EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

A novel voice activity detection based on phoneme recognition using statistical model

Xulei Bao* and Jie Zhu

Abstract

In this article, a novel voice activity detection (VAD) approach based on phoneme recognition using Gaussian Mixture Model based Hidden Markov Model (HMM/GMM) is proposed. Some sophisticated speech features such as high order statistics (HOS), harmonic structure information and Mel-frequency cepstral coefficients (MFCCs) are employed to represent each speech/non-speech segment. The main idea of this new method is regarding the non-speech as a new phoneme corresponding to the conventional phonemes in mandarin, and all of them are then trained under maximum likelihood principle with Baum-Welch algorithm using GMM/HMM model. The Viterbi decoding algorithm is finally used for searching the maximum likelihood of the observed signals. The proposed method shows a higher speech/non-speech detection accuracy over a wide range of SNR regimes compared with some existing VAD methods. We also propose a different method to demonstrate that the conventional speech enhancement method only with accurate VAD is not effective enough for automatic speech recognition (ASR) at low SNR regimes.

1 Introduction

Voice activity detection (VAD), which is a scheme to detect the presence of speech in the observed signals automatically, plays an important role in speech signal processing [1-4]. It is because that high accurate VAD can reduce bandwidth usage and network traffic in voice over IP (VoIP), and can improve the performance of speech recognition in noisy systems. For example, there is a growing interest in developing useful systems for automatic speech recognition (ASR) in different noisy environments [5,6], and most of these studies are focused on developing more robust VAD systems in order to compensate for the harmful effect of the noise on the speech signal.

Plentiful algorithms have been developed to achieve good performance of VAD in real environments in the last decade. Many of them are based on heuristic rules on several parameters such as linear predictive coding parameters, energy, formant shape, zero crossing rate, autocorrelation, cepstral features and periodicity measures [7-12]. For example, Fukuda et al. [11] replaced the traditional Mel-frequency cepstral coefficients (MFCCs) by the harmonic structure information that made a significant improvement of recognition rate in ASR system. Li et al. [12] combined the high order statistical (HOS) with the low band to full band energy ration (LFER) for efficient speech/non-speech segments.

However, the algorithms based on the speech features with heuristic rules have difficulty in coping with all noises observed in the real world. Recently, the statistical model based VAD approach is considered an attractive approach for noisy speech. Sohn et al. [13] proposed a robust VAD algorithm based on a statistical likelihood ratio test (LRT) involving a single observation vector and a Hidden Markov Model (HMM) based hang-over scheme. Later, Cho et al. [14] improved the study in [13] by a smoothed LRT. Gorriz et al. [15] incorporated contextual information in a multiple observation LRT to overcome the non-stationary noise. In these studies, the estimation error of signal-to-noise ratio (SNR) seriously affects the accuracy of VAD. With respect to this problem, the utilization of suitable statistical models, i.e., Gaussian Mixture Model (GMM) can provide higher accuracy. For example, Fujimoto et al. [16] composed the GMMs of noise and noisy speech by Log-Add composition that showed excellent detection accuracy. Fukuda et al. [11] used a large vocabulary with high order GMMs for discriminating the non-speech from speech that made a significant improvement of recognition rate in ASR system.

* Correspondence: qunzhong@sjtu.edu.cn
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

To obtain more accurate VAD, these methods always choose a large number of the mixtures of GMM and select an experimental threshold. But they are not suitable for some cases. To handle these problems, using the GMM based HMM recognizer for discriminating the non-speech from the speech not only can reduce the number of mixtures but also can improve the accuracy of VAD without the experimental threshold.

In this article, the non-speech is assumed as an additional phoneme (named as *'usp'*) corresponding to the conventional phonemes (such as *'zh'*, *'ang'* et al.) in mandarin. Moreover, the speech features, such as harmonic structure information, HOS, and traditional MFCCs which are combined together to represent the speech, are involved in the maximum likelihood principle with Baum-Welch (BW) algorithm in HMM/GMM hybrid model. In the step of discriminating speech from nonspeech, Viterbi algorithm is employed for searching the maximum likelihood of the observed signals. As a result, our experiments show a higher detection accuracy compared with the existing VAD methods on the same Microsoft Research Asia (MSRA) mandarin speech corpus. A different method is also proposed in this article to show that the conventional noise suppression method is detrimental to the speech quality even giving precise VAD results at low SNR regimes and may cause serious degradation in ASR system.

The article is organized as follows. In Section 2, we first introduce the novel VAD algorithm. And then, a different VAD method based on the recursive phoneme recognition and noise suppression methods is given in Section 3. The detail experiments and simulation results are shown in Section 4. Finally, the discussion and conclusion are drawn in Section 5 and Section 6 respectively.

2 The VAD algorithm

2.1 An overview of the VAD algorithm

As well known, heuristic rules based and statistical model based VAD methods respectively have advantages and disadvantages against different noises. We combine the advantages of these two methods together for making the VAD algorithm more robust. The method proposed in this article is shown in Figure 1. We divide this method into three submodules, such as noise estimation submodule, feature extraction submodule and HMM/GMM based classification submodule.

In our study, the MSRA mandarin speech corpus are employed for training the HMM/GMM hybrid models at different SNR regimes (as SNR = 5 dB, SNR = 10 dB et al.) under maximum likelihood principle with BW algorithm firstly. Then, in the VAD process, the SNR of the noisy speech is estimated by the noise estimation submodule, and the corresponding SNR level of HMM/GMM hybrid model is selected. After that, the speech features such as MFCCs, the harmonic structure information and the HOS are extracted to represent each speech/non-speech segment. Finally, the non-speech segments are distinguished from the speech segments by the phoneme recognition using the trained HMM/GMM hybrid model.

Note that, in this article, the typical noise estimation method named minima controlled recursive averaging (MCRA) is employed for the realization of noise estimation submodule, referring to [17] for details.

