Pattern Recognition Letters 32 (2011) 1044–1053

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An improved noise-robust voice activity detector based on hidden semi-Markov models

Yuan Liang [a,*,1], Xianglong Liu [a,1], Yihua Lou [a], Baosong Shan [b]

[a] State Key Laboratory of Software Development Environment, Beihang University, China
[b] School of Mathematics and Systems Science, Beihang University, China

ARTICLE INFO

Article history:
Received 20 May 2010
Available online 21 February 2011
Communicated by P. Franti

Keywords:
Voice activity detection
State duration
Observation distribution
Hidden semi-Markov model
Likelihood ratio test
Forward variable

ABSTRACT

To improve the performance of voice activity detector (VAD) in noisy environments, this paper concentrates on three critical aspects related to noise robustness including speech features, feature distributions and temporal dependence. Based on the statistic on TIMIT and NOIZEUS, Mel-frequency cepstrum coefficients (MFCCs) are selected as speech features, Gaussian Mixture distributions (GMD) are applied to associate the observations in MFCC domain with both speech and non-speech states, and Weibull and Gamma distributions are used to explicitly model noise and speech durations, respectively. To integrate these aspects into VAD, the hidden semi-Markov model (HSMM) as a generalized hidden Markov model (HMM) is introduced first. Then the VAD decision is made according to the likelihood ratio test (LRT) incorporating state prior knowledge and modified forward variables of HSMM. We design a recursive way to efficiently calculate modified forward variables. Finally a series of experiments demonstrate: (1) the positive effect of different robustness-related schemes adopted in the proposed VAD; (2) better performance against the standard ITU-T G.729B, Adaptive MultiRate VAD phase 2 (AMR2), Advanced Front-end (AFE), HMM-based VAD and VAD using Laplacian–Gaussian model (LD–GD based VAD).

1. Introduction

Voice activity detector (VAD) refers to the system of distinguishing active speech from non-speech frames. It has been used for various applications such as speech coding, transmission, enhancement and recognition.

Recently, many attractive statistical model-based VAD algorithms using the likelihood ratio test (LRT) have been developed (Davis et al., 2006; Shin et al., 2007; Sohn et al., 1999). They have made significant contributions to voice activity detection progress, especially the statistical methods based on hidden Markov models (HMMs) (Gazor and Zhang, 2003; Othman and Aboulnasr, 2004). In HMM-based VADs, there are three key aspects related to their performance: speech features (Fujimoto and Ishizuka, 2008), feature distributions (Chang et al., 2006; Gazor and Zhang, 2003) and temporal dependence (Othman and Aboulnasr, 2004; Rabiner, 1989). In this paper we will pay our attention to these three aspects, especially the last one.

For speech features, most VAD algorithms often use discrete Fourier transform (DFT) (Davis et al., 2006; Sohn et al., 1999), discrete cosine transform (DCT) (Gazor and Zhang, 2003) coefficients and Mel-Frequency cepstral coefficients (MFCC) (Fujimoto and Ishizuka, 2008). Our experiments in low signal-to-noise ratio (SNR) environment show that MFCC, known to be good at capturing the important characteristics of speech, performs better than DFT and DCT coefficients in LRT. To further improve the robustness of features, a Wiener filter (Adami et al., 2002) is adopted to reduce noise present in speech before feature extraction. For observation distribution, a common assumption in DCT or DFT domain is that Gaussian distributions (GD) can fit both noise and speech observations (Sohn et al., 1999). However some other works (Chang et al., 2006; Gazor and Zhang, 2003; Shin et al., 2007) have demonstrated that Laplacian (LD) and generalized Gamma distributions (ΓD) are more suitable for approximating voiced speech than GD. According to the statistical results on TIMIT (Garofolo et al., 1993) and NOISEX-92 (Varga and Steeneken, 1993), the coefficients in MFCC domain do follow none of LD, generalized ΓD or GD as they do in DCT and DFT domains, but Gaussian mixture distributions (GMD) with a limit number of components (Liang et al., 2010).

Temporal dependence, usually concerned as a hangover schema, in HMM-based VADs is described by state duration which strictly follows a geometric distribution (GED) (Rabiner, 1989). However in real world this assumption is not always true (Chen et al., 1995; Dong and He, 2007). To alleviate the limitation, researchers have proposed a number of techniques like the continuous variable duration HMM (Othman and Aboulnasr, 2004), the hidden semi-Markov model (HSMM) (Murphy, 2002) and variable

* Corresponding author. Tel.: +86 10 82338138; fax: +86 10 82316736.
E-mail addresses: liangyuan@nlsde.buaa.edu.cn (Y. Liang), xlliu@nlsde.buaa.edu.cn (X. Liu).
[1] Indicates equal contribution.

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.02.015

frame rate approaches (Tan and Lindberg, 2010). It has been proven that these techniques benefit recognition performance (Chen et al., 1995; Dong and He, 2007).

Intuitively, we think that the temporal dependency is invariant to noise type and noise level, so we can model it explicitly to make VAD algorithms more robust to noise. In this paper, we introduce HSMM, a generalized HMM, to explicitly model the state duration. To our knowledge, this is the first work that explicitly models speech durations by combining HSMM and LRT for VAD.

According to statistical test on TIMIT and NOIZEUS (Hu and Loizou, 2008), the durations of speech and non-speech states are respectively modeled by ΓD and the Weibull distribution (WD) (Liang et al., 2010; Liu et al., 2010). To integrate these robustness-related aspects into VAD, we propose a detector based on HSMM incorporating LRT for speech/nonspeech detection. Specifically, the major contributions of this work are as follows:

- We concentrate on three robustness-related aspects of speech and adopt corresponding schemes to improve the performance of VAD: MFCC is selected as speech features in noisy environment and GMD as the feature probability density function (PDF); We first model speech and noise durations explicitly, where Weibull and Gamma PDFs are adopted according to the statistic on TIMIT and NOIZEUS. Experiments validate the positive effects brought by these schemes.
- We first introduce HSMM to integrate these schemes into VAD, and implement an HSMM-based voice activity detector. In this detector, the VAD decision is made according to LRT combining both state prior knowledge and modified forward variables. An efficient way similar to Viterbi of HMMs (Rabiner, 1989) is devised to recursively calculate the modified forward variables, and parameters are adjusted dynamically to improve robustness to the varying noisy environment. Experiments were conducted between the proposed HSMM-based VAD and ITU-T G.729B (Benyassine et al., 1997) VAD, Adaptive MultiRate VAD phase 2 (AMR2) (ETSI, 1998), Advanced Front-end (AFE) (ETSI, 2002), HMM-based VAD (Sohn et al., 1999) and VAD based on Laplacian–Gaussian model (LD–GD based VAD) (Gazor and Zhang, 2003). Experimental results show that HSMM-based VAD achieves the best performance.

This paper is organized as follows: In Section 2 the hidden semi-Markov model is introduced. Section 3 presents the segmental HSMM-based modeling framework for VAD including HSMM components, dynamical parameter adjustment and modified forward variables calculation. Section 4 first gives parameters estimation and then evaluates HSMM-based VAD compared with other algorithms. Finally, conclusions are drawn in Section 5.

2. Hidden semi-Markov models

An HSMM (Murphy, 2002) consists of a pair of discrete-time stochastic processes $\{S_t\}$ and $\{O_t\}$. Similar to HMMs, the observed process $\{O_t\}$ is related to the unobserved or hidden semi-Markovian state process $\{S_t\}$ by the conditional distributions.

