Correspondence

Voice Activity Detection Based on an Unsupervised Learning Framework

Dongwen Ying, Yonghong Yan, Jianwu Dang, and Frank K. Soong, *Fellow, IEEE*

***Abstract*—How to construct models for speech/nonspeech discrimination is a crucial point for voice activity detectors (VADs). Semi-supervised learning is the most popular way for model construction in conventional VADs. In this correspondence, we propose an unsupervised learning framework to construct statistical models for VAD. This framework is realized by a sequential Gaussian mixture model. It comprises an initialization process and an updating process. At each subband, the GMM is firstly initialized using EM algorithm, and then sequentially updated frame by frame. From the GMM, a self-regulatory threshold for discrimination is derived at each subband. Some constraints are introduced to this GMM for the sake of reliability. For the reason of unsupervised learning, the proposed VAD does not rely on an assumption that the first several frames of an utterance are nonspeech, which is widely used in most VADs. Moreover, the speech presence probability in the time-frequency domain is a byproduct of this VAD. We tested it on speech from TIMIT database and noise from NOISEX-92 database. The evaluations effectively showed its promising performance in comparison with VADs such as ITU G.729B, GSM AMR, and a typical semi-supervised VAD.**

***Index Terms*—Model-based Gaussian clustering, sequential Gaussian mixture model (GMM), speech presence probability, unsupervised learning, voice activity detection (VAD).**

I. Introduction

The function of the voice activity detector (VAD) is to distinguish active speech from nonspeech in utterances. It plays an important role in variant speech communication systems, such as speech coding, speech recognition, speech enhancement, and so on. Its accuracy affects their performance. Especially in adverse environments, a robust VAD can significantly improve these systems' performance [1].

Generally speaking, VADs consist of acoustic features and discrimination models. Early algorithms paid more attention to robust acoustic features to distinguish speech/nonspeech. The energy-based features were the most popular one [2]–[8]. The signal-to-noise ratio (SNR) was usually taken as an energy cue for discrimination [3], [4]. Besides these simple ones, some advanced energy-based features, such as Teager Energy [5], long-term speech information [6], [7], were derived by enhancing the discrimination between speech and nonspeech. The second popular feature was the quasi-periodicity of voiced speech [3], [8]–[10]. It can discriminate speech signal from non-periodicity background noises. The third popular feature was the dynamics of speech signal, which was reflected in the variance of power envelopes [3], [11], or SNR [12]. In addition to these popular features, some other ones such as zero-crossing rate [8], high-order statics in linear predictive coding (LPC) residual domain [13], and so on, were utilized to design VAD.

During the last decade, more VADs focused on statistical models to discriminate speech/nonspeech. Most statistical models aimed to construct classifiers for speech/nonspeech classification. The classical classifier made use of the Gaussian statistical model to describe the DFT coefficients [14]–[18]. Later, the speech was more accurately modeled by a Laplacian model after decorrelating signal with an orthogonal transformation [19]. Based on these researches, multiple observations [20] and multiple statistical models [21] were utilized to further improve the classifiers' performance, respectively. In these typical statistical models, the likelihood ratio of speech to nonspeech was a general cue for classification. Besides the classical methods, a few statistical models aimed at finding the change points between speech and nonspeech. The optimal filters for edge detection [22] and GARCH model [23] were employed to locate the changing points.

These statistical models have a common characteristic. They are generally initialized based on an assumption that utterances always begin with nonspeech signal. An initial nonspeech model is established from the first several frames of an utterance. After that, VADs first discriminate each coming frame as speech/nonspeech, and then feed back the discrimination result to update models. In fact, the nonspeech used for initialization can be taken as hand-labeled samples, and so, this model initialization is actually a supervised learning process. Such assumption for initialization is referred to as "nonspeech beginning" assumption in this correspondence. The updating method based on feedback is called decision-directed learning, which is a way of realizing unsupervised learning. In theory, this modeling approach incorporating the supervised learning with the unsupervised one is referred to as semi-supervised learning [24]. It is the common characteristic of model construction in these VADs.

However, the VADs based on semi-supervised learning have a defect in some practical applications. If an utterance begins with speech signal, such assumption will be unsatisfied so that the nonspeech model is invalidly initialized. Serious speech leakage will be caused at the utterance beginning, and more error may be resulted in from the incorrect initialization. In such situation, the unsupervised learning can still work since that assumption is not necessary for it.

Few VADs have taken advantage of the unsupervised learning. In [25] and [26], the noisy speech signal is unsupervisedly clustered into two classes via LBG algorithm based on the energy feature. One class with larger mean is taken as speech, and the other for nonspeech. In [22] and [27], the logarithmic energy probability density functions (pdf) of speech and nonspeech are estimated by model-based clustering. An optimal threshold for discrimination is derived from the pdfs. The methods of unsupervised clustering bring two benefits to VADs. One is that such assumption is unnecessary for them; the other is that the threshold can be self-regulated at the observed data.

Manuscript received April 03, 2010; revised October 28, 2010; accepted December 23, 2010. Date of publication March 10, 2011; date of current version September 30, 2011. This work was supported in part by the China–Japan Joint Research (Intelligent Speech Interfaces For Ubiquitous Communications) under Project 60811140086/F03, in part by the National Science and Technology Pillar Program (2008BAI50B00), in part by the National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014), and in part by the National Thousand Talents Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Brian Mak.

D. Ying and Y. Yan are with the Thinkit Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China (e-mail: yingdongwen@hccl.ioa.ac.cn; yyan@hccl.ioa.ac.cn).

J. Dang is with Tianjin University, Tianjin 300072, China, and also with the Japan Advanced Institute of Science and Technology, Ishikawa 923-1211, Japan (e-mail: jdang.china@gmail.com).

F. K. Soong is with the Speech Group, Microsoft Research Asia, Beijing 100080, China (e-mail: frankkps@microsoft.com).

Digital Object Identifier 10.1109/TASL.2011.2125953

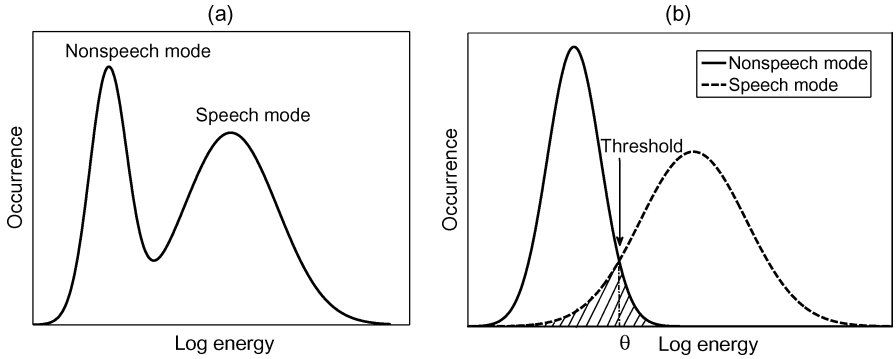

Fig. 1. Schematic illustration of logarithmic energy distribution of a high-SNR subband. (a) Distribution of noisy speech. (b) Distributions of speech and nonspeech.