2.2 Feature extraction

Different features have their own advantages in ASR system. And it is impossible to use one feature to cope with all the noisy environments. Combining some features together for discriminating the speech from nonspeech is a popular strategy in recent years. In this article, three useful features such as harmonic structure information, HOS and MFCCs are combined together to represent the speech signals, since harmonic structure information is robust to high-pitched sounds, HOS is robust to the Gaussian and Gaussian-like noise, and MFCCs are the important features in phoneme recognizer.

2.2.1 Harmonic structure information

Harmonic structure information is a well known acoustic cue for improving the noise robustness, which has been introduced in many VAD algorithms [11,18]. In [11], Fukuda et al. only incorporated the GMM model with harmonic structure information, and made a significant improvement in ASR system. This method assumes that the harmonic structure of pitch information is only included in the middle range of the cepstral coefficients. The feature extraction method is shown in Figure 2.

First, the log power spectrum $y_t(j)$ of each frame is converted into the cepstrum $p_t(i)$ by using the discrete cosine transform (DCT).

$$p_t(i) = \sum_i M_a(i,j) \cdot y_t(j), \tag{1}$$

where $M_a(i, j)$ is the matrix of DCT, and i indicates the bin index of the cepstral coefficients.

Then, the harmonic structure information q_t is obtained from the observed cesptra p_t by suppressing the lower and higher cepstra

$$\begin{aligned} q_t(i) &= p_t(i) \quad D_L < i < D_H, \\ q_t(i) &= \lambda p_t(i) \quad otherwise, \end{aligned} \tag{2}$$

where λ is a small constant.

After the lower and higher cepstra suppressed, the harmonic structure information $q_t(i)$ is converted back

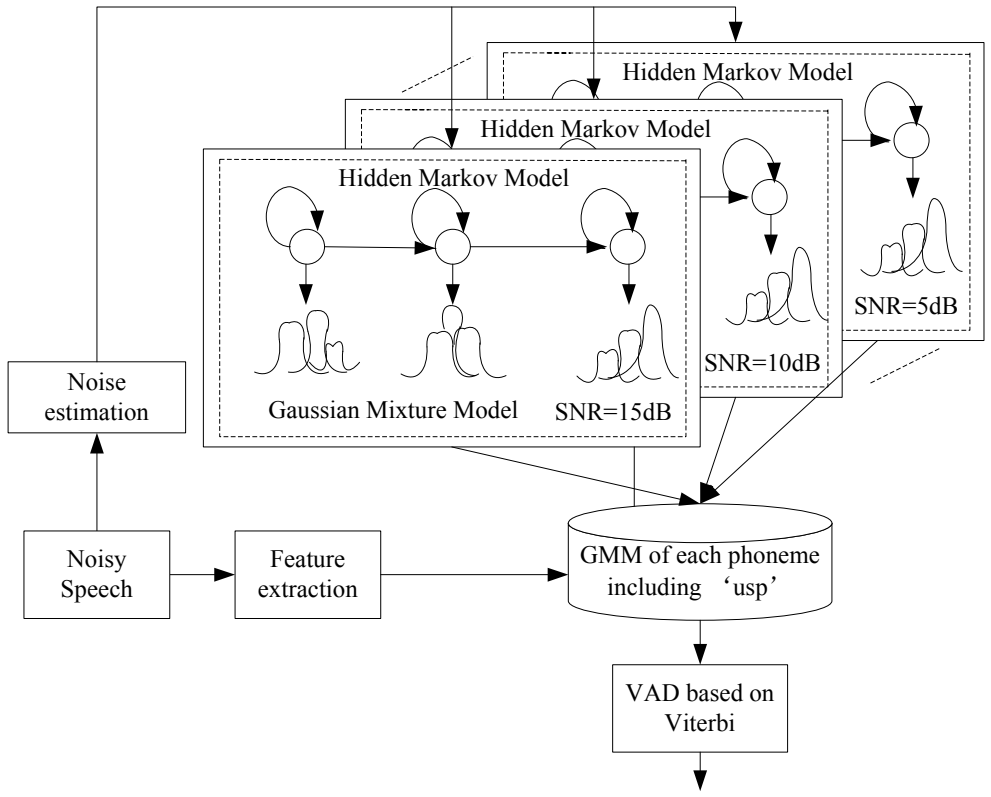

Figure 1 An overview of the proposed VAD algorithm.

to linear domain $w_t(j)$ by inverse DCT (IDCT) and exponential transform. Moreover, the $w_t(j)$ is integrated into $b_t(k)$ by using the K-channel mel-scaled band pass filter.

Finally, the harmonic structure-based mel cepstral coefficients are obtained when $b_t(k)$ is converted into the mel-cepstrum $c_t(n)$ by the DCT matrix $M_b(n, k)$.

$$c_t(n) = \sum_{k=1}^{K} M_b(n,k) \cdot b_t(k), \tag{3}$$

2.2.2 High order statistic

Generally, the HOS of speech are nonzero and sufficiently distinct from those of the Gaussian noise. Moreover, it is reported by Nemer et al. [19] that the skewness and kurtosis of the linear predictive coding (LPC) residual of the steady voiced speech can discriminate the speech from noise more effective.

Assume that $\{x(n)\}$, $n = 0, \pm1, \pm2,...$ is a real stationary discrete time signal and its moments up to order k exist, then the kth-order moment function is given as follows:

$$m_k(\tau_1, \tau_2 \dots \tau_{k-1}) \equiv E[x(n)x(n+\tau_1)\dots x(n+\tau_{k-1})] \tag{4}$$

where $\tau_1, \tau_2,..., \tau_{k-1} = 0, \pm1, \pm2,...$, and $E[\cdot]$ represents the statistical expectation. If the signal has zero mean, then the cumulant sequences of $\{x(n)\}$ can be defined:

Second-order cumulant

$$C_2(\tau_1) = m_2(\tau_1). \tag{5}$$

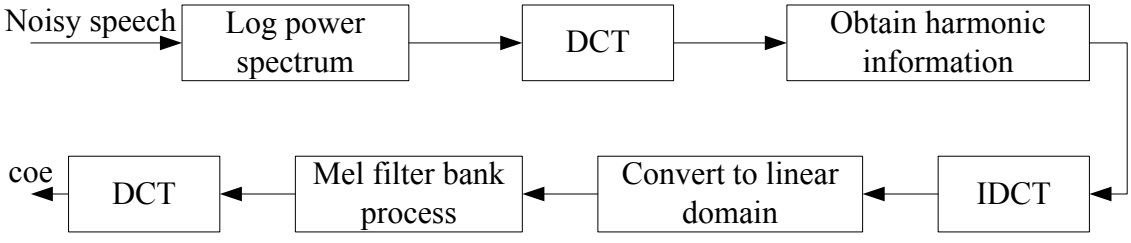

Figure 2 Harmonic structure feature.