Let $O_1^T = O_1 O_2 \cdots O_T$ denote the observation sequence of length T. The same notation is used for the state sequence $S_1^T = S_1 S_2 \cdots S_T$, and $\lambda = (A, B, \pi, \tau)$ denotes the set of model parameters as shown in Fig. 1:

1. N, the number of states $Q_0^{N-1} = (q_0, \ldots, q_{N-1})$ in the model.
2. $\pi = \{\pi_i\}$, where $\pi_i = P(S_1 = q_i)$ is the initial probability of being in state q_i, and $\sum_{i=0}^{N-1} \pi_i = 1$.
3. $A = \{a_{ij}\}$ models probabilistic transitions between different states. For states $q_i, q_j \in Q_0^{N-1}$:

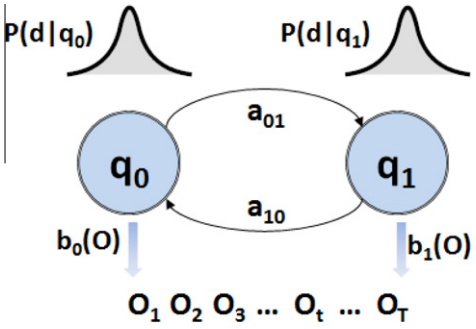

Fig. 1. A hidden semi-Markov model with two states.

$$a_{ij} = P(S_{t+1} = q_j | S_t = q_i) \tag{1}$$

satisfying $\sum_j a_{ij} = 1$, and $a_{ii} = 0$.

4. $B = \{b_i(O_t)\}$ is the PDF for the observation O_t given the state q_i:

$$b_i(O_t) = P(O_t | S_t = q_i), \tag{2}$$

where O_t only depends on S_t.

5. $\tau = \{P(d - q_j)\}$ represents the state duration distribution, which reflects the temporal dependence of states. It models the duration of $\{S_t\}$ remaining in q_j:

$$P(d|q_j) = P(S_{t+d+1} \neq q_j, S_{t+d} = q_j, \ldots, S_{t+2} = q_j | S_{t+1} = q_j, S_t \neq q_j) \tag{3}$$

Similar to HMMs, there also exist three basic problems associated with HSMMs (Murphy, 2002).

(1) Evaluation (scoring): Given the observation sequence O_1^T and an HSMM λ, compute the probability of the observation sequence given the model, i.e., $P(O - \lambda)$.
(2) Learning (training): Estimate the model parameters $\lambda = (A, B, \pi, \tau)$ to maximize $P(O_1^T | \lambda)$.
(3) Recognition (decoding): Given the observation sequence $O_1^t, 1 \leqslant t \leqslant T$ and the model $\lambda = (A, B, \pi, \tau)$, find the most probable states S_t, $1 \leqslant t \leqslant T$. There are two points of view to solve this problem: (a) from the global respective, the whole sequence S_1^T satisfying $\arg\max_{S_1^T} P(S_1^T | O_1^T, \lambda)$ which means that S_1^T will be the most probable state sequence; (b) from the local respective, for real-time applications the most probable state S_t at moment t is concerned, namely S_t satisfying $\arg\max_{S_t} P(S_t | O_1^t, \lambda)$. In this paper we derive modified forward variables to solve the recognition problem from the local respective.

3. HSMM-based voice activity detection

Modern VAD algorithms' hangover schemes using HMMs implicitly describe the state duration effect on likelihood of state transition. In conventional HMMs, the state duration PDF $P(d - q_i)$ is a geometric distribution of d (Rabiner, 1989):

$$P(d|q_i) = a_{ii}^{d-1}(1 - a_{ii}) \tag{4}$$

It has been argued that this is a source of inaccurate duration modeling with the HMMs since this geometrically decaying function of state durations is inappropriate for most real-life applications (Chen et al., 1995). Fig. 2 shows state duration distribution of TIMIT compared with that of HMMs. It can be concluded that the real-world duration distribution (the solid curve) differs from the geometrical functions (dot straight lines). To alleviate the limitation of state duration distribution (Othman and Aboulnasr, 2004) uses a contin-

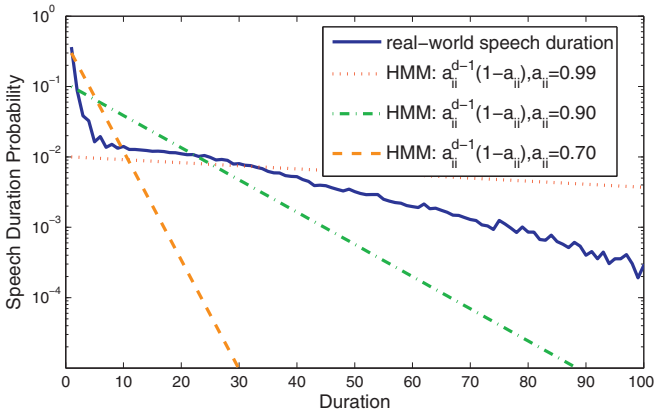

Fig. 2. Speech duration distribution of a TIMIT subset compared with HMMs.

uous transition probability of HMMs to directly control the variable duration in voice activity detection. The experimental evidence demonstrates that this hangover scheme can benefit the detection performance (Gazor and Zhang, 2003; Othman and Aboulnasr, 2004). However, the transition probability intuitively chosen to be exponential function is still inconsistent with real-world situation shown in Fig. 2.

In this paper, HSMM is introduced to explicitly model the duration distribution. Signals composed of speech and noise are regarded as a time duration hidden Markov chain with two states (speech and non-speech), and are modeled by the HSMM $\lambda = (A, B, \tau, \pi)$ shown in Fig. 1. Similar to conventional HMM-based VADs, three critical aspects including speech features, feature distributions and temporal dependence are concerned. Next, we will present components of the segmental HSMM-based modeling framework λ for VAD.

3.1. The state transition layer

For the state transition layer also named hidden layer, three components are concerned: states (speech and non-speech) transition, state durations and observation distributions associated with states.

3.1.1. Transition probability

Fig. 1 shows that in HSMM-based VAD $N = 2$ and a_{ij} satisfies $\sum_{j=0}^{N-1} a_{ij} = 1$ with $a_{ii} = 0$, so the transition matrix $A = (a_{ij})$ is given by:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{5}$$

3.1.2. Duration probability

Signal frames remaining at the same state are regarded as a segment, and sojourn time d of each segmentation is named state duration. For either state the likelihood of transiting to the other varies over its duration, and is modeled by certain distribution $\tau = \{P(d|q_i)\}$. To obtain duration distribution, one way is to estimate them from the training samples. The statistic of both TIMIT and NOIZEUS shows that the durations of noise and speech states share some similar properties and can be modeled well by certain distributions.

The noise state duration, namely the interval between two speech segments, can be viewed as the intermittence of vocalization actions. In engineering, such lifetime can be modeled by the Weibull distribution (WD) which is one of the most widely used lifetime distribution. Therefore, we choose the Weibull distribution to model noise state durations:

$$P(d|q_0) = \frac{k_0}{\omega_0}\left(\frac{d}{\omega_0}\right)^{k_0-1} e^{\left(\frac{d}{\omega_0}\right)^{k_0}}, \tag{6}$$

where $d > 0$ is the length of duration in state q_0, $k_0 > 0$ is the shape parameter and $\omega_0 > 0$ is the scale parameter of WD. In our experiments, duration statistics of TIMIT subset are fitted to it well as shown in Fig. 3(a) (the closer the Q–Q plot (red line) is to $x = y$ (blue[2] dot line), the better the PDF fits the data).