However, there are still two essential problems to be solved in these VADs. First, a mechanism of incrementally updating models is absent. They are not able to run in an online manner because clustering algorithms are usually conducted in an offline manner. So, these VADs cannot be applied to some real-time systems. Second, it is difficult for them to decide whether one or two clusters are to be formed. In case of speech absence or low SNR, miss-detection of speech is serious if two clusters are formed. For these two reasons, more important works need to be done for developing an unsupervised VAD.

Keeping the above problems in mind, we propose a novel VAD based on an unsupervised learning framework. A sequential GMM is presented to realize this learning process at each band. The initialization with EM algorithm plays the role of model-based Gaussian clustering [28], and the updating process for the role of incremental learning. The two components of this GMM respectively represent the speech and nonspeech distributions. According to the GMM, a self-regulatory threshold is yielded to discriminate speech/nonspeech at each subband. The discrimination results of all bands are summarized by a voting procedure.

The proposed algorithm focuses on the statistical modeling framework, which has three distinctions from conventional statistical VADs [14]–[21]. First, the proposed VAD adopts unsupervised learning. Therefore, it does not rely on the assumption of "nonspeech beginning." Second, some constraints to statistical models are introduced into this framework. On one hand, they can shape the special relationships between speech and nonspeech distributions; on the other hand, by using these constraints, the proposed algorithm can automatically decide one or two clusters to be formed so that the unsupervised framework works well. In past researches, few statistical models pay attention to the distribution relationships. Lastly, the *a priori* distributions of speech and nonspeech are considered in decision, while they are ignored in most VADs.

The rest of the correspondence is organized as follows. Section II of the correspondence will introduce the details of the proposed statistical framework. Section III will give the implementation detail of the proposed VAD. In Section IV, an evaluation is conducted to compare the proposed VAD's performance with other leading VADs. The last section will give discussions and conclusions.

II. Unsupervised Learning Framework for VAD

VADs are generally characterized by acoustic features and classifiers. Here, we select the smoothed subband logarithmic energy as the acoustic feature. The input signal is grouped into several Mel subbands in the frequency domain. Then, the logarithmic energy is calculated by using the logarithmic value of the absolute magnitude sum of each subband. Eventually, it is smoothed to form an envelope for classification. Two Gaussian models are employed as the classifier to describe the logarithmic energy distributions of speech and nonspeech, respectively. These two models are incorporated into a two-component GMM. Its parameters are estimated in an unsupervised way. Speech/nonspeech classification is firstly conducted at each subband. Then, all subbands' decisions are summarized by a voting procedure. In the following subsections, we only consider a classifier in a single subband.

A. Modeling Logarithmic Energy Distribution With GMM

To design a classifier, we begin with the logarithmic energy distribution of a subband. Without loss of generality, we firstly investigate a subband with higher SNR, where both speech and nonspeech are present. Its logarithmic energy distribution of a given time period is described by using the occurrence of a histogram. The logarithmic energy distribution is schematically illustrated in Fig. 1(a). As the noise signal is usually supposed to be more stationary than speech signal, the variance of nonspeech logarithmic energy is smaller than that of the speech. So, there exists a sharp peak corresponding to a nonspeech mode; while the other flat peak relates to a speech mode. Since "nonspeech" denotes the noise signal and "speech" for the superposition of noise and clean speech signals in this correspondence, the speech averaged energy is larger than the nonspeech averaged energy. Accordingly, the nonspeech peak locates at the left side of the speech peak. This distribution consisting of speech and nonspeech modes is referred to as *bimodal distribution* in the following sections.

Assuming that both speech and nonspeech log energies obey the Gaussian distribution, the bimodal distribution can be fitted by a two-component GMM, where one component with the smaller mean is identified as the nonspeech mode and the other component for the speech mode. This model is described by the following equations. Let x_k denote the logarithmic energy of a subband at the time k. z is the speech/nonspeech label, $z \in \{0, 1\}$, where 0 denotes nonspeech and 1 for speech. According to the Bayes' rule, we have the equation

$$p(x_k|\lambda) = \sum_z p(x_k, z|\lambda) = \sum_z p(x_k|z, \lambda)p(z) \tag{1}$$

where $p(z)$ is the prior probability of speech/nonspeech, and is actually equal to the weight coefficient $w_z (w_0 + w_1 = 1)$. $p(x_k|z, \lambda)$ represents the likelihood of x_k given the speech/nonspeech model:

$$p(x_k|z, \lambda) = \frac{1}{\sqrt{2\pi\kappa_z}} exp\left\{-(x_k - \mu_z)^2/2\kappa_z\right\} \tag{2}$$

where μ_z and κ_z, respectively, denote the mean and variance. $\lambda \triangleq \{\mu_z, \kappa_z, w_z|z = 0, 1\}$ is the parameter set of the GMM. An interesting point is that, the mean difference $\mu_1 - \mu_0$ represents the *a posteriori*

SNR because μ_1 and μ_0 are, respectively, the averaged logarithmic energy speech and nonspeech.

Let $\mathbf{x} \triangleq \{x_0, x_1, x_2, \ldots, x_M\}$ be a logarithmic energy sequence at a subband. The pdf is given by

$$p(\mathbf{x}|\lambda) = \prod_{k=0}^{M} p(x_k|\lambda). \tag{3}$$

The parameter set λ is estimated by maximizing the above pdf function.

From the GMM, we can obtain both the pdfs of speech and nonspeech logarithmic energy, namely $p(x_k|z = 1, \lambda)p(z = 1)$ and $p(x_k|z = 0, \lambda)p(z = 0)$. These two pdfs are shown in Fig. 1(b). From the two pdfs, we derive an optimal threshold θ to minimize the classification error. The threshold θ satisfies

$$p(\theta|z = 1, \lambda)p(z = 1) = p(\theta|z = 0, \lambda)p(z = 0). \tag{4}$$

Equation (4) is a quadratic equation with one unknown θ. The threshold is one of its roots locating between the two means, namely $\mu_1 > \theta > \mu_0$. The samples with logarithmic energy less than θ are determined as nonspeech, and otherwise as speech. The shadow in Fig. 1(b) denotes the classification error.

B. Sequential GMM Estimation Scheme

The crucial issue of the above model is to estimate the parameter set λ. As the speech and noise are piecewise stationary signal, the GMM parameters should be adapted to signal variation. The estimation consists of an off-line initialization and a sequential updating process. The initial GMM is first established by the EM algorithm, and then incrementally updated with coming data. The parameter set at time k is denoted as $\lambda_k \triangleq \{\mu_{k,z}, \kappa_{k,z}, w_{k,z}|z = 0, 1\}$. λ_0 is the initial parameter set estimated from the first $M+1$ samples by EM algorithm. The noisy logarithmic power is also assumed to confirm the bimodal distribution in this subsection.