Third-order cumulant

$$C_3(\tau_1, \tau_2) = m_3(\tau_1, \tau_2). \tag{6}$$

Fourth-order cumulant

$$C_4(\tau_1, \tau_2, \tau_3) = m_4(\tau_1, \tau_2, \tau_3) - m_2(\tau_1) \cdot m_2(\tau_2 - \tau_3) - m_2(\tau_2) \cdot m_2(\tau_3 - \tau_1) - m_2(\tau_3) \cdot m_2(\tau_1 - \tau_3). \tag{7}$$

Let $\tau_1, \tau_2, \ldots, \tau_{k-1} = 0$, then the higher-order statistics such as variance γ_2, skewness γ_3, kurtosis γ_4, can be expressed as follows respectively:

$$\gamma_2 = E[x^2(n)] = m_2, \tag{8a}$$

$$\gamma_3 = E[x^3(n)] = m_3, \tag{8b}$$

$$\gamma_4 = E[x^4(n)] - 3\gamma_2^2 = m_4 - 3m_2^2. \tag{8c}$$

Moreover, the steady voiced speech can be modeled as a sum of M coherent sine waves, and the skewness and kurtosis of the LPC residual of the steady voiced speech can be written as functions of the signal energy E_s and the number of harmonic M [12]:

$$\gamma_3 = \frac{3}{2\sqrt{2}}(E_s)^{\frac{3}{2}}\left[\frac{M-1}{M}\right], \tag{9}$$

and

$$\gamma_4 = {E_s}^2\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]. \tag{10}$$

2.3 VAD in HMM/GMM model

One of the most widely used method to model speech characteristics is Gaussian function or Gaussian mixture model. The GMM based VAD algorithm has attracted considerable attention for its high accuracy in speech/non-speech detection. However, the number of the mixtures of GMMs must be very large to distinguish the speech from non-speech, which increases the cost of calculation dramatically. Moreover, N-order GMMs can not discriminate the non-speech from speech precisely since the boundary between the speech and non-speech is not clear enough. In this article, we improve this method by regarding the non-speech as an additional phoneme (named as *'usp'*) corresponding to the conventional phonemes (such as *'zh'*, *'ang'* et al.) in mandarin, and using the GMMs based HMM hybrid model to discriminate the non-speech from speech.

In HMM/GMM based speech recognition [20], it is assumed that the sequence of observed speech vectors corresponding to each word is generated by a Hidden Markov model as shown in Figure 3. Here, a_{ij} and $b(o)$ means the transition probabilities and output probabilities respectively. 2, 3, 4 are the states of state sequence $\mathcal{X}$, and O_i represent the observations of observation sequence $\mathcal{O}$.

As well known, only the observation sequence $\mathcal{O}$ is known and the underlying state sequence $\mathcal{X}$ is hidden, so the required likelihood is computed by summing over all possible state sequences $\mathcal{X} = x(1), x(2), x(3), \ldots, x(T)$, that is

$$P(\mathcal{O}|\mathcal{M}) = \sum_{\mathcal{X}} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(O_t) a_{x(t)x(t+1)}, \tag{11}$$

where $x(0)$ is constrained to be the model entry state and $x(T + 1)$ is constraint to be the model exit state. The output distributions are represented by GMMs in hybrid model as

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mathbf{o}_t, \mu_{jm}, \mathbf{\Sigma}_{jm}), \tag{12}$$

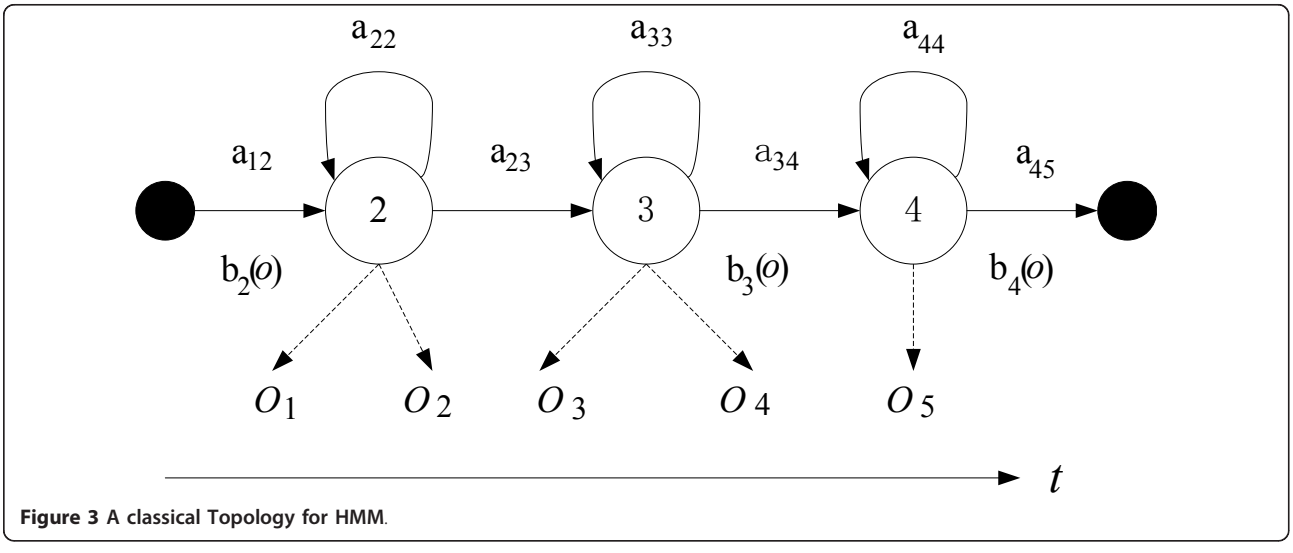

Figure 3 A classical Topology for HMM.