After noise state ends, the duration of speech state can be regarded as its waiting time to the end. The waiting time may depend on its current duration. Such kind of waiting time is usually modeled by the Gamma distribution (ΓD):

$$P(d|q_1) = \frac{1}{\omega_1^{k_1}\Gamma(k_1)} d^{k_1-1} e^{\frac{d}{\omega_1}}, \tag{7}$$

where similarly $k_1 > 0$ and $\omega_1 > 0$ are the shape and the scale parameters of ΓD. The ΓD is useful in reliability models of lifetimes and is frequently used to model waiting times. Usually ΓD is more flexible than the exponential distribution in which the probability of a frame surviving an additional period may depend on its current age. Therefore it is quite suitable for speech duration. As it is depicted in Fig. 3(b), the duration statistics of a TIMIT subset are fitted to ΓD quite well.

To determine parameters of both WD and ΓD, one popular criterion of goodness is to maximize the likelihood function. More details about maximum likelihood estimators for these parameters will be presented in Section 4.2.

3.2. The observation layer

In the observation layer, the observation O_t, the feature vector of the frame at moment t, being a speech or a noise frame is concerned. O_t is often calculated by DCT, DFT, and MFCC processors. We compared these three features in low SNR environments, and the experimental results in Section 4 show that MFCC, which is believed to be capable to capture the important characters of human voice, outperforms DCT and DFT coefficients in the likelihood ratio test. Therefore MFCC is chosen as the speech feature in this paper.

Once the feature has been selected, the conditional likelihood $P(O_t|q_i)$ is estimated based on an observation distribution associated with each state. Experiments on subset of TIMIT and NOISEX-92 show that for both clean speech and noise signals, the coefficients follow none of GD, LD and ΓD (the most common assumption in VAD algorithms) in MFCC domain (see Fig. 4), but Gaussian Mixture distributions (GMD) with a limit number of components. Thereby, we use GMD which can approximate the arbitrary probability density by adjusting its parameters for both speech and non-speech states. The feature O_t consists of K MFCCs o_i ($1 < i < K$) whose PDFs conditioned on state q_i are given by:

$$b_i(O_t) = P(O_t|S_t = q_i; \theta) = \sum_{j=1}^{M} w_{ij}\mathcal{N}(O_t; \boldsymbol{\mu}_{ij}, \Sigma_{ij}) \tag{8}$$

where θ is the parameter set of GMD, w_{ij}, $\boldsymbol{\mu}_{ij}$ and Σ_{ij} are the weight, mean vector and covariance matrix for the jth ($1 \leqslant j \leqslant M$) Gaussian component of state q_i, respectively, and M is the number of components in GMD. These parameters are estimated using the iterative Expectation–Maximization (EM) algorithm. More details about the parameters estimation will be given in Section 4.2.

[2] For interpretation of color in Figs. 1-12, the reader is referred to the web version of this article.

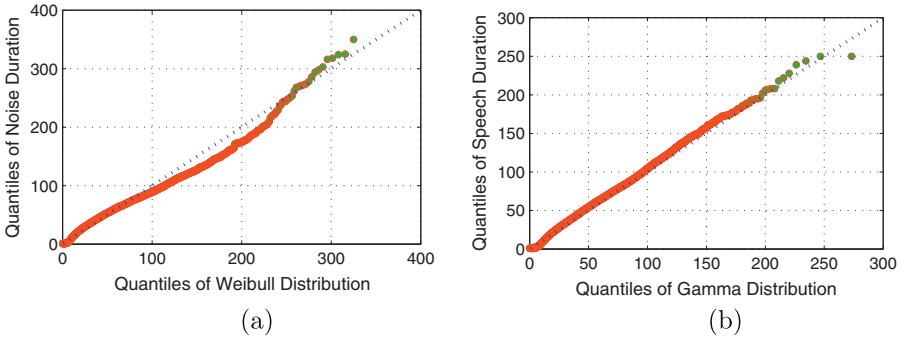

Fig. 3. Quantile–quantile fit: (a) noise duration and (b) speech duration.

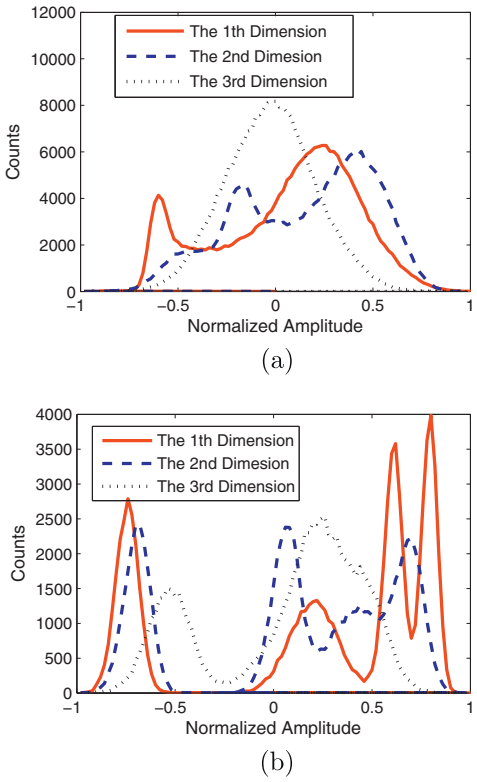

Fig. 4. Counts of MFCCs of: (a) speech and (b) noise.

3.3. Likelihood ratio test

The VAD decision is made according to likelihood ratio test (LRT) depending on a certain threshold. Assuming that each frame is a mixture of speech and uncorrelated additive noise, there are two hypotheses for a VAD to consider for each frame:

$$\begin{cases} H_0 : \textit{speech absent} \\ H_1 : \textit{speech present} \end{cases} \tag{9}$$

To detect which hypothesis holds, a real-time decision of hidden state for each frame can be derived from the likelihood ratio test:

$$LRT(t) = \frac{P(O_1^t|S_t = q_1, \lambda)}{P(O_1^t|S_t = q_0, \lambda)} = \frac{P(S_t = q_0|\lambda)}{P(S_t = q_1|\lambda)} \frac{P(O_1^t, S_t = q_1|\lambda)}{P(O_1^t, S_t = q_0|\lambda)}, \tag{10}$$

where $\frac{P(S_t=q_0|\lambda)}{P(S_t=q_1|\lambda)}$ is the prior probability ratio and $\frac{P(O_1^t,S_t=q_1|\lambda)}{P(O_1^t,S_t=q_0|\lambda)}$ is equal to the posteriori probability ratio. We further assume the stationarity of the HSMM to have $P(S_t = q_i—\lambda) = P(H_i)$ where $p(H_0) + p(H_1) = 1$.

Table 1
Flow of the proposed VAD algorithm.

Step 1: Build HSMM $\lambda = (A, B, \tau, \pi)$ in VAD
Step 2: Estimate B and π with TIMIT using a maximum likelihood estimator
Estimate τ with subset of NOIZEUS using EM algorithm
Step 3: Reduce noise with an instantaneous Wiener filter
Step 4: Extract MFCCs of input signals, set $t = t + 1$ (t is initialized with 0)
Step 5: If $t < P$, go back to step 3
If $t = P$, calculate $LRT(t)$ for the first P frames;
Initialize the threshold η with the mean value of $LRT(t)$
Set the first P frames VAD tags 0; go to step 3
If $P < t \leqslant T$, go to step 6
If $t > T$, quit
Step 6: Calculate $\alpha_t(i)$ and further calculate the $LRT(t)$ for frame t
Step 7: Compare $LRT(t)$ with η, if $LRT(t) > \eta$, set VAD tag 1, else set tag 0
Step 8: Update the parameters of B and update threshold η; go to step 3

Define modified forward variables $\alpha_t(i) = P(O_1^t, S_t = q_i|\lambda)$ of HSMM for probabilities of observation sequence with either state. As presented next, an efficient way can be applied to recursively calculating $\alpha_t(i)$ efficiently.