The following are the typical EM re-estimation formulas

$$\bar{w}_{0,z} = \frac{1}{M+1} \sum_{j=0}^{M} p\left(z|x_j, \lambda_0'\right) \tag{5}$$

$$\bar{\mu}_{0,z} = \frac{\sum_{j=0}^{M} x_j p\left(z|x_j, \lambda_0'\right)}{(M+1)\bar{w}_{0,z}} \tag{6}$$

$$\bar{\kappa}_{0,z} = \frac{\sum_{j=0}^{M} (x_j - \bar{\mu}_{0,z})^2 p\left(z|x_j, \lambda_0'\right)}{(M+1)\bar{w}_{0,z}} \tag{7}$$

where

$$p\left(z|x_j, \lambda_0'\right) = \frac{w_{0,z}' p\left(x_j|z, \lambda_0'\right)}{\sum_z w_{0,z}' p\left(x_j|z, \lambda_0'\right)} \tag{8}$$

where λ_0' is the old parameter set, and $w_{0,z}'$ is the weight coefficient of λ_0'. $\bar{\lambda}_0 \sim \{\bar{w}_{0,z}, \bar{\mu}_{0,z}, \bar{\kappa}_{0,z}\}$ denotes the new parameter set re-estimated from λ_0'. In the next iteration, λ_0' is replaced by $\bar{\lambda}_0$. This iteration continues until EM algorithm converges. The final $\bar{\lambda}_0$ is the initial parameter set λ_0 that we are solving for. Given λ_0, the threshold θ_0 is derived by using (4). Eventually, the first $M+1$ samples are classified by θ_0.

After establishing the initial GMM, the problem arrives at updating this model by coming data. A sequential GMM is an efficient solution to this problem. The basic scheme of realizing a sequential GMM is to incrementally update its parameter set by utilizing the latest K samples [29]. Suppose λ_k is known at the time $k+1$, the parameters in λ_{k+1} are derived by iterative equations (9)-(12), shown at the bottom of the page, where the weight, mean, and variance can be regarded as the zero-, first-, and second-order moments of speech/nonspeech logarithmic energy, respectively.

This sequential scheme is not so desirable for VAD. On one hand, $\{p(x_j|z, \lambda_k)|j = k - K + 1, \cdots, k\}$ has to be calculated at each time k. It will result in heavy computational load. On the other hand, it is not beneficial for GMM to track signal variation because the late and early samples do the same contribution to updating models.

Based on the basic scheme, we propose a novel approach of realizing a sequential GMM. Suppose that the GMM varies with time slowly, $\lambda_k \approx \lambda_{k-1}$ at time k. Accordingly, we have the relationship, $\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) \approx \sum_{j=k-K+1}^{k} p(z|x_j, \lambda_{k-1})$. The summation is approximated by the zero-order moment, $\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_{k-1}) \approx K w_{k,z}$, according to (9). Combining these relationships, we finally have the following equation:

$$\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) \approx K w_{k,z}. \tag{13}$$

Substituting (13) into (9), we obtain

$$w_{k+1,z} = \frac{K w_{k,z} + p(z|x_{k+1}, \lambda_k)}{K+1}. \tag{14}$$

Let $\alpha = K/(K+1)$, we obtain the iterative equation

$$w_{k+1,z} = \alpha w_{k,z} + (1-\alpha) p(z|x_{k+1}, \lambda_k) \tag{15}$$

where α can be considered as a forgetting factor, $0 < \alpha \leq 1$; the conditional probability $p(z|x_{k+1}, \lambda_k)$ is calculated via (12).

With the same principle, the summation item in (10) can be approximated by the first-order moment

$$\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) x_j \approx K w_{k,z} \mu_{k,z}. \tag{16}$$

$$w_{k+1,z} = \frac{\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) + p(z|x_{k+1}, \lambda_k)}{K+1} \tag{9}$$

$$\mu_{k+1,z} = \frac{\sum_{j=k-K+1}^{k} x_j p(z|x_j, \lambda_k) + x_{k+1} p(z|x_{k+1}, \lambda_k)}{\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) + p(z|x_{k+1}, \lambda_k)} \tag{10}$$

$$\kappa_{k+1,z} = \frac{\sum_{j=k-K+1}^{k} (x_j - \mu_{k+1,z})^2 p(z|x_j, \lambda_k) + (x_{k+1} - \mu_{k+1,z})^2 p(z|x_{k+1}, \lambda_k)}{\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) + p(z|x_{k+1}, \lambda_k)} \tag{11}$$

$$p(z|x_k, \lambda_k) = \frac{w_{k,z} p(x_k|z, \lambda_k)}{\sum_z w_{k,z} p(x_k|z, \lambda_k)} \tag{12}$$

Substituting (16) into (10), we obtain

$$\mu_{k+1,z} = \frac{\alpha w_{k,z}\mu_{k,z} + (1-\alpha)p(z|x_{k+1},\lambda_k)x_{k+1}}{w_{k+1,z}}. \quad (17)$$

Accordingly, the summation item in (7) is approximated by the second-order moment

$$\sum_{j=k-K+1}^{k} p(z|x_j,\lambda_k)(x_j-\mu_{k+1,z})^2 \approx K w_{k,z}\kappa_{k,z}. \quad (18)$$

Substituting (18) into (11), we obtain

$$\kappa_{k+1,z} = \frac{\alpha w_{k,z}\kappa_{k,z} + (1-\alpha)p(z|x_{k+1},\lambda_k)(x_{k+1}-\mu_{k+1,z})^2}{w_{k+1,z}}. \quad (19)$$

The sequential scheme consists of (12), (15), (17), and (19). By these equations, λ_{k+1} is derived from λ_k and x_{k+1} at the time $k+1$. It is a first-order recursive process. Then, the time-varying threshold θ_{k+1} is obtained from (4) given λ_{k+1}. Finally, x_{k+1} is classified as speech/nonspeech by θ_{k+1}.

The proposed sequential GMM has two advantages over the basic one. One is that, as the likelihood $\{p(x_j|z,\lambda_k)|j=k-K+1,\cdots,k\}$ is not needed for every time k, the computation efficiency is improved a lot. The other is that, the earlier frames are forgotten with time going, and the later frames play a more important role. For this reason, the tracking capability of the proposed algorithm is more powerful than the basic one. Actually, some other schemes were presented to realize the sequential GMM [30], [31]. Compared with them, the proposed scheme makes a compromise among computational efficiency, performance, and memory requirement.

C. Constraints to GMM

In high-SNR subbands, the logarithmic energy confirms the bimodal distribution. However, in low-SNR subbands, the speech mode is unobvious in the logarithmic energy histogram. Especially when speech signal is absent, there exists only the nonspeech mode. This distribution comprising only a nonspeech mode is called as *unimodal distribution*. If the unimodal distribution is fitted by the two-component GMM, a serious error of miss-detection will be resulted in. So, the proposed two-component GMM must be improved to correctly fit the unimodal distribution together with the bimodal one.

