Voice Activity Detection Based on a Sequential Gaussian Mixture Model

Dongwen Ying*, Junfeng Li*, Qiang Fu*, Yonghong Yan*, and Jianwu Dang†
* The Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences
E-mail: {yingdongwen, lijunfeng, qfu, yyan}@hccl.ioa.ac.cn
† Tianjin University, School of Computer Science and Technology
E-mail: jdang.china@gmail.com

***Abstract*—Voice activity detection (VAD) is a basic component of noise reduction algorithms. In this paper, we propose a voice activity detector based on a sequential Gaussian Mixture Model (SGMM) in log-spectral domain. This model comprises two Gaussian components, which respectively describe the speech and nonspeech log-power distributions. The initial distributions are firstly established by EM algorithm, and then sequentially updated in an on-line manner. From the SGMM, a self-regulatory threshold for discrimination is derived at each subband. 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.**

I. Introduction

Voice activity detection (VAD) is an indispensable component for noise reduction. It plays an important role in variant speech communication systems. Generally speaking, VADs consist of acoustic features and discrimination models.

Early algorithms paid more attention to robust acoustic features to distinguish speech/nonspeech. Three types of acoustics features are widely used. The energy-based features were the most popular one. The SNR (Signal to Noise Ratio) was usually taken as an energy cue for discrimination [3], [4]. The second popular feature was the quasi-periodicity of voiced speech [3], [5]–[7]. 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], [8] or SNR [9].

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 [10]–[12]. 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. These VADs based on this assumption 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 [14], [15], 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 [13], [16], 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. However, there are still two essential problems to be solved in these VADs. Firstly, a mechanism of incrementally updating models is absent. They are not able to run in an on-line manner because clustering algorithms are usually conducted in an off-line manner. So, these VADs can not be applied to some real-time systems. Secondly, 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 [**?**], 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.

II. Sequential Gaussian Mixture Model

This algorithm is concerned on a single band in the following subsections, where the band index r is omitted for the interests of brevity.

A. Modeling Power Sequence with GMM

Without loss of generality, we first consider a high-SNR frequency band with speech and nonspeech signals. It is assumed that both speech and nonspeech power obeys the Gaussian distribution. This model is described by the following equations.

Let x_k denote the log power 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

$$p(x_k|\lambda) = \sum_z p(x_k, z|\lambda) = \sum_z p(x_k|z, \lambda)p(z) \quad (1)$$

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.

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

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 difference of the two means represents the posteriori SNR because μ_1 and μ_0 are respectively the averaged logarithmic energy of 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 probability density function (PDF) is given by

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

The parameter set λ is estimated by maximizing this likelihood function. μ_0 is the estimate of the noise level of the $M + 1$ frames.

B. Sequential Estimation of GMM Parameters

How to sequentially estimate the GMM parameter set is a crucial point for this algorithm. Our approach of estimating it comprises an initialization process and an updating process. The initial GMM is firstly established by the typical EM algorithm, and then sequentially estimated at each time instant. The parameter set at time k is denoted as $\lambda_k \triangleq \{\boldsymbol{\mu}_k, \boldsymbol{\kappa}_k, \boldsymbol{w}_k\}$. λ_0 is the initial one estimated from the first P samples in an off-line manner. In this section, we give the details of sequential estimation.

The following are the basic way to sequentially update GMM by utilizing every K frames. Suppose λ_k is known at the time $k + 1$, the parameters in λ_{k+1} are derived by the following sequential equations:

$$w_{k+1,z} = \frac{\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) + p(z|x_{k+1}, \lambda_k)}{K + 1} \quad (4)$$

$$\mu_{k+1,z} = \frac{\sum_{j=k-K+1}^{k} x_j p(z|x_j, \lambda_k) + x_{k+1} p(z|x_{k+1}, \lambda_k)}{\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) + p(z|x_{k+1}, \lambda_k)} \quad (5)$$

$$p(z|x_k, \lambda_k) = \frac{w_{k,z} p(x_k|z, \lambda_k)}{\sum_z w_{k,z} p(x_k|z, \lambda_k)} \quad (7)$$