Therefore, the likelihood ratio can be rewrote as:

$$LRT(t) = \frac{P(H_0)}{P(H_1)} \frac{\alpha_t(1)}{\alpha_t(0)} \tag{11}$$

The iteration can be conducted according to the above formula until $\alpha_T(i)$ is computed. Then a VAD decision can be made based on LRT. If $LRT(t) \geqslant \eta$ where η is a threshold, then H_1 holds; otherwise H_0 holds. Table 1 outlines the proposed VAD algorithm step by step. Details about how to choose η adaptive to different environment will be presented in Section 3.5.

3.4. Forward variables

Similar to forward variables of HMMs (Yu and Kobayashi, 2003), modified forward variables $\alpha_t(i) = P(O_1^t, S_t = q_i|\lambda)$ are defined from local respects as discussed in Section 2. In HSMM-based VAD they represent the probabilities of the observation sequence O_1^t with state q_i given parameters λ of the HSMM:

$$\begin{aligned} \alpha_t(i) = P(O_1^t, S_t = q_i|\lambda) &= \sum_{d=1}^{D} P(O_1^t, S_t = q_i, D_t = d|\lambda) \\ &= \sum_{d=1}^{D} \sum_{d'=0}^{d} P(O_1^t, S_t = q_i, D_t = d, D_t^* = d'|\lambda) \\ &= \sum_{d=1}^{D} \sum_{d'=0}^{d} \sum_{j \neq i} \alpha_{t-d'}^*(j) a_{ji} P(d|q_i) \Pi_{s=t-d'+1}^{t} b_i(O_s). \end{aligned} \tag{12}$$

where $t = 1, 2, \ldots, T$, $i = 0, 1$, D_t and D_t^* are the duration that current state lasts and the duration that current state has last till the

moment t, and $\alpha_t^* = P(O_1^t, S_t = q_i | S_{t+1} \neq q_i, \lambda)$ that means at moment t duration in q_i will finish, namely $D_t = D_t^*$ (Murphy, 2002). According to this definition, α_t^* can be derived from the following recursive equation:

$$\begin{aligned}\alpha_t^*(i) &= P(O_1^t, S_t = q_i | S_{t+1} \neq q_i, \lambda) = \sum_{d=1}^{D} P(O_1^t, S_t = q_i, D_t \\ &= d | S_{t+1} \neq q_i, \lambda) = \sum_{d=1}^{D} P(O_1^t, S_t = q_i, D_t = d, D_t^* = d | \lambda) \\ &= \sum_{d=1}^{D} \sum_{j \neq i} \alpha_{t-d}^*(j) a_{ji} P(d|q_i) \Pi_{s=t-d+1}^{t} b_i(O_s). \end{aligned} \tag{13}$$

Initially $\alpha_t^*(i)$ are initialized as follows:

$$\alpha_t^*(i) = \begin{cases} \pi_i P(d = 1|q_i) b_i(O_1), & t = 1 \\ 0, & t \leqslant 0. \end{cases} \tag{14}$$

Since α_t^* can be easily obtained by iteration according to (13), $\alpha_t(i)$ will be efficiently calculated. Theoretically, by limiting the maximum duration to D, the time complexity will be $O(N^2DT)$ (Yu and Kobayashi, 2003). We can see this modified variable brings no much more complexity than inference of HMMs. Performance comparison between modified and traditional forward variables is made in Section 4.

3.5. Adaptive LRT threshold

After obtaining $\alpha_t(i)$ with (12), *LRT* of every frame can be calculated by (11), and then it is compared with the threshold $\eta(t)$ to make the final decision. In order to make the algorithm more robust and adaptive to the varying noise environment, $\eta(t)$ should take both historical and current *LRT* of noise into consideration. Inspired by the profile of *LRT* in Fig. 5(b), we devise a dynamical η which is initialized as 0 and adaptive to the current *LRT* value from the average of historical *LRT* values of noise frames. We assume that the first P frames are noise frames, and simply adjust η for $t > P$ ($P = 15$ in this paper) as follows:

$$\begin{cases} \eta(t+1) = \rho_i v(t) + (1 - \rho_i) LRT(t), & t > P \\ \eta(t) = v(t), & t \leqslant P \end{cases} \tag{15}$$

where

$$i = I_{(\eta(t),\infty)}(LRT(t)) \tag{16}$$

and

$$v(t) = \frac{1}{t} \sum_{0}^{t-1} LRT(t), \quad t \geqslant 1 \tag{17}$$

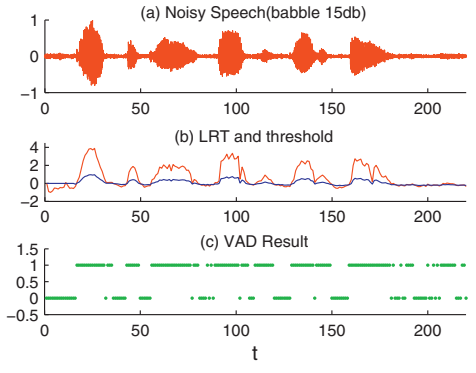

Fig. 5. Likelihood ratio and threshold of noisy speech (babble 15 dB). (a) Normalized amplitude of raw speech; (b) the *LRT* and threshold of the speech and (c) detection results, 1 represents speech, 0 represents non-speech.

I is an indicator function and $LRT(0) = 0$. Values of ρ_i are listed in Table 3. Here, we choose a larger ρ_0 and smaller ρ_1, because intuitively we think *LRT* values for speech frames have greater effect on η than that for noise frames to make η adaptive to different situations. Also to make η adjusted slightly and thus stable for long speech, both values of ρ_i should be set to keep $1 - \rho_i$ small (In this paper $\rho_i > 0.5$). As shown in Fig. 5, the threshold can closely approximate the real *LRT* value of noise frames and dynamically varies as *LRT* changes.

4. Experiments

4.1. Data

A noisy speech corpus (NOIZEUS) (Hu and Loizou, 2008) is used to evaluate the proposed algorithm. NOIZEUS contains a set of thirty 8 kHz voice streams from three male and three female speakers. Thirty sentences were selected from the IEEE database so as to include all phonemes in the American English language, and were corrupted by eight different real-world noises at different average SNR levels between 15 dB and 0 dB. The noise was taken from the AURORA database and includes train, babble, car, and street noise. We made reference decisions for a clean speech material by labeling manually at every 10 ms frame.

First, a Wiener filter is used to reduce the additive noise to the signals. Wiener filter is widely used in signal processing to reduce the amount of noise presented in a speech by comparing with an estimated noiseless speech. In this paper, to meet the real-time demand, we adopt an instantaneous Wiener filter designed by ICSI, OGI, and Qualcomm (Adami et al., 2002). Second, a MFCC processor is applied to extract acoustic features of the filtered speech. The feature extraction conditions are listed in Table 2.

Two evaluations were conducted. First, a series experiments were conducted to evaluate the effects brought by features choosing, duration distributions modeling and observation distributions modeling. We also observed the effects of forward variables modification, and noise filtering. Second, performance comparisons were made between the proposed VAD system and G.729B (Benyassine et al., 1997), encoder Adaptive MultiRate VAD phase 2 (AMR2) (ETSI, 1998), Advanced Front-end (AFE) (ETSI, 2002), HMM-based (Sohn et al., 1999) and LD–GD based (Gazor and Zhang, 2003) VAD systems.