For such purpose, it is important to discriminate the unimodal distribution from the bimodal one. The unimodal distribution associates with low SNR, and vice versa for the bimodal one. Since the mean difference represents the posteriori SNR, it can be used as a cue to distinguish the unimodal/bimodal distributions. A threshold δ of the mean difference is set for discrimination. So, a bimodal distribution should satisfy the relationship

$$\mu_{k,1} > \delta + \mu_{k,0}.$$

There is a similar relationship between the speech and nonspeech variances of the bimodal distribution. Under the assumption that the noise signal is more stationary than speech signal, the variances should satisfy the relationship

$$\kappa_{k,1} > \kappa_{k,0}.$$

The parameter set for the bimodal distribution can be estimated by EM algorithm, but the GMM parameters for the unimodal distribution have to be estimated in a special way. Since the unimodal distribution associates with low SNR, all samples can be approximately considered to be nonspeech. So, the mean and variance of the nonspeech component can be taken as that of all samples. For the unimodal distribution, the parameters of the speech component are impossible to be estimated from the real data. We construct a virtual speech component, where its mean and variance are, respectively, set as $\delta + \mu_{k,0}$ and $\kappa_{k,0}$.

Therefore, whether in cases of bimodal or unimodal distributions, the GMM parameters should satisfy the relationship

$$\mu_{k,1} = \max\{\mu_{k,1}, \mu_{k,0} + \delta\} \quad (20)$$

$$\kappa_{k,1} = \max\{\kappa_{k,0}, \kappa_{k,1}\} \quad (21)$$

where δ makes a tradeoff between weak speech spectral components and strong nonspeech spectral components. In high-noise environments, a large δ is beneficial to reject the latter; while a small δ is advantageous to detect the former in low-noise environments.

It is worthwhile noting that the mean constraint will cause the convergence failure of EM algorithm. When the virtual speech component is constructed under the unimodal distribution, the distribution center of the virtual component $\mu_{k,0} + \delta$ will deviate far away from that of all samples. As a result, $p(x_k|z=0,\lambda_k) \gg p(x_k|z=1,\lambda_k)$, and so, according to (12), $p(z=0|x_k,\lambda_k) \gg p(z=1|x_k,\lambda_k)$. In the re-estimation process, $w_{k,1}$ will approach to 0, and so the denominators of (6) and (7) will be zero when estimating the speech component. Finally, the EM algorithm will not converge if the constraint of (20) is activated. This phenomena also occurs in the sequential process. To guard against it, another constraint needs to be placed on the weight coefficients:

$$\begin{aligned} w_{k,1} &= \max\{w_{k,1}, \epsilon\} \\ w_{k,0} &= 1 - w_{k,1} \end{aligned} \quad (22)$$

where ϵ is close and greater than zero. In EM algorithm, the re-estimation will be terminated when this constraint is activated.

These constraints come from the distribution relationships between speech and nonspeech. They play different roles in this framework. Besides shaping the distribution relationship, the mean constraint in (20) decides how many clusters to be formed. When it is activated, all samples are clustered into one class; otherwise, two classes will be formed. The weight constraint in (22) is a slave to it. It will follow the mean constraint to be activated. The constraint function of (23) is to shape the variance relationship. All constraints to GMM are embedded into the initialization and updating processes. The details of applying these constraints are shown in the next section. The constrained model can deal with both unimodal and bimodal distributions.

We respectively design two tests on a high-SNR subband and a low-SNR subband of a noisy utterance to demonstrate the functions of these constraints. In the tests, only the offline fitting is used for demonstration. The envelope of the high-SNR subband is shown in Fig. 2(a). Fig. 2(b) plots its histogram, where the logarithmic energy confirms the bimodal distribution. The fitting result is labeled on the histogram. As the mean difference is much larger than δ, these constraints are not activated. The optimal threshold 1 is labeled on the envelope of Fig. 2(a). Comparing the determined boundary with the corresponding oracle in Fig. 2(f), one can see that the classifier works well in the high-SNR subband. It should be clarified that not all speech signal is present in this subband.

As a further verification of constraints, we apply them on a low-SNR subband in Fig. 2(c). The histogram in Fig. 2(d) and (e) illustrates that its logarithmic energy obeys the unimodal distribution. We first fit this distribution without constraints, as shown in Fig. 2(d), and then fit it with constraints, as shown in Fig. 2(e). The derived thresholds are labeled on the envelopes of Fig. 2(c). The corresponding boundaries are illustrated in Fig. 2(f), where both the boundary 2 and 3 come from the low-SNR band and boundary 1 from the high-SNR band. One can see

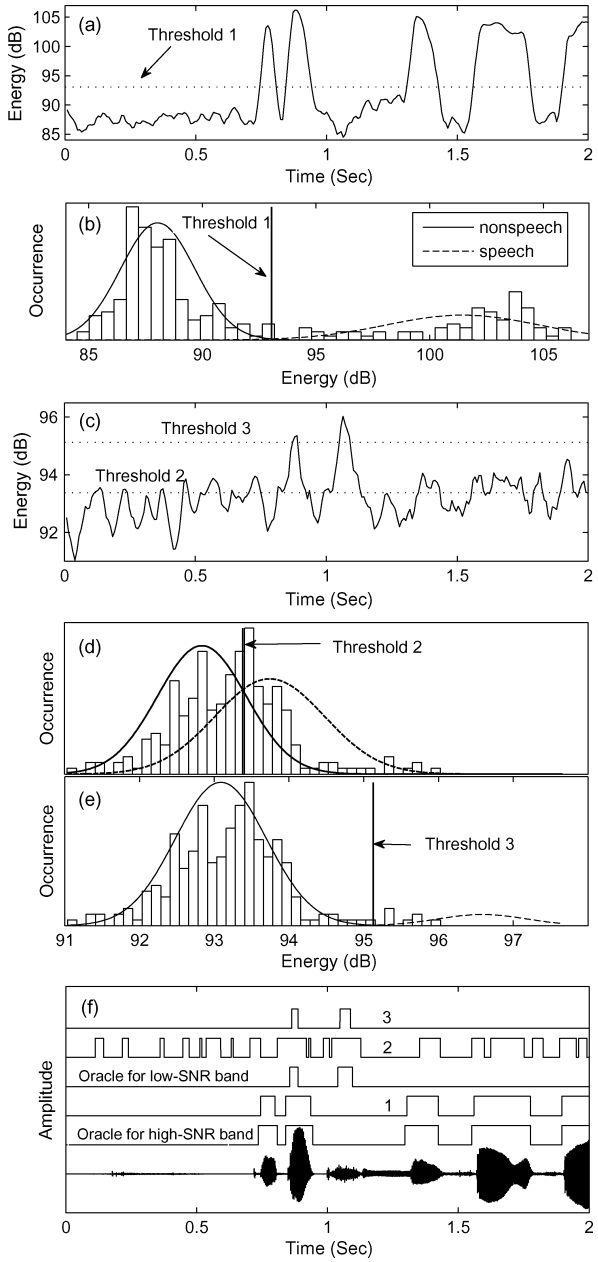

Fig. 2. Demonstration of the offline fitting. (a) A logarithmic energy envelope at a high-SNR subband. (b) The fitting result of the high-SNR subband labeled on its histogram. (c) A logarithmic energy envelope of a low-SNR subband. (d) Fitting result of the low-SNR subband without constraints. (e) Fitting result of the low-SNR subband with constraints. (f) VAD boundaries determined by thresholds 1, 2, and 3.

that miss-detection of speech is very serious without constraints while it is less likely to occur with constraints. From these tests, one can see that the EM algorithm with constraints can deal with both the bimodal and unimodal distributions.

