オンライン変分ベイズ学習に基づくモデル比較を用いた音声区間検出

クーナポ　ダビド[1]　　渡部晋治[2]　　中村篤[2]　　河原　達也[1]

1. 京都大学　情報学研究科 知能情報学専攻
〒 606-8501 京都市左京区吉田本町
2.NTT コミュニケーション基礎科学研究所
〒 619-0237 京都府相楽郡精華町光台 2-4

あらまし　　教師なし・オンラインの音声区間検出 (VAD) 方法を提案する。オンライン EM は学習データのない未知の環境にも適用できる枠組みであるが、雑音のみの区間や音声のみの区間が連続すると、モデルの更新が適切に行われないという問題があった。これに対して、提案手法は変分ベイズ EM(VB-EM) 学習に基づいており、その過程で得られる自由エネルギー (Free Energy) をモデルの信頼度比較に利用するものである。VB-EM をオンライン学習に定式化し、モデルパラメータとモデル信頼度の推定を同時・逐次的に行う。CENSREC-1-C を用いた音声区間検出の評価実験により、提案手法が従来のオンライン EM よりも有意に効果的であることを確認した。

Using Online Model Comparison in the Variational Bayes Framework: an Application to 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, Soraku-gun, Kyoto 619-0237, Japan

Abstract　We propose an unsupervised online method for Voice Activity Detection (VAD). The online EM can be applied to any new environments with no training data, but falls in unreliable estimations when the noise-only or speech-only segments last for a long time. The proposed method is based on the Variational Bayes (VB) approach to EM algorithm, and uses Free Energy, which is computed during the estimation process, to assess the model reliability in parallel. An online variation of the VB-EM is formulated for sequential estimation of both model parameters and model comparison measure. An experimental evaluation using the CENSREC-1-C database demonstrates that the proposed method significantly outperformed the conventional online EM method.

1 Introduction

Voice Activity Detection (VAD), which automatically detects speech segments from audio signals, is an important task for many speech applications. It is used for example as a front-end for Automatic Speech Recognition (ASR) [1]. For ASR in noisy environments, the number of insertion errors becomes large [2], and the noise-robust VAD is crucial for the overall performance of the system.

Most VAD algorithms consist of two sub-parts: one which performs feature extraction, and the other for classification itself. In a supervised context, various methods have been studied, such as SVM [3], linked HMM [4] and GMM [5]. The best performance is obtained when the training data and new data have similar distribution; if there is a mismatch between training and unseen data, model adaptation is needed. Another class of methods is based on signal processing approaches, with an explicit noise-signal 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 adapted to the SNR, which is estimated separately. As noted in [6], those state machines often rely on some heuristics for the noise floor estimation. The goal of this study is to develop a statistical model for online classification, with a reliability measure stated as a statistical model comparison problem.

We assume a feature relevant for VAD, such as energy and High Order Statistics (HOS, [7], [8]), is available. If we consider each class (speech and non speech) to follow a normal distribution, the observed feature can be modeled as binary mixture of Gaussian distributions. The classification problem can be reduced to the online estimation of the parameters of a mixture model; we adopt an online derivation of the EM algorithm based on stochastic approximation [9], [10]. To adapt the statistical model to environmental changes, we incorporate assessment of the model's reliability using a Bayesian approach.

For practical computation of the posterior in the Bayesian context, we use the Variational Bayes (VB) framework [11]. The VB framework provides an explicit approximation of the log-evidence called the free energy, which can be used for model comparison [12]. Online extension of the Variational Bayes based on the stochastic approximation [13] of the free energy [14] can be 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, hence both classifier parameters and reliability are estimated from the same statistical model.

The organization of the paper is as follows. Section 2 introduces online EM for unsupervised, online classification in the context of mixture models. Section 3 reviews the VB-EM framework for explicit computation of the free energy, for model comparison. Based on the stated equivalence between the VB-EM procedure and direct minimization of the parametrized free energy, we review the online extension of the VB-EM using a stochastic approximation of the parametrized free energy, for online model comparison. 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 we assume unsupervised classification without training data, the classification often relies on thresholding the feature [6]. The threshold is estimated and updated from the background noise level, and the frame-level speech/non-speech classification is converted to speech boundaries using a hangover scheme. This is the most straightforward method for unsupervised classification.