4.2. Parameter estimation

In previous section, we elaborate the setup of HSMM $\lambda = (A, B, \pi, \tau)$ in VAD. Before evaluating HSMM-Based VAD, first discuss its parameter estimation in detail.

4.2.1. State probabilities and duration density estimation

The initial π_i can be estimated from state duration statistics of TIMIT. From the stochastic viewpoint, speech/non-speech states transition can be regarded as an on–off process according to which the probability π_i in state q_i can be estimated by $\frac{m_i}{\sum_j m_j}$ initially, where m_i ($m_0 = \omega_0 \Gamma(1 + \frac{1}{k_0})$ and $m_1 = \omega_1 k_1$) are average durations of noise and speech, respectively.

Table 2
Feature extraction conditions.

Sampling frequency	8000 Hz (16 bit)
Pre-emphasis	$1 - 0.97z^{-1}$
Feature parameters	20 Mel cepstral coefficients
Frame length	10 ms
Window type	Hamming
Filterbank type	triangular

Table 3
HSMM parameter estimation.

i	π_i	k_i	ω_i	ρ_i
0	0.49	19.152	1.021	0.95
1	0.51	0.979	20.763	0.75

To calculate π_i, we need to estimate parameters k_i and ω_i of WD and ΓD. One popular criterion of goodness is to maximize the likelihood function (Nelson, 2004). Maximum likelihood estimation (MLE) involves calculating the values of the parameters that give the highest likelihood given the particular observations, namely state durations here. The maximum likelihood estimators of k_i and ω_i on TIMIT are achieved shown in Table 3.

4.2.2. Observation conditional distribution estimation

To calculate the parameters θ of observation PDF, we also adopt MLE to find a solution satisfying:

$$\theta = \arg\max_{\theta} \prod_{t=1}^{T} P(O_t|q_i;\theta) \tag{18}$$

Firstly, k-means algorithm is used to divide the labeled training data of NOISEUS into M clusters for state q_i to get the initial values for μ_{ij} and σ_{ij} of each cluster, and w_{ij} is initialized with $1/M$. As a tradeoff between the performances and computational cost, in practice M is set with 8, which is sufficient to model observation distributions according to our observations. σ_{ij} is chosen to be a diagonal matrix based on the empirical evidence that diagonal matrices outperform full matrices and the fact that the density modeling of an Kth order full covariance mixture can equally be achieved using a larger order, diagonal one (Reynolds, 1995).

After initiation, EM algorithm is used to re-estimate the parameters and 10 iterations are used for parameter convergence (Reynolds, 1995). For GMD, EM-estimation yields following equations:

$$\begin{aligned}
\boldsymbol{\mu}_{ij} &= \frac{\sum_{t=1}^{T}\psi_{ij}(O_t)O_t}{\sum_{t=1}^{T}\psi_{ij}(O_t)} \\
\Sigma_{ij} &= \frac{\sum_{t=1}^{T}\psi_{ij}(O_t)(O_t-\boldsymbol{\mu}_{ij})^T(O_t-\boldsymbol{\mu}_{ij})}{\sum_{t=1}^{T}\psi_{ij}(O_t)} \\
w_{ij} &= \frac{1}{T}\sum_{t=1}^{T}\psi_{ij}(O_t)
\end{aligned} \tag{19}$$

where $\psi_{ij}(O_t)$ is the posterior probability of the jth Gaussian component for state q_i:

$$\psi_{ij}(O_t) = \frac{w_{ij}\mathcal{N}(O_t;\boldsymbol{\mu}_{ij},\Sigma_{ij})}{\sum_{k=1}^{M} w_{ik}\mathcal{N}(O_t;\boldsymbol{\mu}_{ik},\Sigma_{ik})} \tag{20}$$

4.3. Metrics

For VAD, two error types exist: false miss and false alarm. False miss refers to true speech frames misclassified as non-speech by the VAD. False alarm happens when the VAD declares a frame as speech but it is non-speech. In this paper, the performance is first evaluated in terms of false missed (clipping) rate P_c, and false alarmed (false detection) rate P_e:

$$\begin{aligned}
P_c &= \frac{\#\ \text{speech frames that are mistakenly classified as noise}}{\#\ \text{total speech frames}} \\
P_e &= \frac{\#\ \text{noise frames that are mistakenly classified as speech}}{\#\ \text{total noise frames}}
\end{aligned} \tag{21}$$

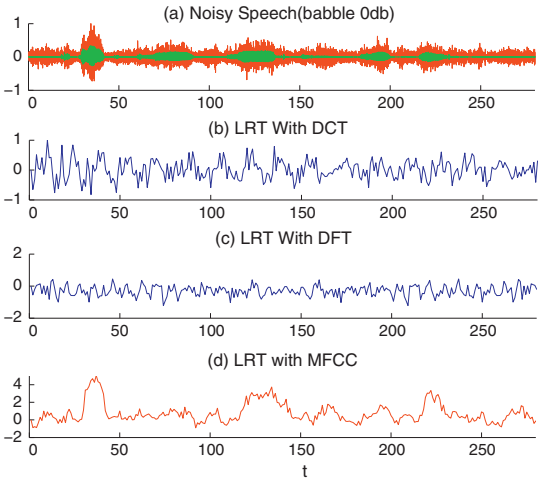

Fig. 6. Likelihood ratio test on noisy speech. (a) Normalized amplitude of clean and noisy speech; (b) LRT in DCT domain; (c) LRT in DFT domain and (d) LRT in MFCC domain.

P_c and P_e have a trade-off relationship and are controlled by the threshold or certain parameters. So we use Detection Error Tradeoff (DET) (Martin et al., 1997) as a measure tool. The DET plot shows the probability of false miss (P_{fm}) as a function of the probability of false alarm (P_{fa}) on a normal deviate scale. It is powerful to discriminate detection or classification performance in the literature.

4.4. Different aspects evaluation

4.4.1. Feature

We conducted a series of experiments with DCT, DFT coefficients and MFCC respectively to evaluate their performances in the VAD algorithm. Fig. 6 shows that the curve of *LRT* in MFCC domain is much closer to the shape of the clean speech spectra (the green curve in Fig. 6(a)) than those in DCT and DFT domains. We analyzed the correlation r between different LRT value sequences and absolute amplitudes sampled periodically from the clean speech, where LRT with MFCC gives much higher correlation to clean speech ($r = 0.5132$) than the other two (0.1210 and 0.0799). Because the closer the profile of *LRT* is to that of clean speech amplitude, the larger the degree is to which it can reflect the dis-

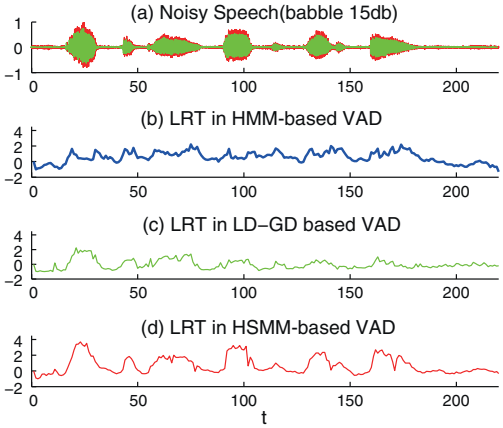

Fig. 7. Likelihood ratio test on noisy speech. (a) Normalized amplitude of clean and noisy speech; (b) LRT in HMM-based VAD; (c) LRT in LD–GD based VAD and (d) LRT in proposed HSMM-based VAD.