III. VAD System Implementation

The previous section shows speech/nonspeech discrimination in a single subband. The decision based upon one subband cannot yield a reliable VAD. We combine all subband decisions in a voting procedure. It is worthwhile mentioning the reason for treating each subband individually with a univariate GMM instead of jointly with a multivariate GMM. The major reason rests on the unsupervised learning, where the clustered models should be identified as speech/nonspeech according

```
For the first M+1 frames                                   } Initialization
    For each Mel subband
        Extract a logarithmic energy envelope.
        Establish a GMM by EM with constraints.
        Determine the threshold from GMM using Eq. 4.
        Tune the threshold using Eq. 24.
        Classify of M+1 samples as speech/nonspeech.
    End
    Summarize all sub-bands' classification by voting.
    Discriminate speech/nonspeech by hangover scheme.      }
End
For each new coming frame at time k+1                      } Updating
    Do FFT and calculate x_{k+1} at each Mel subband.
    For x_{k+1} at each sub-band
        Calculate p(z | x_{k+1}, λ_k) with Eq. 12.
        Update the weight coefficients with Eq. 15.
        Constrain the weight coefficients with Eq. 22.
        Update the means with Eq. 17.
        Constrain the means with Eq. 20.
        Update the variances with Eq. 19.
        Constrain the variances with Eq. 21.
        Determine the threshold from GMM using Eq. 4.
        Tune the threshold using Eq. 24.
        Determine x_{k+1} as speech/nonspeech.
    End
    Summarize all sub-band results by voting.
    Discriminate the k+1 frame by hangover scheme.
End
```

Fig. 3. Process of VAD decision.

to their statistical characteristics. With the univariate GMM, the cluster identification can be easily done by comparing the scalar value of the two cluster means. If each frame were taken as a decision unit by using the multivariate GMM, the voting procedure would be unnecessary. In such way, however, there is no guarantee that all elements of one cluster mean are greater than those of another one. As a result, the clusters are difficult to be identified based on the mean vector. For this reason, the approach of "univariate GMM + voting procedure" is more appropriate than the multivariate GMM in this algorithm.

The detail of implementing the proposed VAD is described in Fig. 3, where the EM algorithm with constraints is illustrated in Fig. 4. Several points need to be considered for practical applications. The first point is to extract the logarithmic energy envelope of a subband. The noisy speech signal is chopped into frames by a Hanning window. Then, it is transformed into frequency domain with fast Fourier transform (FFT), and grouped into N mel-scale subbands. For the ℓth subband, its logarithmic energy is calculated as follows:

$$\bar{x}_k = 10\log_{10}\left[\frac{1}{f_{\ell+1}-f_\ell}\sum_{j=f_\ell}^{f_{\ell+1}-1}|Y_{k,j}|^2\right] \tag{23}$$

where $Y_{k,j}$ is the jth DFT coefficient of the kth frame, and f_ℓ is the frequency bin index corresponding to the ℓth mel scale, $\ell = 0, 1, \cdots, N$. Finally, the sequence $\{\bar{x}_k | k = 0, 1, 2, \cdots\}$ is smoothed by using a five-point medium filter to form the envelope $\{x_k | k = 0, 1, 2, \cdots\}$.

The second point is to tune the subband threshold. The general criterion of classification is to minimize the classification error, but, to prevent speech leakage, VADs in many speech systems emphasize more on the accuracy of speech detection. For a given binary classifier, it is well known that the detection accuracy of one class can be improved

```
Initialize GMM by using unsupervised clustering.
While GMM likelihood is increasing
    If w_{0,1} < ε
        w_{0,1} = ε and w_{0,0} = 1 − ε.
        Terminate the iteration.
    End
    Calculate p(z | x_k, λ'_0) for all z & x_k with Eq. 8.
    Calculate new weights with Eq. 5.
    Calculate the new means with Eq. 6.
    Constrain the speech mean with Eq. 20.
    Calculate the new variances with Eq. 7.
    Constrain the speech variance with Eq. 21.
    λ'_0 = λ̄_0.
End
λ_0 = λ'_0.
```

Fig. 4. EM algorithm with constraints.

```
For the k-th frame
    If votes exceed the voting threshold
        speech_flag(k) = 1.
        bc = bc + 1.
        If bc > b_th
            h = h_cnt.
        End
    Else
        bc = 0.
        h = h − 1.
        If h ≤ 0
            speech_flag(k) = 0.
            h = 0.
        Else
            speech_flag(k) = 1.
        End
    End
End
```

Fig. 5. Hangover scheme.

TABLE I
VALUES OF PARAMETERS USED IN THE IMPLEMENTATION OF THE PROPOSED ALGORITHM, FOR A SAMPLING RATE OF 8 kHz

$\alpha = 0.99$	$\delta = 3.5$	$\epsilon = 0.03$	$M = 60$
$bc = 4$	$h_{cnt} = 5$	N=8	
frame length: 20 ms		frame shift: 10 ms	

by sacrificing that of the other one. Here, the threshold is lowered to improve the speech accuracy at the cost of decreasing the nonspeech accuracy:

$$\grave{\theta}_k = \gamma(\theta_k - \mu_{k,0}) + \mu_{k,0} \tag{24}$$

where $0 < \gamma \leq 1$. At each subband, the discrimination is conducted based on the tuned threshold $\grave{\theta}_k$.

The third point is the hangover scheme used to prevent speech leakage in further. The hangover scheme does this by reducing the risk of a low-energy portion of speech signal being falsely rejected. It smooths the decision result just similar to that used in the AMR2 standard [3]. This scheme is shown in Fig. 5, where bc is the burst counter, h for the hangover counter, b_{th} for the burst threshold, and h_{cnt} for the hangover length.

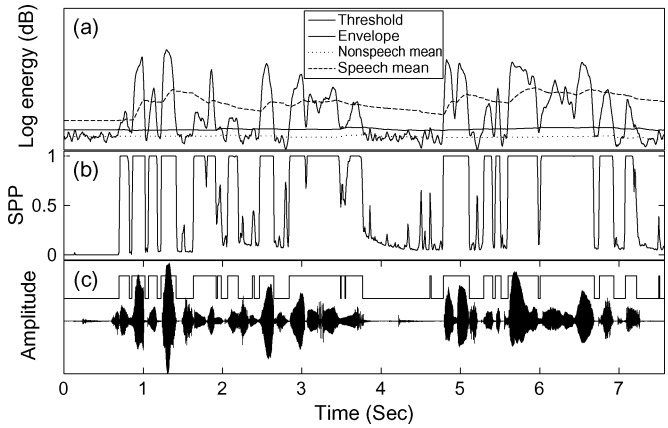

Fig. 6. Speech/nonspeech discrimination at one subband. (a) A logarithmic energy envelope. (b) Speech presence probability. (c) VAD boundary.