where M is the number of mixture components, c_{jm} is the weight of mth component and $\mathcal{N}(\mathbf{o}, \mu, \mathbf{\Sigma})$ is a multivariate Gaussian with mean vector μ and covariance matrix $\mathbf{\Sigma}$, that is

$$\mathcal{N}(\mathbf{o}, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{o}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{o}-\mu)}, \quad (13)$$

where n is the dimensionality of $\mathbf{o}$.

In the GMM/HMM based VAD method, we use the same method which is usually employed in ASR system by phoneme recognition. In first step, each phoneme (including the conventional phonemes and the non-speech phoneme) in GMM/HMM hybrid model are initialized. Then the underlying HMM parameters are re-estimated by Baum-Welch algorithm. In the step of discrimination, Viterbi algorithm is employed for searching the maximum likelihood of the observed signals, which can be referred to [20] for details. Note that, in our method, the triphones which are essential for ASR are not adopted here, because we think that the monophones based recognition is appropriate for discriminating the speech from the nonspeech.

3 A recursive phoneme recognition and speech enhancement method for VAD

It is mentioned that the Minimum Mean Square Error (MMSE) enhancement approach is much more efficient than other approaches in minimizing both the residual efficient and the speech distortion. Moreover, the non-stationary music-like residual noise after MMSE processing can be regarded as additive and stationary noise approximately, which ensures that some simplified model adaption method [14].

Let $S_k(n)$, $N_k(n)$, $Z_k(n)$ denote the kth spectral component of the nth frame of speech, noise and observed signal, respectively. And assume $A_k(n)$, D_k, $R_k(n)$ are the spectrum amplitude of $S_k(n)$, $N_k(n)$, $Z_k(n)$. Then the estimate $\hat{A}_k(n)$ of $A_k(n)$ can be given as [14]:

$$\hat{A}_k(n) = \frac{1}{2}\sqrt{\frac{\pi \xi_k}{\gamma_k(1+\xi_k)}} M(a; c; x) \cdot R_k(n), \quad (14)$$

where a = -0.5, c = 1, $x = -\gamma_k \xi_k/(1 + \xi_k)$, and $M(a; c; x)$ is the confluent hypergeometric function. ξ_k and γ_k are interpreted as the *a priori* and *a posteriori* SNR, respectively. The estimation of *a priori* and the *a posteriori* can be deemed as follow:

$$\hat{\xi}_k(n) = \alpha \frac{\hat{A}_k^2(n-1)}{\lambda_d(k, n-1)} + (1-\alpha) P(\gamma_k(n) - 1), \quad (15)$$

$$\gamma_k(l) = \frac{|Z_k(l)|^2}{\lambda_d(k)}, \quad (16)$$

where the noise variance $\lambda_d(k)$ is updated according to the result of VAD.

Generally, we always use the VAD based speech enhancement method for noise suppression before speech recognition. And it seems that the denoised speech is the optimal choice for ASR. If so, we may also can obtain a more accurate result of change point detection when we use the VAD method in the denoised speech. Following this idea, we propose a different VAD method which integrate our proposed VAD method (mentioned in Section 2) with the MMSE speech enhancement method, as shown in Figure 4.

The main steps of the proposed method are listed as follows (suppose the HMM/GMM models have been constructed).

1 The robust features which are mentioned above are extracted for representing each frame.
2 The change point detection between speech and non-speech is estimated by the phoneme recognition using the trained HMM/GMM model.
3 The variance of the noise is updated when the non-speech detected, *a priori* and *a posterior* of each frame are then calculated using the Equation (15) and (16).
4 The estimation $\hat{A}_k(n)$ is calculated using the Equation (14).

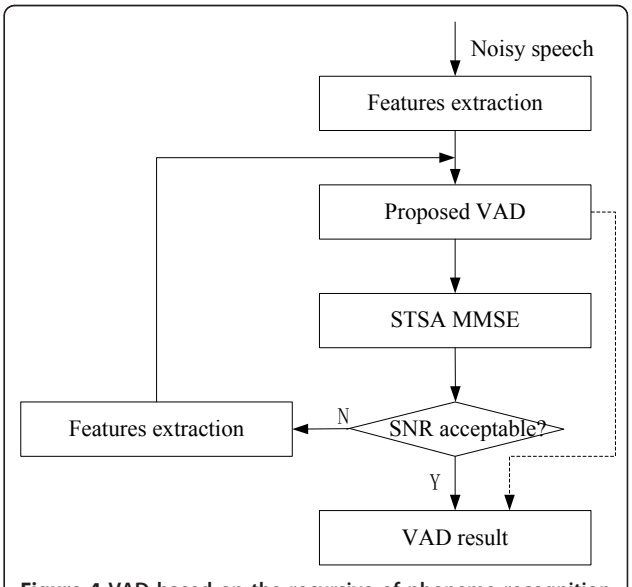

Figure 4 VAD based on the recursive of phoneme recognition and speech enhancement.

5 Estimate the SNR of the denoised speech to justify whether the SNR is larger than 15 dB or not. If the SNR is less than 15 dB, then back to step 1, else the result estimated in step 2 is the final VAD result.

4 Experimental results

In this section, the performances of the proposed method are evaluated. The MSRA mandarin corpus test data that has 500 utterances with 0.74 h length is used as the test set, and the training set from MSRA has 19688 utterances with 31.5 h length, referring to [21] for details.