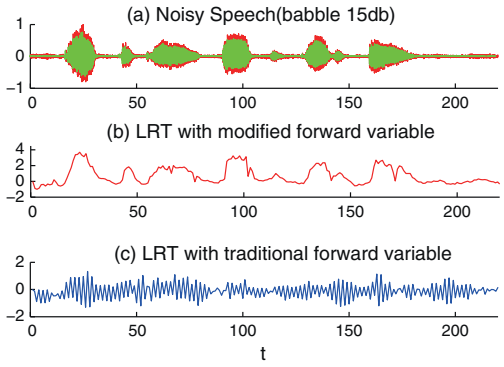

Fig. 8. Likelihood ratio test with: (a) modified forward variable and (b) traditional forward variable.

crimination between noise and speech, therefore the better the detection results will be. So LRT with MFCC as speech features can easily and accurately make VAD decisions on noisy speech as energy-based methods do on clean speech. We believe that this happens partially because MFCC can capture the important characters of speech which is usually invariant to the noise. We use MFCC as the speech feature in the following experiments.

4.4.2. Duration distribution

To study how the duration distribution influences the accuracy of VAD, we compared the proposed VAD with a HMM-based VAD which uses GED for duration distribution. In this control experiment, all the parameters and processes are the same except the state transition probabilities. For HMMs, these probabilities are set as $A = [0,2,0.8;0.1,0.9]$ (Fujimoto et al., 2007) that gives the best performance; for HSMM, the probabilities are calculated by WD and ΓD. With the assumption that the duration is unaffected by noise, parameters of WD and ΓD are trained with a subset of TIMIT database containing 4400 clean utterance spoken by 440 speakers from eight major dialect regions of the United States.

We compared the profiles of *LRT* and the VAD results between HMM-based and proposed VAD. Similar to Figs. 6, 7(b) and (d) show that with the same observation PDF, the *LRT* profile of the

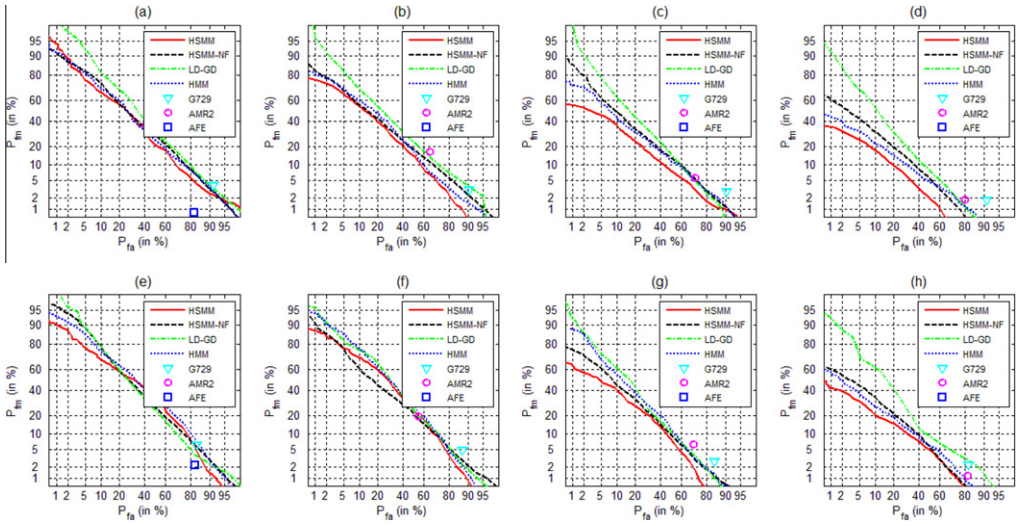

Fig. 9. Comparison of different VADs: (a) babble 0 dB; (b) babble 5 dB; (c) babble 10 dB; (d) babble 15 dB; (e) train 0 dB; (f) train 5 dB; (g) train 10 dB and (h) train 15 dB.

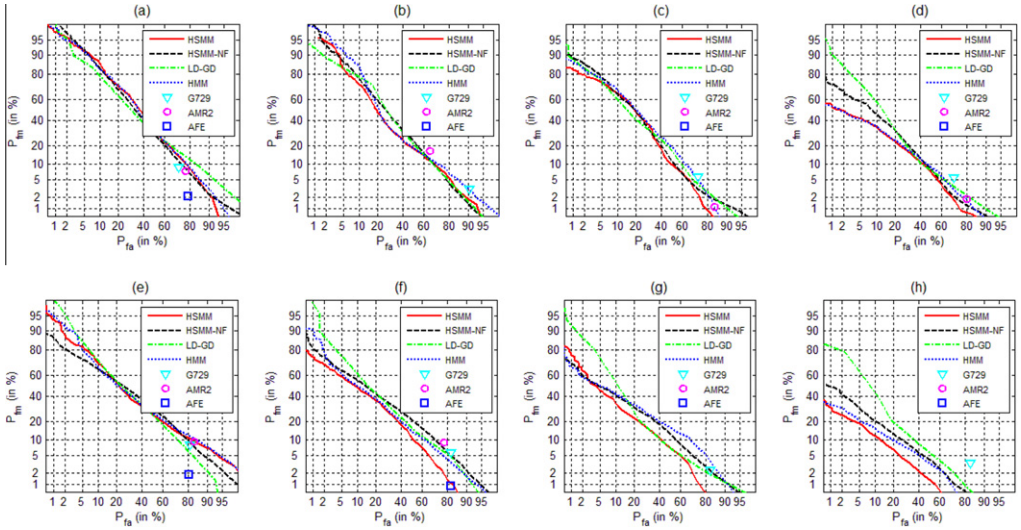

Fig. 10. Comparison of different VADs: (a) Car 0 dB; (b) Car 5 dB; (c) Car 10 dB; (d) Car 15 dB; (e) Street 0 dB; (f) Street 5 dB; (g) Street 10 dB and (h) Street 15 dB.

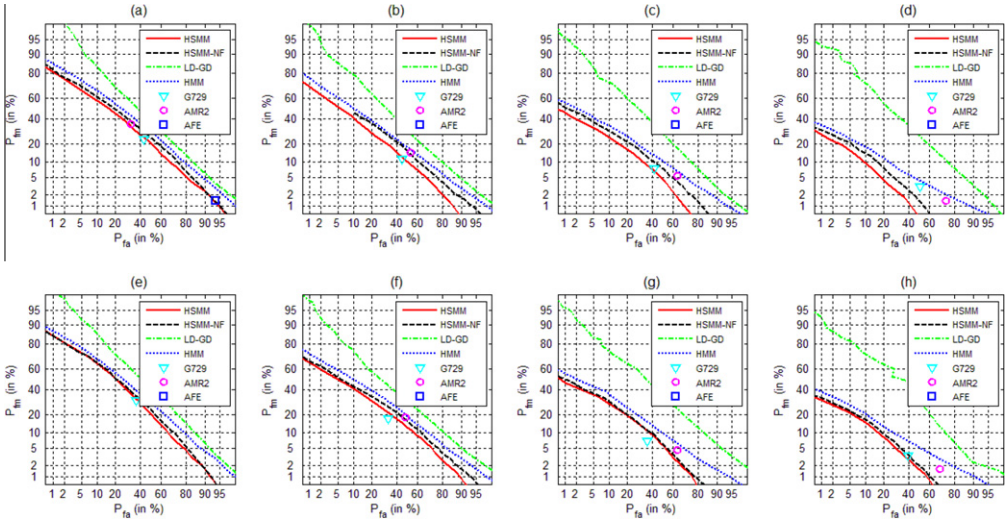

Fig. 11. Comparison of different VADs on long utterances: (a) babble 0 dB; (b) babble 5 dB; (c) babble 10 dB; (d) babble 15 dB; (e) train 0 dB; (f) train 5 dB; (g) train 10 dB and (h) train 15 dB.

proposed VAD is closer to that of the clean speech amplitude (green curves in Fig. 7(a)). Also in Fig. 9, DET curves of the proposed VAD much closer to the low left corner, which means that the proposed VAD outperforms HMM-based VAD. So we can safely conclude that WD and ΓD (HSMM-based VAD) can more accurately model the duration distribution than GED (HMM-based VAD).