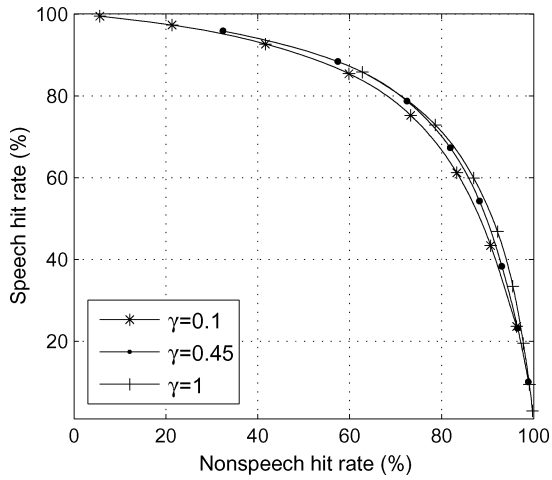

Fig. 7. ROC curves given various γ.

The last point is to apply these constraints in maximum likelihood estimation. Kuhn–Tucker necessary conditions seem to be theoretically sound to constrained maximization, but it is too complex to guarantee its reliability, and it will result in a heavy computation load. Compared with it, the proposed way is more cost-effective and reliable.

The used parameters determined by preliminary experiments are listed in Table I. We find out that, in all parameters, the coefficient γ tuning subband thresholds is the most sensitive to performance. The subband number N influences the VAD's performance and computational complexity. They will be investigated in the following section by experiments. The initialization window length M has an effect on the real-time capability. A small M is beneficial to this capability. For some utterances beginning with speech, however, the small M may make noise signal unavailable for initialization so that the noise model is initialized invalidly by speech samples. For this reason, M should be large enough to guarantee some nonspeech samples to be used for initialization. Combining these considerations, $M = 60$ is appropriate for practical applications.

The key point of the proposed algorithm is the classification at each subband. Fig. 6 demonstrates discrimination at a subband. The hangover scheme is not performed to better show the fitting process. Note that, in this demonstration, there are no speech samples to be available for initialization, where all samples in the initialization window are used to estimate the nonspeech component. The initial speech component of the GMM is virtual. With the coming speech samples, this

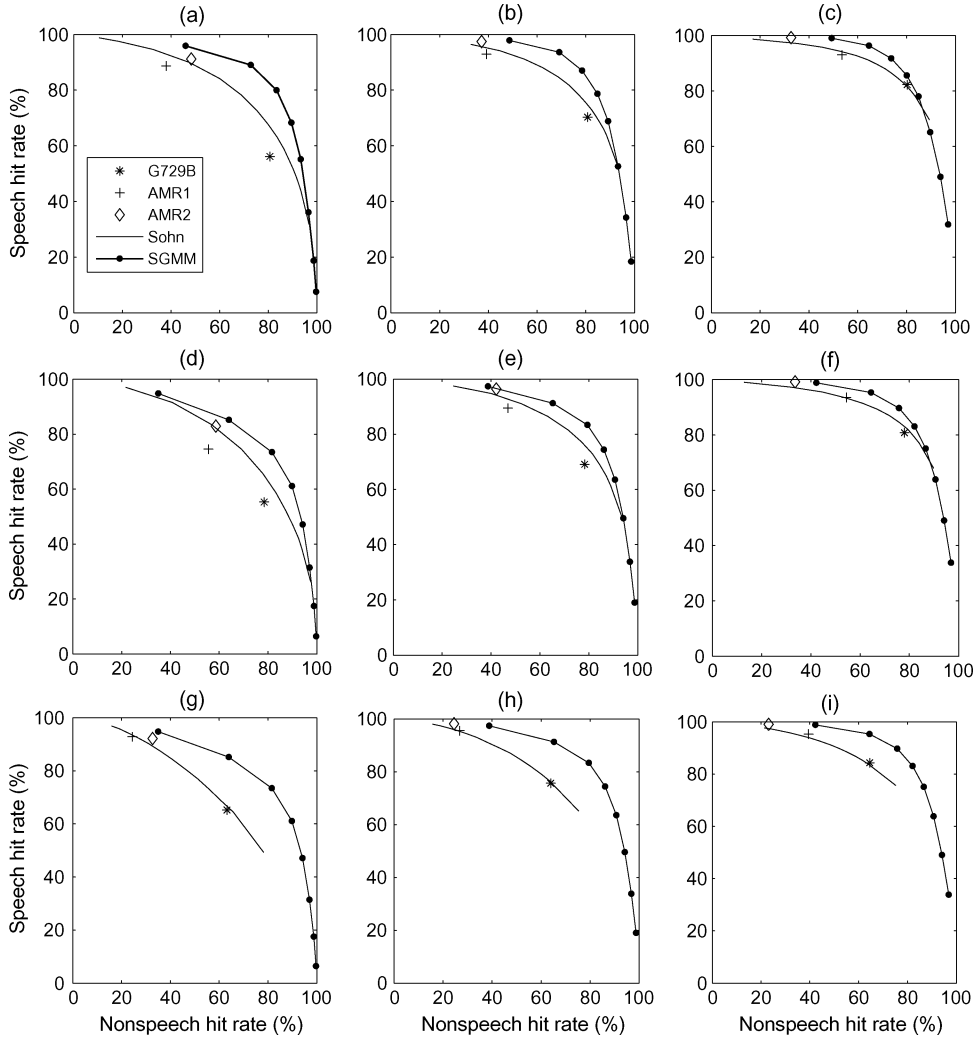

Fig. 9. ROC curves under different noises (columns) and SNRs (rows). (a) 0-dB white noise. (b) 5-dB white noise. (c) 10-dB white noise. (d) 0-dB F16 cockpit noise. (e) 5-dB F16 cockpit noise. (f) 10-dB F16 cockpit noise. (g) 0-dB babble noise. (h) 5-dB babble noise. (i) 10-dB babble noise.

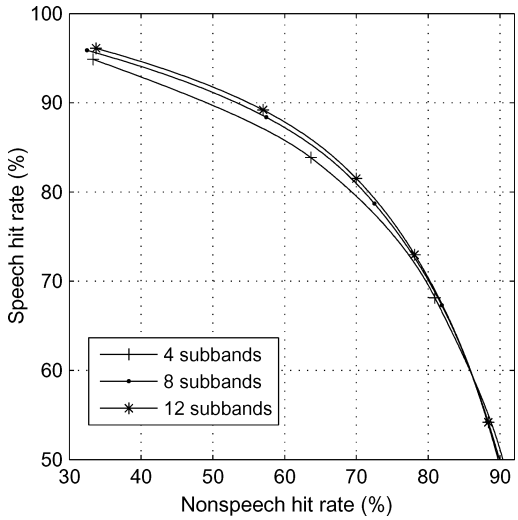

Fig. 8. ROC curves given various number of subbands.

virtual component evolves into a real one. This example gives a demonstration of how the speech and nonspeech means, threshold $(\hat{\theta}_k)$, and SPP vary with the coming samples.