In this article, the feature parameters for the HMM/GMM hybrid model based VAD are extracted at intervals of 20 ms frame length and 10 ms frame shift length, composed of 13th order harmonic structure information features, 1st order skewness, 1st order kurtosis, 12th order log-Mel spectra with energy and its Δ, leading to an HMM set with 5 states.

To illustrate the statistical properties of speech signals, we take one of the test utterances as an example, shown in Figure 5a. As we can see, the proportion of voiced speech to unvoiced speech is almost 3:1.

Three different types of experiments are considered here. First, we want to find out whether the increase of the number of the GMM mixtures can improve the accuracy of VAD. Then, we compare the proposed VAD method with some existing VAD methods to determine whether the proposed method is more robust to the noise. And in the last experiment, we use a different method to demonstrate that the conventional noise suppression method is detrimental to the speech quality even giving precise VAD results at low SNR regimes.

4.1 Relationship between the VAD accuracy and the number of mixtures

Figure 5b,c depicts the results of VAD by HMM/GMM hybrid model at non-stationary noise environments. The number of the mixtures of GMM here is 4. The non-stationary noise is downloaded from http://www.freesound.org.

From Figure 5b, we can find the proposed VAD method is very robust to the high SNR noise since the detection of change point is almost completely correct. And the result of the detection accuracy is also excellent when the SNR is low as shown in Figure 5c.

Less number of the mixtures not only can save the time of discriminating the unvoiced speech from voiced speech, but also can reduce the memory of storing the GMM parameters. So, with acceptable accuracy of VAD, the number of the mixtures are the less the better.

In order to investigate the precision of the proposed method in different GMMs mixture number, we take all the 500 test utterances as examples to obtain the probabilities of accurate VAD detection P_a at different kinds of noise with different SNRs.

$$P_a = \sum_{i=1}^{N} |o(i) - d(i)|/N, \tag{17}$$

where N is the total number of the corpus frames, $o(i)$ = 0, 1 denotes the labeled speech/non-speech segments, and $d(i)$ = 0, 1 denotes the estimated speech/non-speech segments.

Figure 6 and Table 1 give the VAD results of the proposed method from different mixtures of GMMs at different kinds of noise environments with different SNRs.

In Figure 6, the *ylabel* denotes the accuracy of VAD, and the *xlabel* denotes the SNR regimes.

In Table 1, we give another three noise environments as non-stationary noise environments, in-car noise environment and city street noise environment for test the proposed VAD algorithm, where the noise environment is named as *NE* for short.

Examining Figure 6 and Table 1, we note some interesting points:

- When the noise is Gaussian or Gaussian-like noise, such as gaussian white noise in Figure 6, the

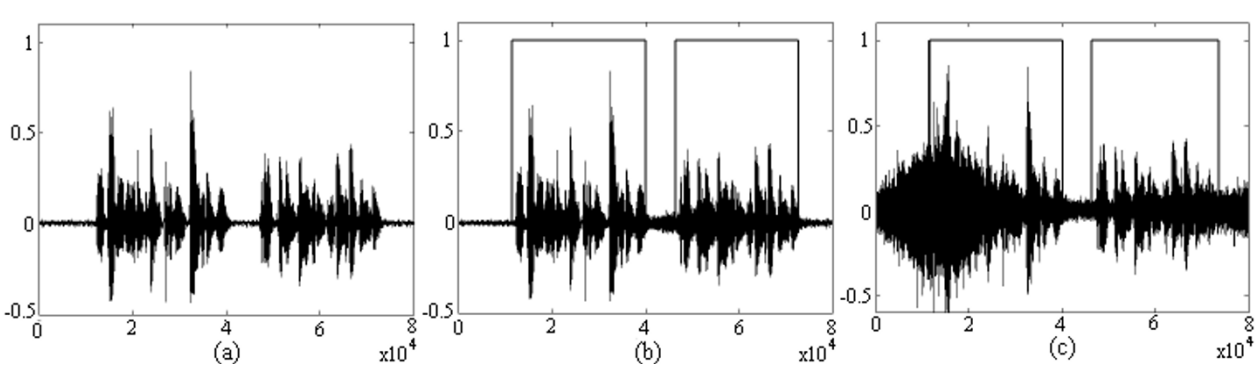

Figure 5 An example of the HMM/GMM based VAD with car passing noise. **(a)** Clean speech, **(b)** SNR = 15 dB, **(c)** SNR = 5 dB.

performance of the proposed VAD algorithm is excellent even at low SNR regimes. However, when meets the non-stationary noise, the algorithm is not robust enough at low SNRs.

- When the number of the mixtures of GMMs increases, the accuracy of the proposed VAD seems to not increase by the same rules. As seen from Table 1 and Figure 6, when the SNR is high, the performance of low order GMMs is better than the performance of the higher order GMMs.
- The VAD algorithm in Gaussian white noise and city street noise have much better performances than in other noises. This also demonstrates the HOS is robust to the Gaussian/Gaussian-like noise.
- The mix4 has much stable result than any other mixtures in most noisy environments using the phoneme recognition method based on HMM/GMM hybrid model.

Table 1 VAD results from different GMM orders at different kinds of environments with different SNRs

NE	SNR (dB)	mix1 (%)	mix2 (%)	mix4 (%)	mix8 (%)	mix16 (%)
n-stat	-5	77.31	78.90	78.46	83.19	83.17
	0	78.22	81.33	82.28	83.40	85.74
	5	81.37	82.66	81.65	82.93	84.13
	10	85.45	85.97	86.47	88.70	90.66
	15	87.91	91.33	92.67	91.91	92.82
	20	97.01	96.78	96.49	95.94	95.21
In car	-5	94.68	94.80	94.81	94.91	94.94
	0	95.70	95.78	95.72	95.70	95.50
	5	96.51	96.36	96.52	95.70	95.58
	10	96.80	96.57	96.58	96.23	95.84
	15	97.49	97.36	97.12	96.53	95.90
	20	97.45	97.40	97.16	96.47	95.91
Street	-5	88.85	89.21	91.42	94.91	94.94
	0	93.84	93.90	94.33	94.49	94.36
	5	95.15	95.40	95.24	95.33	95.05
	10	96.02	96.32	95.94	95.86	95.46
	15	96.68	95.60	97.03	96.38	95.60
	20	97.15	97.32	97.27	96.61	95.47

4.2 Comparative analysis of the proposed VAD algorithms

In order to gain a comparative analysis of the proposed VAD performance under different environments such as the vehicle and street, several classic VAD schemes are also evaluated. The results are summarized in Table 2, where the MOLRT is a method proposed by Lee [22]. The number of the mixtures in the proposed scheme is 4 according to the result of Table 1.