where $p(z = 1|x_k, \lambda_k)$ denotes SPP. The weight, mean, and variance can be regarded as the zero, first, and second order moments of speech/nonspeech logarithmic energy, respectively. This updating method is not so desirable. On one hand, $\{p(x_j|z, \lambda_k)$ for all j 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 such considerations, we propose a novel approach of sequentially updating GMM. Suppose that the GMM varies with time slowly, $\lambda_k \approx \lambda_{k-1}$ at time k. Accordingly, we have the relationship $\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) \approx \sum_{j=k-K+1}^{k} p(z|x_j, \lambda_{k-1})$. According to Eq. 4, the summation is approximated by the zero-order moment, $\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_{k-1}) \approx K w_{k,z}$. Combining these relationships, we finally have the following equation

$$\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) \approx K w_{k,z}. \quad (8)$$

Substituting Eq.8 into Eq. 4, we obtain

$$w_{k+1,z} = \frac{K w_{k,z} + p(z|x_{k+1}, \lambda_k)}{K + 1}. \quad (9)$$

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

$$w_{k+1,z} = \alpha w_{k,z} + (1 - \alpha) p(z|x_{k+1}, \lambda_k) \quad (10)$$

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 using Eq. 7.

With the same principle, the summation item in Eq.5 can be approximated by the 1st-order moment

$$\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) x_j \approx K w_{k,z} \mu_{k,z}. \quad (11)$$

Substituting Eq.11 into Eq. 5, we obtain

$$\mu_{k+1,z} = \frac{\alpha w_{k,z} \mu_{k,z} + (1 - \alpha) p(z|x_{k+1}, \lambda_k) x_{k+1}}{w_{k+1,z}}. \quad (12)$$

Accordingly, the summation item in Eq. 6 is approximated by the 2nd-order moment

$$\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k)(x_j - \mu_{k+1,z})^2 \approx K w_{k,z} \kappa_{k,z}. \quad (13)$$

Substituting Eq.13 into Eq. 6, we obtain

$$\kappa_{k+1,z} = \frac{\alpha w_{k,z} \kappa_{k,z} + (1 - \alpha) p(z|x_{k+1}, \lambda_k)(x_{k+1} - \mu_{k+1,z})^2}{w_{k+1,z}}. \quad (14)$$

The sequential estimate consists of Eqs. 7, 10, 12, and 14, where the SPP is derived from GMM, and then feeded back to

$$\kappa_{k+1,z} = \frac{\sum_{j=k-K+1}^{k}(x_j - \mu_{k+1,z})^2 p(z|x_j, \lambda_k) + (x_{k+1} - \mu_{k+1,z})^2 p(z|x_{k+1}, \lambda_k)}{\sum_{j=k-K+1}^{k} p(z|x_j, \lambda_k) + p(z|x_{k+1}, \lambda_k)} \tag{6}$$

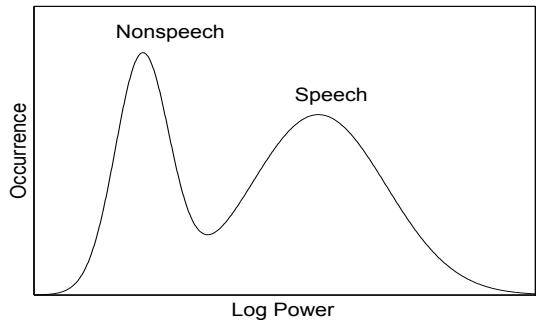

Fig. 1. Schematic illustration of the log-power distribution of noisy speech with high SNR.

update GMM. By these equations, the model λ_{k+1} is estimated from λ_k and x_{k+1}. In this iteration, the earlier frames are forgotten with time going, and the later frames plays a more important role. $\mu_{k+1,0}$ is the noise estimate given λ_k and x_{k+1}.

C. Constraints to the GMM

The above model is only applicable to high-SNR bands with both speech and nonspeech signal. But at low-SNR bands, the speech signal may be unobvious or absent, and so it is difficult to model the power with the two-component GMM. For this reason, some constraints must be introduced to deal with the low-SNR bands.

