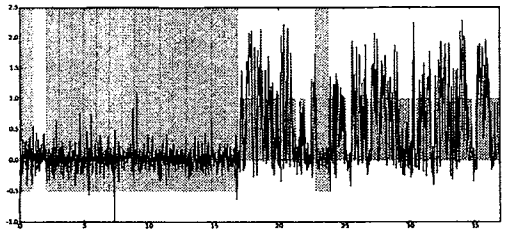

Fig. 4 Clean speech example: we compute Free Energy every second, and sections where Free Energy is maximal for one component model are grayed.

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 %

5. Evaluation in VAD performance

As an experimental evaluation, we tested the proposed method on a public database, CENSREC-1 [14]. This database consists of noisy continuous digit utterances in Japanese. The recordings were done in two kinds of noisy environments (street and restaurant), and high (SNR $>$ 10 dB) and low (-5 $\leqq$ SNR $\leqq$ 10 dB) SNRs. For each of these conditions, close and remote recordings were available [14]; in this study, we used the close recordings as the HOS feature is more suited to close talking speech. The results are given by frame error rates: False Alarm Rate (FAR: ratio of noise frames detected as speech by the number of noise frames), False Rejection Rate (FRR: ratio of speech frames detected as noise 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 on the same dataset 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.

Although it performed reasonably well on the tested dataset for the VAD problem, some improvements can be proposed; in particular, the usage of sections for the Free Energy evaluation is rather ad-hoc. A more elegant approach would be to derive an online version of the Variational Bayes algorithm for mixtures, such as the Free energy would be updated frame by frame, and as such everything (reliability assessment from the Free energy and the classification) would derive naturally from the same model: this will be the object of further studies.

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.

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] Izhak Shafran and Richard Rose, "Robust speech detection and segmentation for real-time ASR applications," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, 2003, vol. 1, pp. 432–435.

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

[6] Elias Nemer, Rafik Goubran, and Samy Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions On Speech And Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[7] Ke Li, M. S. S. Swamy, and M. Omair Ahmad, "An improved voice activity detection using high order statistics," *IEEE Transactions on speech and audio processing*, vol. 13, pp. 965–974, 2005.

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

[9] Olivier Cappe, 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.

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

[11] 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.

[12] 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.

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

[14] 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.