If we use a statistical framework instead, presence/absence of speech can be regarded as the realization of a binary random variable C, and the feature as the realization of a random variable (or vector for multi-dimensional features) x. If we assume each class is Gaussian, the observation model is a GMM, and estimation can be tackled using the Expectation-Maximization (EM) [15] applied to latent models. Unfortunately, each iteration of the EM algorithm requires the whole dataset, and hence cannot be used for online classification where the observations came recursively. An online extension has been proposed to the EM algorithm recently [16], [17], and we applied it to VAD in [18]. 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.

2.1 EM Algorithm

The Maximum Likelihood Estimation (MLE) is hard to compute explicitly for latent models, and EM algorithm is a popular method for optimizing the likelihood directly. Given N IID observations

$x \triangleq x_1, \ldots, x_N$, the likelihood L of θ is defined as:

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

The key principle of EM applied to the MLE framework is to build a function $Q(\theta)$ which is easier to maximize than the observed likelihood $L(\theta)$, while its maximization will give a reasonable estimate of the MLE applied directly to L. The standard EM algorithm defines the function Q as the expected log-likelihood of the complete data (x,h) conditionally on the observation x only:

$$Q_{\theta_i}(\theta) \triangleq E[\ln p(x,h;\theta)|x;\theta_i] \quad (2)$$

$$\theta_{i+1} \triangleq \arg\max_{\theta} Q_{\theta_i}(\theta) \quad (3)$$

where θ_i is the parameter estimated at the i^{th} iteration. Iteratively running Eq. (2) and (3) gives a sequence $\{\theta_i\}$ which converge to a local maximum of complete data likelihood L in general settings. In particular, if the complete data (x,h) follow a density in the (Natural) Exponential Family[1] (EF, [19]):

$$p(x;\theta) \triangleq \int p(x,h;\theta)dh \quad (4)$$

$$p(x,h;\theta) \triangleq \langle s(x,h),\theta\rangle + s_0(x,h) - \psi(\theta) \quad (5)$$

where s is a function of x of the same dimension as θ and is a sufficient statistics for θ, $\langle .,. \rangle$ the scalar product, ψ a function of θ and s_0 another function of x, the computation of Q is reduced to the computation of $s(x,h)$ under the density $p(h|x)$. Noting f the function:

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

The EM algorithm (Eq. (2) and (3)) can then be written as follows:

$$\overline{s}(x_n;\theta_i) \triangleq E[s(x_n,h_n)|x_n,\theta_i] \quad (6)$$

$$\overline{s}(x;\theta_i) \triangleq \frac{1}{N}\sum_{n=1}^{N} \overline{s}(x_n;\theta_i) \quad (7)$$

$$\theta_{i+1} = f(\overline{s}(x;\theta_i)) \quad (8)$$

2.2 Online EM

When the observation come one after another, and the classification needs to be done for each observation, the EM algorithm cannot be used as it is: each iteration of the E step (2) needs all the data at once. Online extensions of the EM algorithm have been suggested, first to alleviate the relatively intensive computational and memory cost at the time EM algorithm was getting popular, and later for online estimation problems. A recent approach is based on recursively approximating Q itself, while keeping the M step essentially the same [16], [17]. The online approximation $\hat{Q}$ of Q is based on the following recursion:

$$\hat{Q}_n(\theta) = \hat{Q}_{n-1}(\theta) + \gamma_n \left[E[\ln p(x_n,h_n;\theta)|x_n;\theta_{n-1}] - \hat{Q}_{n-1}(\theta)\right] \quad (9)$$

where γ_n is a learning parameter. The M step is kept the same as for the offline EM, that is $\hat{\theta}_n$ is set as the maximum of $\hat{Q}$; each iteration of this procedure is repeated once for each new observation x_n (the iteration index and the sample index are now the same). When the complete data are in the EF, the online update equation (9) can be written as:

$$\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1}\left(\overline{s}(x_n;\hat{\theta}_n) - \hat{s}_n\right) \quad (10)$$

$$\hat{\theta}_{n+1} \triangleq f(\hat{s}_{n+1}) \quad (11)$$

The properties of this algorithm, including theoretical considerations on convergence can be found in [10]. In particular, it is proved that the online update in Eq. (9) converges to a stationary point of the Kullback Leibler between the observation density and the model density. We can also note that Eq. (10) does not rely on any matrix inversion, nor does it consider the complete data likelihood and that at each step, θ_n automatically satisfies the parameter constraints, which is not always the case of methods based on updating the parameters themselves (See section 2.4 of [10] for an example with a Poisson Mixture). Since $\overline{s}(x_n)$ only depends on the observation at time n, this procedure defines a practical online estimation for every model where the offline EM is applicable (where both $\overline{s}$ and f can be computed explicitly and efficiently).

2.3 Application to Voice Activity Detection

When applied to the estimation of a binary mixture of Gaussian distributions, online EM can be used for concurrent noise/speech level estimation, where each class (speech and noise) is assumed to be normally distributed. It was shown to give reasonable results in [18], but 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 is highly unreliable;

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

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 Variational Bayes approach

The statistical model used in Sec. 2 can be seen as a a binary mixture, whose state changes in time. To alleviate problems mentioned above, we propose to use the Bayesian framework for inference, in particular for model comparison; that is it is used to compare whether the data are better explained by a model with one or two components.

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:

$$p(\theta, h|x, m) = \frac{p(x, h|\theta, m)p_0(\theta|m)}{p(x|m)} \quad (12)$$

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

$$p(x|m) = \int p(x, h|\theta, m)p_0(\theta)dhd\theta \quad (13)$$

which is called the evidence. Although The evidence is of no interest when computing posterior (since it depend neither on θ or h), it is useful when considering model comparison:

$$p(m|x) := p(x|m)p_0(m)/p(x) \quad (14)$$

To make computation tractable, we use the Variational Bayes framework (VB [11]) which restricts the posterior $q(\theta, h) := p(\theta, h|x, m)$ to a simpler functional form, making integrals involved in Bayesian computation tractable for a large class of models, of which Gaussian mixtures are a particular case. The essence of Variational Bayes is to provide a tractable lower bound of the marginalized likelihood. For any function $\tilde{q}(h, \theta)$ over the hidden data h and parameter θ, the Kullback-Leibler divergence between $\tilde{q}$ and the true posterior $q := p(h, \theta|x, m)$ can be computed as follows:

$$\begin{aligned} KL(\tilde{q}||q) &:= \int \tilde{q}(\theta, h) \ln \frac{\tilde{q}(\theta, h)}{p(\theta, h|x, m)} d\theta dh \\ &:= \ln p(x|m) - F_m(q_\theta, q_h) \geq 0 \quad (15) \end{aligned}$$

where the Free energy F_m is defined as:

$$F_m := \int \tilde{q}(\theta, h) \ln \frac{p(x, h, \theta|m)}{\tilde{q}(\theta, h)} d\theta dh \quad (16)$$

and the inequality (15) is by definition of the Kullback-Leibler divergence, and a consequence of the Jensen inequality applied to the concave function log. Inequality (15) shows that F_m is a lower bound of the 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 marginalized log-likelihood; as Bayesian model comparison is based on evaluating $p(x|m)$ for different models, if F_m is tight enough, it may be used in place of the marginalized likelihood. One can show in particular that in the limit of a large number of samples, F_m and the Bayesian Information Criterion (BIC) are the same [20]: F_m can be considered as a generalization of the BIC in that regard.