Several constraints come from the special relationships between speech and nonspeech distributions [16]. Fig. 1 schematically illustrates the typical log-power histogram of a noisy band, which consists of a speech and nonspeech distributions. Assuming that the noise signal is usually more stationary than speech signal, the variance of nonspeech power is smaller than that of the speech. So, there exists a sharp peak corresponding to nonspeech μ_0 in the histogram; while the other flat peak corresponds to speech μ_1. Since "speech" denotes the superposition of noise and clean speech signals in this paper, the average of nonspeech power is smaller than that of speech. The mean difference $\mu_1 - \mu_0$ represents the posterior SNR of this band. These relationships shown in Fig. 1 are shaped by three constraints.

Firstly, the relationship between μ_1 and μ_0 is shaped by the following constraint

$$\mu_{k,1} = \max\{\mu_{k,1}, \mu_{k,0} + \delta\}, \tag{15}$$

where $\delta > 0$. This constraint makes the averaged posterior SNR no less than δ. If the actual posterior SNR is less than δ, this constraint will set the speech distribution center to be $\mu_{k,0} + \delta$, which is greater than the actual one. As a result, the speech likelihood $b(x_k|s_k = 1, \lambda_k)$ is decreased for most samples. Eventually, some weak speech components dominated by noise signal will be taken as nonspeech to be used for noise estimation. Here, δ makes a tradeoff between weak speech spectral components and strong nonspeech spectral components. A large δ will result in more weak speech components to be used for noise estimate. In contrast, a small δ will make the strong nonspeech components to be unavailable for estimation. So, a large δ is beneficial to high noise environments, and a small δ to low noise environments. According to our preliminary experiments, we find out that $\delta = 5$ can achieve a good tradeoff in general noisy conditions.

It should be noted that, in the initialization and updating process, when the constraint of Eq. 15 decreases a lot the likelihood $b(x_k|z = 1, \lambda_k)$, $p(z = 0|x_k, \lambda_k) \gg p(z = 1|x_k, \lambda_k)$. With several iterations, $p(z = 1|x_k, \lambda_k)$ will approach to zero. So, the denominators will be zero when solving the speech mean and variance, which results in the failure of this algorithm. To prevent this failure, we introduce the second constraint

$$\begin{aligned} w_{k,1} &= \max\{w_{k,1}, \epsilon\} \\ w_{k,0} &= 1 - w_{k,1}. \end{aligned} \tag{16}$$

where ϵ is near and greater than zero. When the condition in Eq. 16 is not satisfied, the iteration in EM algorithm will terminate. This constraint is a slave to the one in Eq. 15.

Thirdly, according to the variance relationship, the following constraint is introduced.

$$\kappa_{k,1} = \max\{\kappa_{k,0}, \kappa_{k,1}\} \tag{17}$$

With these three constraints, SGMM can run under not only high-SNR frequency bands, but also low-SNR bands. Even when speech signal is absent, the mean and variance of the speech component are respectively $\mu_{k,0} + \delta$ and $\kappa_{k,0}$. At that situation, the speech distribution is not estimated from the real data. It is a virtual one that is constructed according to the nonspeech component. From the point view of clustering, all the data is clustered into the nonspeech class in low-SNR or speech-absent situations.

III. IMPLEMENTATION OF THE ALGORITHM

The above section describes noise estimation in one frequency band. This process is conducted in parallel at each band. Before estimation, the log power is smoothed by a five-point median filter. The constraints are applied after each parameter is updated. For example, in the updating process, Eq.10 is followed by Eq.16, Eq.12 by Eq.15, and Eq.14 by Eq.17. In initialization, these constraints are utilized in the same way. Actually, the Kuhn-Tucker necessary condition is the major way of constrained maximization. Compared with it, the proposed way for applying constraints is more cost-effective.

It is worthwhile clarifying that the SGMM can be correctly initialized even if speech samples are not utilized. At that

situation, these constraints will construct a virtual speech component, which can convert into a real one when the SGMM is updated with coming speech samples. For the signal with a sampling rate of 16 kHz, the parameters of this SGMM is set as $\alpha = 0.97, \delta = 5, \epsilon = 0.03, P = 60$. The signal is chopped into frames with a hanning window, 16 ms frame length, 8 ms frame shift.