It is seemed that for all the testing cases, the performance of the proposed VAD is better than that of the G.729B VAD, the LRT by Sohn and MOLRT by Lee, except for the case of the non-stationary noise with a SNR of -5 dB, where the performance of the proposed VAD is slightly worse than that of the MOLRT based VAD. In case of the stationary noise, the accuracy of the proposed VAD is higher than 90% in any SNR level.

4.3 VAD based on the recursive method

In our study, VAD based ASR system is not studied, but we do another experiment to find out whether the

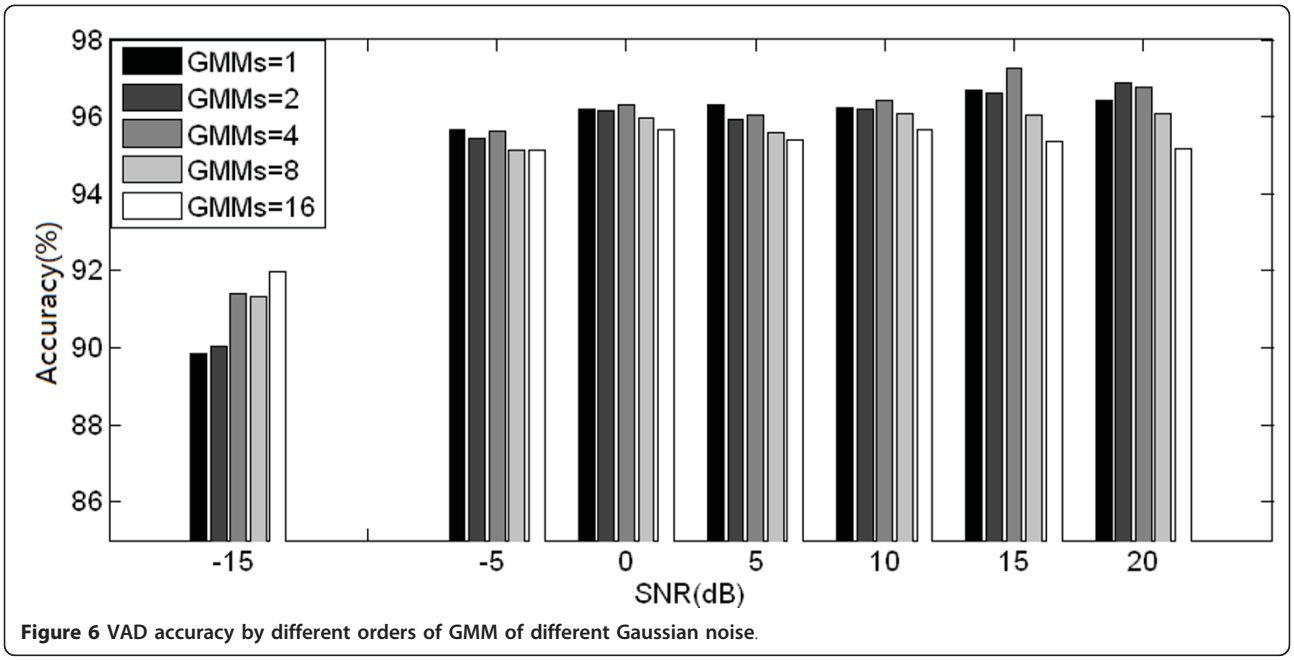

Figure 6 VAD accuracy by different orders of GMM of different Gaussian noise.

Table 2 Comparison results at different kinds of environments with different SNRs

NE	SNR (dB)	Proposed (%)	G.729B (%)	Sohn (%)	MoLRT (%)
White	-5	95.64	71.64	71.71	88.33
	0	96.31	72.55	79.75	94.91
	10	96.41	73.68	86.17	94.75
	20	96.77	74.65	87.58	93.17
Non-stat	-5	78.46	70.60	81.74	80.56
	0	82.28	70.61	83.08	81.47
	10	86.47	70.92	84.15	82.75
	20	96.49	71.84	85.23	84.69
In car	-5	94.81	71.22	83.78	79.79
	0	95.72	72.31	85.03	87.98
	10	96.58	74.62	85.98	92.41
	20	97.16	75.70	85.78	91.30
Street	-5	91.42	76.29	79.95	84.71
	0	94.33	76.28	82.32	87.65
	10	95.94	75.61	84.39	90.09
	20	97.27	75.38	85.59	90.17

integration of proposed VAD with the conventional speech enhancement can recover the clear speech at low SNR regimes or not.

We take Figure 5a as the speech prototype, and the VAD results at different noise environments are shown in Figures 7 and 8. In Figures 7 and 8a, the VAD results are obtained according to the proposed VAD algorithm, and Figures 7 and 8b show the VAD results based on the integration method.

Examining Figures 7 and 8, we can conclude some interesting points:

- When comparing Figure 7a with Figure 8a, the proposed VAD algorithm is much more robust to the stationary noise than the non-stationary noise.
- Comparing Figure 7a with Figure 7b, and comparing Figure 8a with Figure 8b, we can find if the accuracy of the VAD algorithm is very high, the combination method can keep the VAD accuracy, else the performance will degrade dramatically.