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, we will restrict ourselves to densities within the EHF, as in Section 2, that is $p(x, h|\theta, m)$ will be given by Eq. (5). 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 [19]:

$$\ln p_0(\theta|\tau_0, \alpha_0) \propto \langle \theta, \alpha_0 \rangle - \tau_0 \psi(\theta) \quad (17)$$

where τ_0, α_0 are the hyper-parameters: τ_0 is a scalar, and can be interpreted as the pseudo count of the prior, that is for N observations, the ratio $\tau_0/(\tau_0 + N)$ represents the weight of the prior relatively to the total number of observation $\tau_0 + N$ in the posterior $p(\theta|x_1, \ldots, x_N)$; the vector α has the same dimensions as θ. The key idea of the Variational Bayes framework is to optimize the negative free energy with respect to $\tilde{q}$, but by limiting the possible forms for $\tilde{q}(\theta, h) \approx \tilde{q}_\theta(\theta)q_h(h)$. In this context, maximization of $-F_m$ is reduced to a set of two coupled equations, similar to the EM algorithm [11]; at iteration i, the updated hyper-parameters

[2] For mixture of Gaussian, m may represents the number of Gaussian in the task addressed in this work.

τ_{i+1}, α_{i+1} are computed according to the following equations:

$$\tilde{q}(h; \overline{\theta}_{i+1}) = \prod_{n=1}^{N} \tilde{q}(h_n; \overline{\theta}_{i+1}) \tag{18a}$$

$$\begin{cases} \ln \tilde{q}_\theta(\theta) := \langle \theta, \alpha \rangle - \tau \psi(\theta) \\ \tau := \tau_0 + N \\ \alpha := \alpha_0 + \bar{s}(x; \overline{\theta}_{i+1}) \end{cases} \tag{18b}$$

where we note:

$$\overline{\theta}_{i+1} := E_{\tilde{q}_\theta}[\theta] := \int \theta \tilde{q}_\theta(\theta) d\theta \tag{19}$$

$$\tilde{q}(h_n; \overline{\theta}_{i+1}) = p(h_n | x_n, \overline{\theta}_{i+1}) \tag{20}$$

$$\bar{s}(x_n; \overline{\theta}_{i+1}) := E_{\tilde{q}_{h_n}}[s(x_n, h_n) | x_n] \tag{21}$$

$$= \int s(x_n, h_n) \tilde{q}_{H_n}(h_n; \overline{\theta}_{i+1})$$

$$\bar{s}(x; \overline{\theta}_{i+1}) := \frac{1}{N} \sum_{n=1}^{N} \bar{s}(x_n; \overline{\theta}_{i+1}) \tag{22}$$

As mentioned in [21], and as explicitly carried on in [14], those equations can be retrieved from the direct optimization of a parametrized free energy F^p, where $\tilde{q}$ has been replaced by its parametric form as defined in the equation (18). Thus, the VB-EM equation (18) can be written as:

$$\overline{\theta}_{i+1} := \overline{\theta}(\alpha_i, \tau_i) \tag{23}$$

$$\begin{pmatrix} \tau_{i+1} \\ \alpha_{i+1} \end{pmatrix} := g(\bar{s}(x; \overline{\theta}_{i+1})) : \tag{24}$$

where g is a function which can be explicited from direct optimization of F_m^p (See [14] for more details).

3.3 Online VB-EM

The online extension of the VB method is thus in principle similar to the online extension in the EM applied to the MLE; F^p is recursively approximated by $\hat{F}^p$ in a similar fashion as Q was by $\hat{Q}_n$, and the hyper-parameters are updated as in the traditional VB-EM procedure, replacing $\bar{s}$ by $\hat{s}$ in Eq. (24). This gives a recursive update of the hyper-parameters. At sample $n+1$, this is solved as:

$$\hat{s}_{n+1} := \hat{s}_n + \gamma_{n+1} \left[\bar{s}(x_{n+1}; \overline{\theta}_{n+1}) - \hat{s}_n \right] \tag{25}$$

$$\alpha_{n+1} := \hat{s}_{n+1} + \alpha_0 \tag{26}$$

Those online updates of hyper-parameters can be used to compute $\hat{F}^p$ itself, giving an online model comparison measure.