IV. Evaluation

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 whole set is hand labeled from phone transcriptions. We connect every two sentences into a longer utterance. The white and babble noises from NOISEX-92 database 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 [5] (referred to as G729), and a soft VAD proposed by Sohn [11] (denoted as Sohn). As the sampling rate of the AMR and G729B is 8000 Hz, all the data is resampled to 8000 Hz for a fair comparison.

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). 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.

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. 2 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 Sohn VAD are the most similar to that of SGMM. But the Sohn VAD which employs the assumption of "nonspeech beginning" to initialize the nonspeech model. One can see that the SGMM based on the unsupervised learning framework runs better than Sohn VAD, as shown in Fig. 2. 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 0.6s signal of each long utterance. Thus, some utterances will begin with speech signal. The experiment result is shown in Fig. 3, 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.

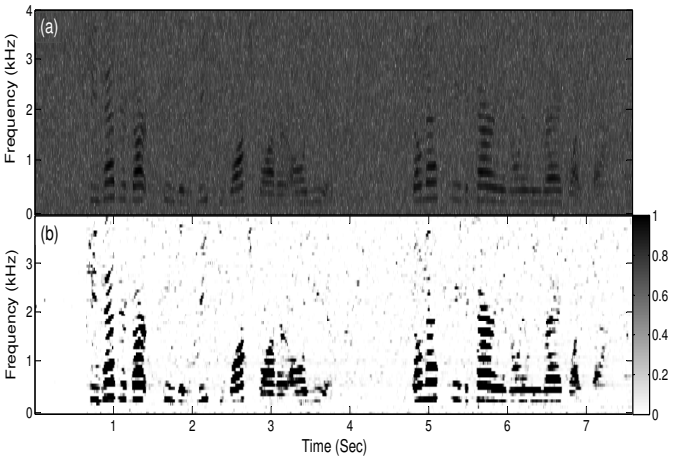

Fig. 4. Informal evaluation of speech presence probability: (a) Noisy speech spectra; (b) Speech presence probability.

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. 4(a) shows the spectra of an utterance corrupted by white noise at SNR 0 dB, and the color gray of Fig. 4(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 Conclusions

In this paper, we present a VAD based on sequential Gaussian mixture model. 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 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