4.4.3. Observation distribution

We also compared the proposed VAD with LD–GD based VAD to study the effects of modeling observation distribution. In LD–GD based VAD, LD and GD are adopted for speech and noise state respectively, all the other conditions are the same with proposed VAD. Assuming that observation distributions will be affected by noise, noisy data set is used to train the their parameters. We divided NOIZEUS into two parts: the training part is comprised of the first five speakers' utterances corrupted with babble, train, car, and street noises at 0 dB, 5 dB, 10 dB, and 15 dB; the rest utterances are used as the testing data. Fig. 7(c) and (d) shows that with the same state duration PDF, the *LRT* profile of the proposed VAD can better discriminate the noise and speech. Moreover, the VAD results in Fig. 9 show that the proposed algorithm has higher VAD accuracy than LD–GD based VAD, except the VAD results for train 0 dB. These improvements are obvious with different SNRs.

4.4.4. Forward variables

To observe the performance of the modified forward variable, a comparison was made between the likelihood tests with modified and traditional forward variables. The results in Fig. 8 show that using modified variable improve the performance of proposed VAD in likelihood ratio test, indicating that modified variable is more accurate for HSMM to calculate the probability of observation sequence given a state.

4.4.5. Noise filtering

Finally evaluation on how much the proposed VAD gains from applying Wiener filter was made. We compared the performance of the proposed algorithm with and without filtering. Curves labeled "HSMM" (with filtering) and "HSMM-NF" (without filtering) in Figs. 9–11 show that by filtering noises, the VAD accuracy has been generally improved, especially at the high SNRs. But for low SNRs (0 dB and 5 dB), "HSMM-NF" sometime outperforms "HSMM". This is because that wiener filter can better estimates the noise power at high SNRs and thus eliminate the additive noise, while for low SNRs, noise might be mistaken as speech and speech might be filtered as noise, which degrades detection performance. The average detection results between HSMM-based algorithm with and without filtering are listed in Table 4.

Comparing to the VAD results without noise filtering before feature extraction, both P_e and P_c in the VAD using Wiener filtering are reduced, especially P_c decreased by at least 21%. Thereby, it can be concluded that noise filtering helps to improve the accuracy of the proposed VAD.

4.5. Algorithm comparison

In the above we have studied several aspects in the proposed algorithm. Next, VAD performance comparison is made among the proposed VAD, AFE VAD, G.729B VAD, AMR2 VAD, HMM-based VAD and LD–GD based VAD. The G.729B performs on 80 samples/frame, and AMR2 is 160 samples/frame. For all the algorithms, wiener filter is first used to reduce noise.

Table 5 shows the average performance of the proposed VAD compared with AFE, G.729B, AMR2, HMM-Based and LG–DG based VAD. On average the proposed VAD owns a lower false missed rate P_c (with 46.0% improvement over G.729B and 75.45% over AMR), and improves the false alarmed rate P_e by 18.37% and 8.33% than G.729B and AMR2 respectively. Though AFE has the lowest P_c, the P_e of 82.8% means it almost classifies all frames to speech, making it impractical for applications in noisy environments. HMM-based VAD owns the lowest P_e, however this happens at the cost of high false missed rate, which should be maintained low to avoid speech loss. So HMM-based VAD is inappropriate for the applications concerning about speech signals. LD–GD based VAD performs generally better than AMR2, G.729B, HMM-based VAD on average

Table 4
HSMM-based VAD results.

Noise	VAD with filter		VAD without filter		Improvement (%)	
	P_c	P_e	P_c	P_e	P_c reduce	P_e reduce
Babble	2.4	66.4	4.9	68.0	50.5	2.4
Car	1.7	69.6	3.4	75.5	48.0	7.9
Street	3.4	64.8	6.0	66.5	43.7	2.5
Train	3.3	68.0	4.2	70.8	21.8	3.9

Table 5
Performance of the proposed VAD compared with G.729B, AMR2, HMM-based and LD–GD based VAD.

Environment		AFE		G.729B		AMR2		HMM		LD–GD		Proposed	
Noise	SNR	P_c	P_e	P_c	P_e	P_c	P_e	P_c	P_e	P_c	P_e	P_c	P_e
Babble	0	0.8	82.1	3.8	91.4	41.6	37.2	10.0	64.6	4.5	73.7	3.5	75.7
	5	0.3	82.6	3.0	90.6	16.6	64.8	5.9	62.6	5.1	70.5	2.2	69.1
	10	0.0	84.0	2.8	90.5	5.7	71.9	5.9	52.9	2.5	70.3	2.6	65.7
	15	0.0	84.7	1.7	91.1	1.8	80.0	6.4	36.9	2.2	66.0	1.4	55.1
Car	0	2.1	78.7	13.6	68.6	22.1	55.5	7.2	71.6	7.8	68.4	3.4	78.9
	5	0.2	83.0	8.7	71.8	7.6	76.8	3.5	66.8	5.4	64.4	1.0	74.1
	10	0.0	82.3	5.7	72.9	1.1	84.4	4.8	54.8	3.0	63.3	2.0	62.2
	15	0.0	83.8	5.4	70.5	0.4	89.3	1.4	57.8	2.3	61.8	0.5	63.1
Street	0	1.8	80.5	8.2	81.1	9.3	83.8	16.9	56.1	5.2	73.2	6.5	65.9
	5	1.0	82.5	5.8	82.0	8.9	77.6	8.2	61.2	3.5	71.3	3.4	71.1
	10	0.0	83.5	2.5	83.0	0.1	99.1	6.2	52.1	3.9	70.4	2.3	58.7
	15	0.0	83.5	3.5	83.5	0.5	93.2	4.5	46.9	2.1	66.5	1.3	63.2
Train	0	2.2	82.5	6.2	83.9	32.6	51.2	17.5	56.5	5.6	70.5	7.8	67.9
	5	0.5	82.5	4.8	87.3	19.4	54.9	7.4	64.9	4.3	69.3	2.9	73.2
	10	0.0	85.1	2.7	84.8	6.3	69.9	6.6	54.1	2.7	69.5	1.5	64.1
	15	0.0	85.3	2.2	82.4	1.2	81.6	3.1	59.7	2.2	68.5	0.9	66.9
Average	–	0.56	82.8	5.0	82.2	11.0	73.2	7.2	57.5	3.9	68.6	2.7	67.1

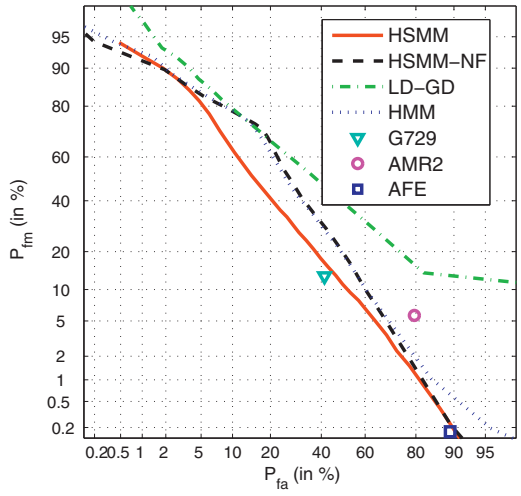

Fig. 12. Comparison of different VADs on mixed utterances.