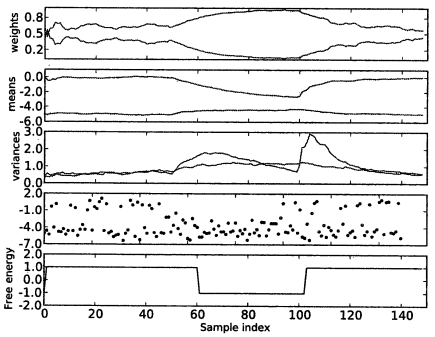

Figure 1: Online-VB-EM procedure applied on simulation data: the weights, means, variances, samples are displayed. The bottom axis is the relative order between F_1 and F_2: if positive, $F_2 > F_1$, and $F_1 > F_2$ otherwise. The sampled model state is changed at sample 50 and 100.

3.4 Example

An example of this procedure is showed on Figure 1, where we sample data from an artificial binary mixture (sample displayed on the second bottom plot). The first 50 samples are sampled from a well separated mixture, then almost overlapping from sample 50 to 100, and then back to the first state starting at sample 100. The weights, means and variances are updated online as well. We run VB-EM for both models with one and two components, and evaluate the online free energy in each case; the bottom axis shows value 1 when $\hat{F}_2^p > \hat{F}_1^p$ and -1 when $\hat{F}_1^p > \hat{F}_2^p$. It is observed that online free energy can track model changes, at least on this simple example.

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, as in [22]), we run the estimation for models with one and two components at the same time as well as F_1 and F_2; we always update the classifier assuming a model with two components, but when $F_1 > F_2$, we assume the signal contains only noise for those sections, although the model keeps being updated.

We evaluate this method on the CENSREC-1-C dataset [23] This database consists of noisy con-

Table 1: Results of the proposed VAD, compared to online-EM based, without model selection

Proposed method	FAR	FRR
High SNR, Average	4.6 %	4.4 %
Low SNR, Average	4.1 %	5.0 %
Without model selection	FAR	FRR
High SNR, average	8.7 %	8.0 %
Low SNR, average	9.5 %	9.6 %

tinuous 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 $\leq$ SNR $\leq$ 10 dB) SNRs. For each of these conditions, close and remote recordings were available [23]; 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) and False Rejection Rate (FRR: ratio of speech frames detected as noise divided by the number of speech 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.

5 Conclusions

A new scheme to improve the reliability of online classification based on online VB-EM has been proposed. It uses online free energy, an online approximation of log-evidence in the Variational Bayes framework, to assess the classifier online. The method is intended to replace the state machines, and thus can be applied to problems other than VAD, providing a simple statistical solution without relying on heuristics.

References

[1] Lawrence R. Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[2] Brian Kingsbury, George Saon, Lidia Mangu, Mukund Padmanabhan, and Ruhi Sarikaya, "Robust speech recognition in noisy environments: The 2001-IBM Spine evaluation system," in *ICASSP*, 2002.

[3] 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.

[4] Sumit Basu, "A linked-HMM model for robust voicing and speech detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, 2003.

[5] 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.

[6] 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.

[7] 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.

[8] 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, no. 5, pp. 965–974, September 2005.

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

[10] Olivier Cappé and Eric Moulines, "Online em algorithm for latent data models," 2008.

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

[13] Harold J Kushner and G. George Yin, *Stochastic approximation algorithms and applications*, Springer-Verlag, 1997.

[14] Masa-aki Sato, "Online Model Selection Based on the Variational Bayes," *Neural Computation*, vol. 13, pp. 1649–1681, 2001.

[15] 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 (Methodological)*, vol. 39, pp. 1–38, 1977.

[16] Masa-aki Sato and Shin Ishii, "On-line EM algorithm for the normalized Gaussian network," *Neural Computation*, vol. 12, pp. 407–432, 2000.

[17] O Cappé, M. Charbit, and E. Moulines, "Recursive EM algorithm with applications to DOA estimation," in *Proceedings of 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006*, 2006.

[18] David Cournapeau and Tatsuya Kawahara, "Voice activity detection based on high order statistics and online em algorithm," *IEICE Transactions on Information and Systems*, vol. 12, December 2008.

[19] David Cox, *Principles of Statistical Inference*, Cambridge University Press, 2006.

[20] M.J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London., 2003.

[21] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.

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

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