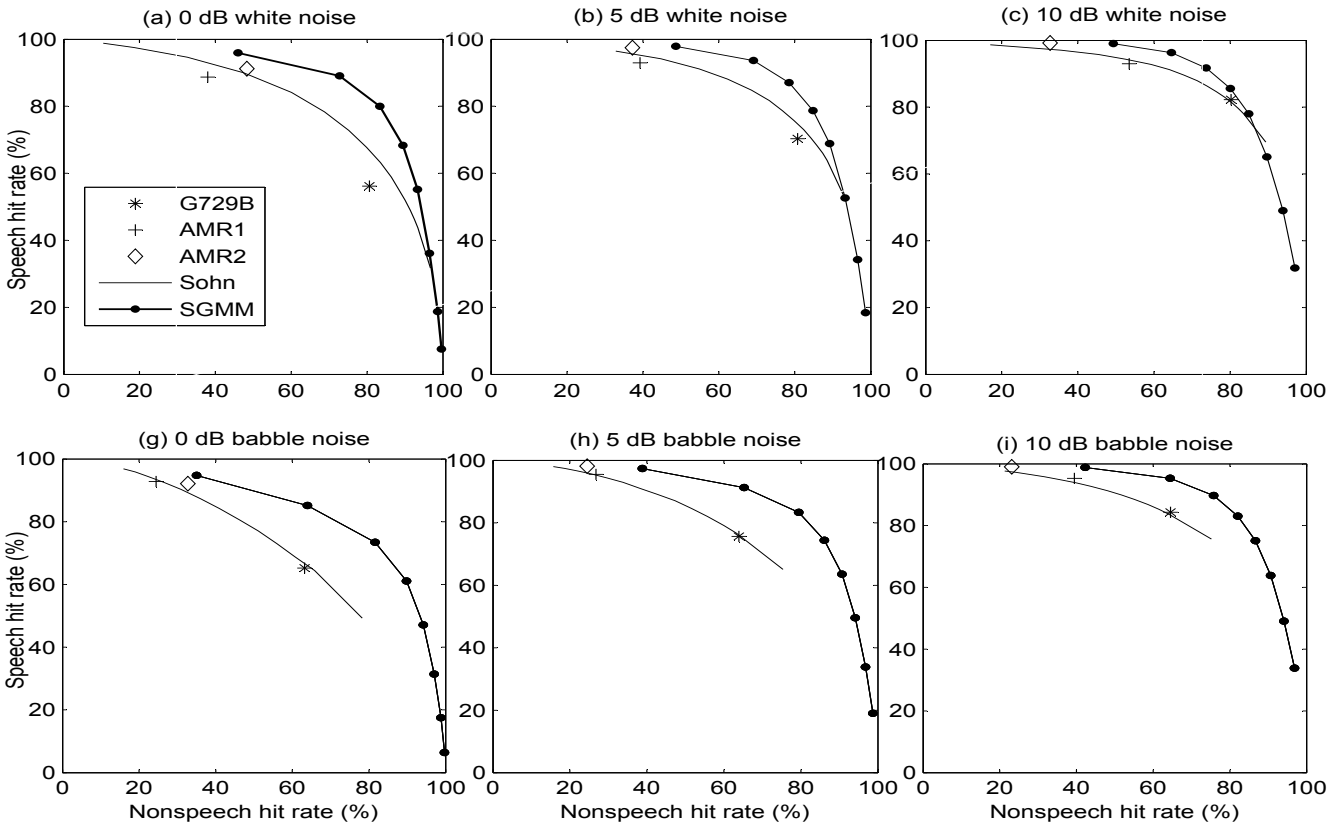

Fig. 2. ROC curves under different noises (columns) and SNRs (rows).

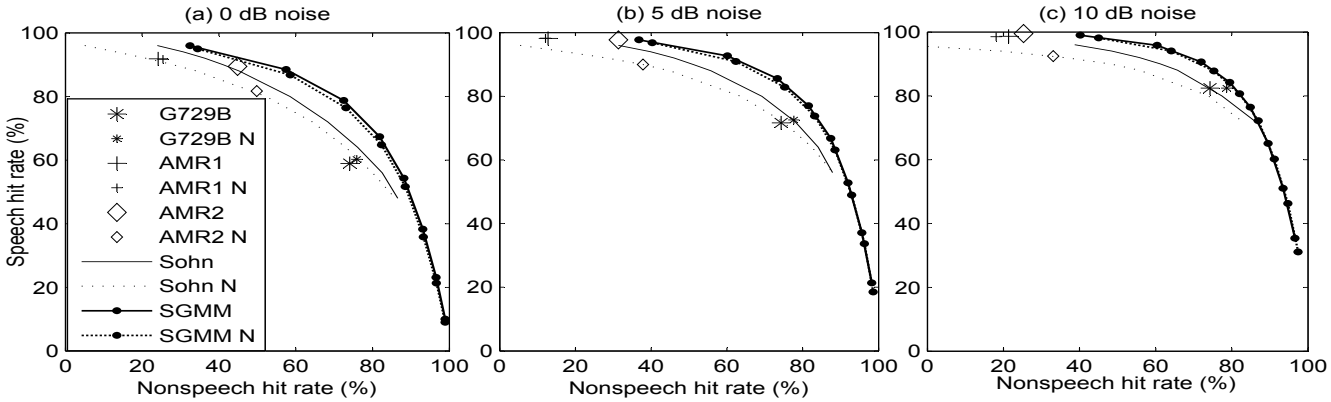

Fig. 3. ROC curves of the data set satisfying the assumption of "nonspeech beginning" vs. that unsatisfying this assumption.

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 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.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275, 11161140319). This work is also supported by National Thousand Talents Program.

REFERENCES

[1] R. Jeannes, 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, April 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. 29, no. 4, pp. 777–785, Aug. 1981.

[3] ETSI, "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," ETS1 EN 301 708 Recommend., 1999.

[4] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," *ETSI ES 202 050 Recommend.* , 2002.

[5] ITU, "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, International Telecommunication Union, 1996.

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

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

[8] M. Marzinzik, 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.

[9] 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, March 2006.

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

[11] 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.

[12] 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.

[13] 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, March 2002.

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

[15] 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.

[16] D. V. Campernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language*, vol. 3, pp. 151–168, 1989.

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