IV. Evaluation

In this section, we evaluate the proposed VAD's performance. It is compared with other leading VADs on the TIMIT database [32]. The proposed VAD is referred to as *SGMM* in the following sections.

A. Experimental Conditions

As a large-scale data set is helpful to give a convincing evaluation of this VAD, we use the TIMIT TEST corpus, consisting of 1680 utterances from 168 individual speakers. The data set encompasses all phonemes and eight different dialects of English. The whole set is hand labeled from phone transcriptions. We connect every two sentences into a longer utterance. Three typical noises, namely white, F16 cockpit and babble noises, from the NOISEX-92 database [33] are artificially added to the test set at variant SNR conditions.

In order to gain a comparative analysis of the SGMM performance, several modern VAD algorithms are also evaluated. These algorithms are the two ETSI AMR VADs options 1 and 2 [3] (denoted, respectively, as AMR1 and AMR2), the ITU G.729 Annex B VAD [8] (referred to as G729), and a soft VAD proposed by Sohn [15] (denoted as Sohn). The implementations of AMR and G.729 are taken from the authors' C implementations, respectively [34], [35]. As the sampling rate of the AMR and G729B is 8000 Hz, all the data is resampled to 8000 Hz for a fair comparison.

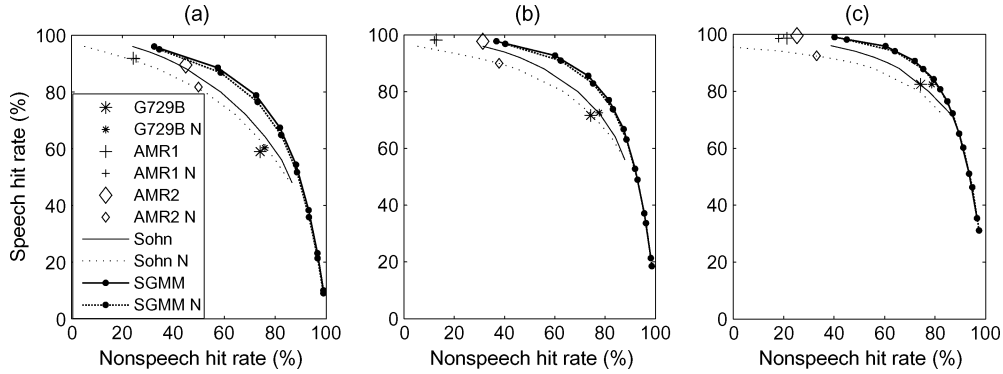

Fig. 10. ROC curves of the data set satisfying the assumption of "nonspeech beginning" versus that unsatisfying this assumption. (a) 0-dB noise. (b) 5-dB noise. (c) 10-dB noise.

In our experiments, the detection performance is assessed in terms of the speech hit rate (HR1) (i.e., the ratio of the correctly detected speech frames to all speech frames) and nonspeech hit rate (HR0) (i.e., the ratio of correctly detected nonspeech frames to all nonspeech frames). A receiver operating characteristics (ROC) curve gives a full description of the relationship between HR0 and HR1. The SGMM ROC curves are obtained by tuning the voting threshold.

B. Determination of Parameter γ and N

Before showing comparative results, the selection of the appropriate γ is considered. Fig. 7 shows the influence of the coefficient γ, where the ROC curves are obtained from all 0-dB noisy utterances. First, we set γ as 1, where $\grave{\theta}_k = \theta_k$ and the classification error of each subband is minimal. Then, we gradually decrease γ to 0.1. One can see that, the performance of SGMM becomes worse because the classifier error of each subband is enlarged. However, the maximal HR1 corresponding to the lowest voting threshold "1" becomes larger and larger. At the end, γ is determined as a tradeoff between the maximal HR1 and SGMM performance. From this experiment, we find out that $\gamma = 0.45$ achieves an optimal tradeoff, where the maximal HR1 approaches to 95.9% while the performance decreases a little bit.

With the same method, the influence of the subband number N on the proposed VAD's performance is investigated, as shown in Fig. 8. N is tuned from 4 to 12 while other parameters are kept invariant. Increasing the number of subbands is beneficial to improve the VAD performance since the frequency resolution is increased. When $N > 8$, the VAD performance is improved a little bit. When $N > 12$, no additional improvements are reported. From this experiment, one can see that $N = 8$ yields the best tradeoff between computational cost and performance.

C. Speech/Nonspeech Discrimination Experiments

We design two experiments to evaluate the discrimination capability of the SGMM VAD. The first experiment is to compare the VADs' performance at general conditions, where the data set satisfies the assumption of "nonspeech beginning" Fig. 9 shows the ROC curves at variant noisy conditions. The eight working points in each SGMM ROC curve, respectively, correspond to the voting thresholds from 1 to 8.

The acoustic feature and the statistical model of the Sohn VAD are the most similar to that of SGMM. Its feature is the spectral amplitude and its model is the Gaussian one, but the Sohn VAD is a typical semi-supervised one, which employs the assumption of "nonspeech beginning" to initialize the nonspeech model and the decision-directed learning to update speech and nonspeech models. The semi-supervised learning has the inherent advantage in nonspeech model initialization under the assumption of "nonspeech beginning." Therefore, from a pure theoretical point of view, the semi-supervised approach of model construction would be better than the unsupervised one. However, the SGMM based on the unsupervised learning framework still runs better than Sohn VAD, as shown in Fig. 9. This result shows that the SGMM VAD does better than typical semi-supervised VADs in the whole process of model construction. The standardized VADs such as AMR and G729 extract several acoustical features to fully utilize the property of speech signal for speech/nonspeech discrimination. They combine these features together by fuzzy rules. Comparing with these standardized VADs, the proposed VAD shows promising performance.

The purpose of the second experiment is to evaluate the influence of the "nonspeech beginning" assumption on the VADs' performance. We compare their performance on the data sets of satisfying and unsatisfying the assumption. The latter is obtained by cutting off the first 0.6-s signal of each long utterance. Thus, some utterances will begin with speech signal. The experiment result is shown in Fig. 10, where the symbol "N" denotes the ROC curve of the data set without this assumption. As the Sohn VAD is a semi-supervised one, the assumption is crucial to it. AMR2 VAD also utilizes the semi-supervised way to track background noise. So, the performance of AMR2 and Sohn VADs are affected by this assumption. In contrast, since SGMM VAD is an unsupervised one, its performance changes a little bit. G.729 and AMR1 VADs utilize neither of the semi-supervised and unsupervised learning. So, their performance is not affected by this assumption.