5 Discussion

Some VAD algorithms which have been demonstrated robust to the noise are introduced to the ASR system, and the performance of the speech recognition seems not bad in high SNR level. For example, Fukuda combines the VAD algorithm with Wiener filter before ASR. However, we think that there are something more should be done before ASR. So, we first propose a novel VAD algorithm based on HMM/GMM hybrid model, which is confirmed further by the following experiment to be more robust in many noise environments. Then we combine the proposed VAD with the speech enhancement algorithm for change point detection to find out what should be done before ASR.

The novel VAD algorithm proposed in this article is based on the phoneme recognition using HMM/GMM hybrid model, which is much different from the existing VAD methods. In our study, different GMMs orders are considered to improve the VAD accuracy, but it seems that the accuracy could not be improved when the orders become higher.

In order to gain a comparative analysis of the proposed VAD performance under different environments, several classic VAD schemes are also evaluated. And the results show that the proposed VAD method is more useful than the existing methods.

We propose a different detection method to indirectly show the reason why the performance of the ASR system are not well accepted at low SNR regimes, named 'A recursive phoneme recognition and speech

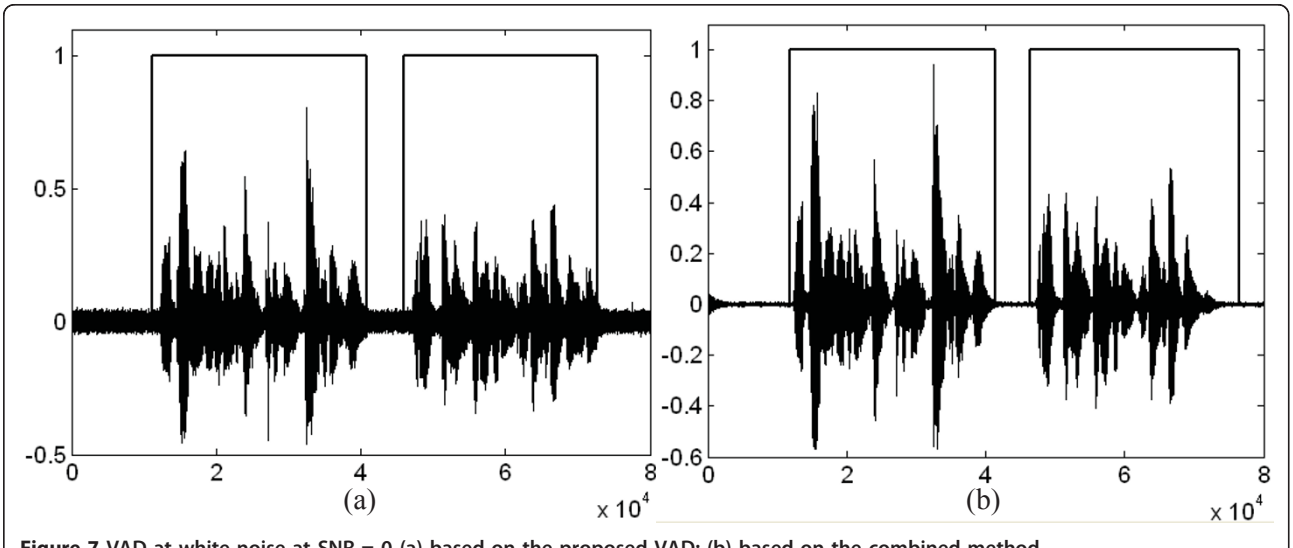

Figure 7 VAD at white noise at SNR = 0 (a) based on the proposed VAD; (b) based on the combined method.

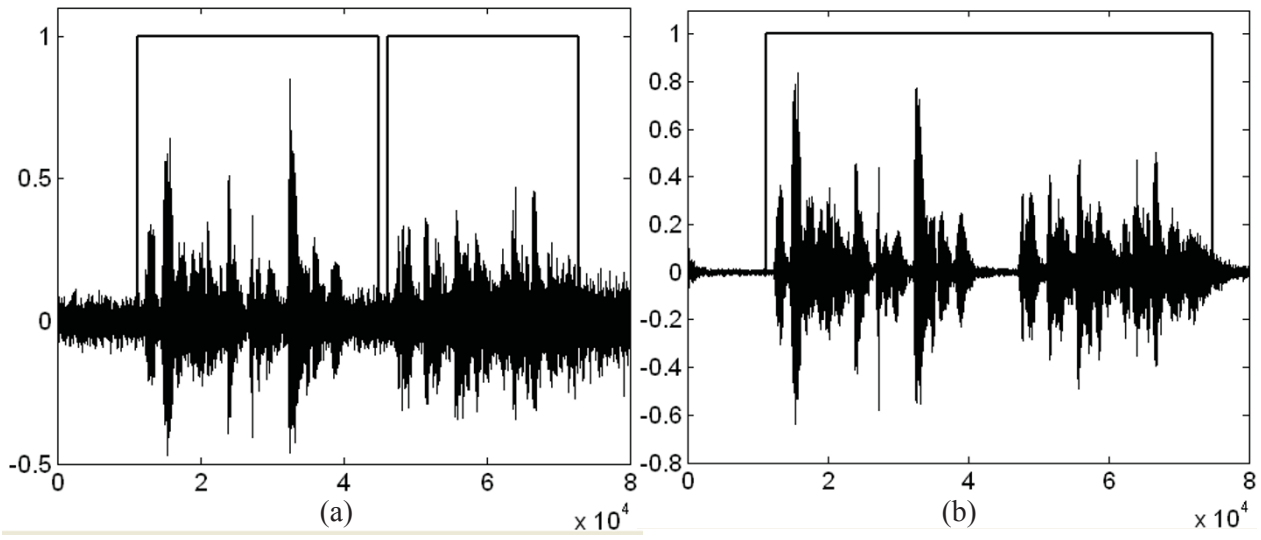

Figure 8 VAD at non-stationary noise at SNR = 0 (a) based on the proposed VAD; (b) based on the combined method.

enhancement method for VAD'. And the experimental result is shown in the Section 4.3. Some points are concluded:

- If the accuracy of the VAD is more than 95%, the noise can be suppressed well with the little speech distortion. And it is helpful for ASR.
- When the accuracy drops down, the speech can not be recovered well in the noisy speech, despite that the noise of unvoiced speech can be suppressed. Apparently, the performance of speech recognition will degrade, and become even worse than the speech recognition without noise suppression.