P_c and P_e (except for a higher P_e than HMM-based VAD), but worse than the proposed VAD.

Figs. 9 and 10 show the performance comparison on data of different noise types and SNR levels. For most cases, curves of HSMM-based VAD are the closest ones to the left low corner, which means that the proposed algorithm gives the best performance. Moreover, in some cases even the HSMM-based VAD without noise filtering can work better than other VADs with filtering. Similar to the conclusion from Table 5, although sometimes AFE achieves a performance point much closer to the left low corner, the probability of false alarm is much higher (note that, due to this reason some points of AFE fall outside the figures), which is undesired for applications in noisy environments.

To verify the performance on long utterance, we conducted two experiments on the utterances of 50–80 s. The configurations including both parameters and the process are the same with the experiments above. The data set for the first experiment contains eight long utterances of 80 s, each of them is constructed by joining 30 short utterances with the same noise at the same SNR from NOIZEUS. The data set for the second experiment contains three utterances (52, 55 and 56 s) randomly concatenated from the data with different noises at different SNRs in NOIZEUS. Detection results, that HSMM-based VAD outperforms all the other algorithms in this paper, respectively shown in Figs. 11 and 12 are consistent with those in Fig. 9, which indicates that the proposed algorithm can work both well and stably for both short and long utterances.

On the whole, the proposed VAD provides a favorable performance: (1) an improvement due to explicitly modeling the duration distribution, which means that the transmission probability modeled by WD and ΓD in the HSMM is more accurate than the constant transmission probability used in HMM; (2) lower false alarmed and false missed rates than LD–GD based VAD, reflecting GMD can better approximate the observation distribution of noisy speech than LD–GD model; (3) considerable performance gain from modified forward variable and noise filtering; (4) achieves an overall performance improvement to G.729B, AMR2, AFE, HMM-based and LD–GD based VAD systems.

5. Conclusion

In this paper, we focus on three robustness related aspects of VAD including speech features, feature distributions, and temporal dependence. To integrate these aspects in VAD, we proposed a novel VAD algorithm based on HSMM combined with LRT. This is the first work that explicitly models speech durations by combining HSMM and LRT for VAD. We apply an instantaneous Wiener filter to noise reduction, and then use a MFCC processor to extract features. Weibull and Gamma PDFs are applied to non-speech and speech duration distributions respectively, and Gaussian Mixture PDFs are used for both speech and non-speech states for modeling observation distributions. A series of experiments show that these schemes help to improve the accuracy of VAD in noisy environment respectively, and the whole system achieves better performance than G.729B, AMR2, AFE, HMM-based and LD–GD based VAD.

Further work can concentrate on the analysis of the duration mechanism in real world, and the calculation of optimal thresholds for likelihood ratio tests.

Acknowledgment

The authors thank reviewers for their valuable comments. They also would like to thank Professor Shun-Zheng Yu of Sun Yat-Sen

University, Jared O'Connell of Aarhus University, Storsjö Martin from the Opencore-amr-devel mailing list, and Wei Su and Zhao Wang of Beihang University for discussion on forward variables and implementation.

This work is supported by the National High-Technology Research and Development Program (2007AA010301-02 and 2009AA043303) and National Major Project of China "Advanced Unstructured Data Repository" (2010ZX01042-002-001-00).

References

Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivadas, S., 2002. Qualcomm-ICSI-OGI features for ASR. In: Proc. ICSLP, vol. 1, pp. 4–7.

Benyassine, A., Shlomot, E., Su, H.Y., 1997. ITU Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. IEEE Comm. Mag., 64–73.

Chang, J.H., Kim, N.S., Mitra, S.K., 2006. Voice activity detection based on multiple statistical models. IEEE Trans. Signal Process. 54 (6).

Chen, M.Y., Kundu, A., Srihari, S.N., 1995. Variable duration hidden Markov model and morphological segmentation for handwritten word recognition. IEEE Trans. Image Process. 4 (12).

Davis, A., Nordholm, S., Togneri, R., 2006. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. IEEE Trans. Audio Speech Lang. Process. 14 (2).

Dong, M., He, D., 2007. A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology. Mech. Systems Signal Process. 21 (5), 2248–2266.

ETSI, 1998. Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels. General description. European Standard (Telecommunications series), GSM 06.94 version 7.1.1.

ETSI, 2002. Speech processing, transmission and quality aspects (STQ). Distributed speech recognition; Advanced front-end feature extraction algorithm. Compression algorithms. ETSI, Sophia Antipolis, France, ETSI ES 202 050 Rec.

Fujimoto, M., Ishizuka, K., Kato H., 2007. Noise robust voice activity detection based on statistical model and parallel non-linear Kalman filtering. In: Proc. ICASSP (4), pp. 797–800.

Fujimoto, M., Ishizuka, K., 2008. Noise robust voice activity detection based on switching Kalman filter. IEICE Trans. Inform. Systems E91-D (3).

Garofolo, J.S., Lamel, L.F., et al., 1993. DARPA TIMIT acoustic–phonetic continuous speech corpus. US Dept. of Commerce, NIST, Gaithersburg, MD, USA.

Gazor, S., Zhang, W., 2003. A soft voice activity detector based on a Laplacian–Gaussian model. IEEE Trans. Speech Audio Process. 11 (5), 498–505.

Hu, Y., Loizou, P., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Speech Audio Process. 16 (1), 229–238.

Liang, Y., Liu, X.L., Zhou, M., 2010. A Robust voice activity detector based on Weibull and Gaussian mixture distribution. In: Proc. ICSPS, pp. V2.26-28.

Liu, X.L., Liang, Y., Lou, Y.H., Li, H., Shan, B.S., 2010. Noise-robust voice activity detector based on hidden semi-Markov models. In: Proc. ICPR, pp. 81–84.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki M., 1997. The DET curve in assessment of detection task performance. In: Proc. Eur. Conf. on Speech Communication and Technology, pp. 1895–1898.

Murphy, K.P., 2002. Hidden semi-Markov models (HSMMs). unpublished notes.

Nelson, W., 2004. Applied Life Data Analysis. Wiley-Blackwell. pp. 313–346.

Othman, H., Aboulnasr, T., A semi-continuous state transition probability HMM-based voice activity detection. In: ICASSP, 2004, pp. 821–824.

Rabiner, L.R., 1989. A tutorial on hidden Markov model and selected applications in speech recognition. IEEE Proc. 77, 257–286.

Reynolds, D.A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. 17 (1), 91–108.

Shin, J.W., Chang, J.H., Kim, N.S., 2007. Voice activity detection based on a family of parametric distributions. Pattern Recognition Lett. 28 (11), 1295–1299.

Sohn, J., Kim, N.S., Sung, W.Y., 1999. Statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (1).

Tan, Z.H., Lindberg, B., 2010. Low-complexity variable frame rate analysis for speech recognition and voice activity detection. IEEE J. Select. Top. Signal Process. 4 (5), 798–807.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition, IINOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Comm., vol. 12, pp. 247–251.

Yu, S.Z., Kobayashi, H., 2003. An efficient forward–backward algorithm for an explicit duration hidden Markov model. IEEE Signal Process. Lett. 10 (1), 11–14.