D. Informal Evaluation to Speech Presence Probability

In addition to discriminating frames, the SGMM VAD can provide the SPP in the time-frequency domain. At each subband, the SPP sequence $\{p(z = 1|x_k, \lambda_k)|k = 0, 1, 2, \cdots\}$ describes the speech activity in a soft manner. The SPP is informally evaluated by comparing the time–frequency SPP with the noisy spectrogram. Fig. 11(a) shows the spectra of an utterance corrupted by white noise at SNR 0 dB, and the color gray of Fig. 11(b) denotes SPP. For the sake of comparison, each subband consists of only one frequency bin. From this comparison, one can see the speech spectral structure is described clearly by the time–frequency SPP.

V. Discussions and Conclusion

In this correspondence, we present a statistical framework based on unsupervised learning to model the speech and nonspeech distributions in frequency domain. It considers not only the distribution relationships between speech and nonspeech, but also the *a priori* distributions of speech and nonspeech signals. This framework outperforms conventional statistical models because of its advantages in both the initialization process and the sequential process. In initialization, both the speech and nonspeech models are simultaneously constructed based on the criterion of maximum likelihood. This initialization does not rely on the assumption of "nonspeech beginning." Whether or not speech

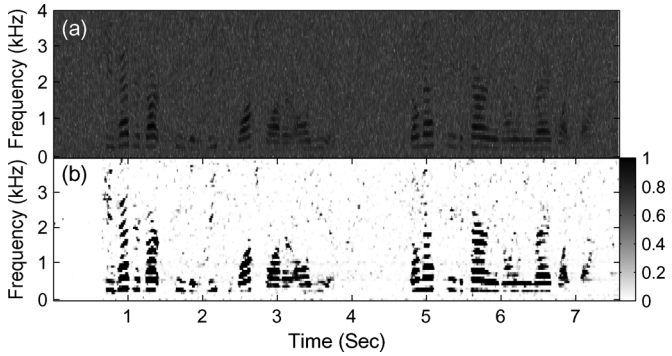

Fig. 11. Informal evaluation of speech presence probability. (a) Noisy speech spectra. (b) Speech presence probability.

signal is present in the utterance beginning, the proposed model can be correctly initialized. Thus, this VAD is more practical than conventional ones.

In the updating process, the advantage is shown in two aspects. One aspect is the soft manner of updating statistical models. The "soft" degree is controlled by the SPP. In contrast, most VADs utilize a "hard" updating manner. The speech/nonspeech model is either updated or not. The soft updating method of SGMM VAD is more reasonable than that of the conventional ones. The other aspect concerns the decision feedback. Due to the speech sparsity in the frequency domain, not all frequency components of a speech frame are occupied by speech signal. Hence, it is better to describe the speech presence of each component, and to feed them back respectively. However, most VADs only gives speech presence information in the frame level. The more detailed information in each frequency component is absent. Hence, the nonspeech information in the speech frames is unavailable for updating models. On the contrary, as the SGMM VAD can provide the frequency domain SPP, the nonspeech information in speech frames can be employed to update models. Therefore, this statistical framework can more accurately model signals than conventional ones. This is another reason for using the univariate GMM instead of the multivariate GMM. Especially in noise reduction applications, this advantage of the proposed VAD is obvious. Due to these advantages, the proposed VAD performs better than typical semi-supervised VADs even when the assumption of "nonspeech beginning" is satisfied. The experiments confirm its superiority.

The proposed algorithm uses only a simple acoustic feature for classification. In fact, other features that satisfy the bimodal distribution of Fig. 1(a) can also be applied to this unsupervised framework. By using advanced features, this VAD is expected to be further improved by fully employing speech properties. The proposed VAD is just one application of the unsupervised learning framework. This framework can be further considered for other applications such as speech enhancement, noise power tracking and subband SNR estimation.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and Dr. X. Lu for their helpful comments.

REFERENCES

[1] R. Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, no. 3, pp. 245–254, Apr. 1995.

[2] F. Lamel, R. Rabiner, E. Rosenberg, and G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 4, pp. 777–785, Aug. 1981.

[3] *Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels*, ETS1 EN 301 708 Rec., ETSI, 1999.

[4] *Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, ETSI ES 202 050 Rec., ETSI, 2002.

[5] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator," *IEEE Signal Process. Lett.*, vol. 8, no. 1, pp. 10–12, Jan. 2001.

[6] J. Ramírez and J. C. Segura *et al.*, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3, pp. 271–287, 2004.

[7] J. Ramírez and J. C. Segura *et al.*, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1119–1129, Nov. 2005.

[8] *Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear rrediction. Annex B: A silence compression scheme for G.729 optimized for terminals conforming to recommend*, V.70, ITU, 1996.

[9] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Elect. Eng. 1992*, pp. 377–380, 1992.

[10] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," in *Proc. SAPA'06*, Pittsburgh, PA, 2006, pp. 65–70.

[11] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 109–118, Feb. 2002.

[12] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 412–423, Mar. 2006.

[13] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, Mar. 2001.

[14] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, vol. 1, pp. 365–368.

[15] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[16] Y. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, no. 10, pp. 276–279, Oct. 2001.

[17] J. Górriz, J. Ramírez, E. Lang, and C. Puntonet, "Jointly Gaussian PDF-based likelihood ratio test for voice activity detection," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1565–1578, Nov. 2008.

[18] J. Ramírez, J. C. Segura, M. C. Benítez, Á. de la Torre, and A. Rubio, "A new Kullback–Leibler VAD for speech recognition in noise," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 666–669, Feb. 2004.

[19] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian–Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498–505, Sep. 2003.

[20] J. Ramírez and J. C. Segura, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, Oct. 2005.

[21] J. Chang, N. Kim, and S. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[22] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 146–157, Mar. 2002.

[23] R. Tahmasbi and S. Rezaei, "Change point detection in GARCH models for voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1038–1046, Jul. 2008.

[24] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006, ch. 1, pp. 1–12.

[25] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Commun.*, vol. 34, pp. 141–158, 2001.

[26] Y. Shi, F. K. Soong, and J. L. Zhou, "Auto-segmentation based partitioning and clustering approach to robust end pointing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. 793–796.

[27] D. V. Campernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Comput. Speech Lang.*, vol. 3, pp. 151–168, 1989.

[28] R. Xu and D. Wunsch, "Survey of clustering algorithm," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[29] O. Arandjelovic and R. Cipolla, "Incremental learning of temporally-coherent Gaussian Mixture Models," in *Proc. BMVC*, 2005.

[30] Q. Huo and C. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 2, pp. 161–172, Mar. 1997.

[31] V. Krishnamurthy and J. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback–Leibler information measure," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2557–2573, Aug. 1993.

[32] J. S. Garofolo, Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database Nat. Inst. Standards Technol. (NIST). Gaithersburg, MD, prototype as of Dec. 1988.

[33] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[34] Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi Rate (AMR) Speech; ANSI-C Code for AMR Speech Codec, 1998.

[35] ITU, Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic Code-Excited Linear Prediction. Annex I: Reference Fixed-Point Implementation for Integrating G.729 CS-ACELP Speech Coding Main Body With Annexes B, D and E, Int. Telecommun. Union, 2000.