From Table 1, we have found the accuracy of the VAD is well accepted in most environment at any SNRs. However, the VAD accuracy can not be improved much when the noise is suppressed by the speech enhancement method, as shown in Figure 8. It also means the speech enhancement method damage the speech a lot during the suppression of the noise at low SNRs. If we could keep the quality of the source signal by speech enhancement method, the clear speech can be recovered.

6 Conclusion

In this article, we propose a phoneme recognition based VAD method that follows the idea of phoneme recognition. Note that, the proposed method is much different from others since HMM/GMM based phoneme recognition is only used for VAD here while others use phoneme recognition for ASR or some other applications. Some sophisticated features are combined to represent the speech segments. Experiments performed on MSRA mandarin speech data set confirm the advantage. We compare the proposed algorithm with some popular VAD methods, and results exhibit the good performance of the proposed algorithm. In the section of 'VAD based on the recursive method', we also find more study should be done in the future. First, more robust VAD algorithm should still be pursued. Second, noise estimation algorithm should be introduced to the ASR system to forecast the noise component of noisy speech. Third, Some limitations should be set to reduced the distortion of speech. Last, more robust speech enhancement algorithm is desired.

Competing interests
The authors declare that they have no competing interests.

Received: 19 September 2011 Accepted: 9 January 2012
Published: 9 January 2012

References

1. JH James, B Chen, L Garrison, Implementing VoIP: a voice transmission performance progress report. IEEE Commun. Mag. **42**(7), 36–41 (2004)
2. C Wang, K Sohraby, R Jana, J Lusheng, M Daneshmand, Voice communications over ZigBee networks. IEEE Commun. Mag. **46**(1), 121–127 (2008)
3. J Chien, C Ting, Factor analyzed subspace modeling and selection. IEEE Trans. Audio, Speech and Lang. Process. **16**(1), 239–248 (2008)
4. Y Shao, C Chang, A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system. IEEE Trans. Systems, Man, and Cybernetic. **37**(4), 877–889 (2007)
5. J Ramirez, JC Segura, JM Gorriz, L Garcia, Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. IEEE Trans. Audio, Speech and Lang. Process. **15**(8), 2177–2189 (2007)
6. S Yamamoto, K Nakadai, M Nakano, *et al*, Real-time robot audition system that recognizes simultaneous speech in the real world, in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 5333–5338 (2006)

7. M Fujimoto, K Ihizuka, T Nakatani, A voice activity detection based on the adaptive integration of multiple speech features and signal decision scheme, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* 4441–4444 (2008)
8. G Evanglelopulos, P Maragos, Multiband modulation energy tracking for noisy speech detection. IEEE Trans. Audio, Speech and Lang. Process. **14**(6), 2024–2038 (2006)
9. J Padrell, D Macho, C Nadeu, Robust speech activity detection using LDA applied to FF parameters, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. **1**, 557–560 (2005)
10. M Asgari, A Sayadian, M Frahadloo, EA Mehrizi, Voice Activity Detection Using Entropy in Spectrum Domain, in *Telecommunication Networks and Applications Conference,* 407–410 (2008)
11. T Fukuda, O Ichikawa, M Nishimura, Improved voice activity detection using static harmonic features, in *Proceeding of the IEEE International Conference on Acoustics Speech and Signal Processing,* 4482–4485 (2010)
12. K Li, MNS Swamy, OM Ahmad, An improved voice activity detection using higher order statistics. IEEE Trans. Speech and Audio Process. **13**(5), 965–974 (2005)
13. J Sohn, NS Kim, W Sung, A statistical model-based voice activity detection. IEEE Signal Process. Lett. **16**(1), 1–3 (1999)
14. YD Cho, K Al-Naimi, A Kondoz, Improved voice activity detectionbased on a Smoothed statistical likelihood ratio, in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*. **2**, 737–740 (2001)
15. JM Gorriz, J Ramirez, EW Lang, CG Puntonet, Jointly Gaussian PDF-Based Likelihood Ratio Test for Voice Activity Detection. IEEE Trans. On Audio, Speech and Lang. Process. **16**(8), 1565–1578 (2008)
16. M Fujimoto, K Ishizuka, H Kato, Noise Robust Voice Activity Detection based on Statistical Model and Parallel Non-linear Kalman Filtering. in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*. **4**, 797–800 (2007)
17. I Cohen, B Berdugo, Noise estimation based by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process, Lett. **9**(1), 12–15 (2002). doi:10.1109/97.988717
18. K Ishizuka, T Nakatni, M Fujimoto, N Miyazaki, Noise robust voice activity detection based on periodic to aperiodic component ratio. Speech Commun. **52**(1), 41–60 (2010). doi:10.1016/j.specom.2009.08.003
19. E Nemer, R Goubran, S Mahmoud, Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Trans. Speech Audio Process. **9**(3), 217–231 (2001). doi:10.1109/89.905996
20. S Young, D Kershaw, J Odell, D Ollason, V Valtchev, P Woodland, The HTK Book. Available from http://htk.eng.cam.ac.uk/docs/docs.shtml
21. HH Xu, J Zhu, GY Wu, An efficient multistage ROVER for automatic speech recognition. in *roceeding of the IEEE International Conference on Multimedia and Expo,* 894–897 (2009)
22. LN Tan, BJ Borgstrom, A Alwan, Voice activity detection using harmonic frequency components in likelihood ratio test. in *Proceeding of the IEEE International Conference on Acoustics Speech and Signal Processing,* 4466–4469 (2010)

doi:10.1186/1687-4722-2012-1
Cite this article as: Bao and Zhu: **A novel voice activity detection based on phoneme recognition using statistical model.** *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:1.