Noise-Robust Voice Activity Detector Based on Hidden Semi-Markov Models

Xianglong Liu*, Yuan Liang, Yihua Lou, He Li, Baosong Shan
State Key Laboratory of Software Development Environment
Beihang University, Beijing 100191, P.R.China

Abstract

This paper concentrates on speech duration distributions that are usually invariant to noises and proposes a noise-robust and real-time voice activity detector (VAD) using the hidden semi-Markov model (HSMM) to explicitly model state durations. Motivated by statistical observations and tests on TIMIT and the IEEE sentence database, we use Weibull distributions to model state durations approximately and estimate their parameters by maximum likelihood estimators. The final VAD decision is made according to the likelihood ratio test (LRT) incorporating state prior knowledge and modified forward variables. An efficient way that recursively calculates modified forward variables is devised and a dynamic adjustment scheme is used to update parameters. Experiments on noisy speech data show that the proposed method performs more robustly and accurately than the standard ITU-T G.729B VAD and AMR2.

1 Introduction

Voice activity detector (VAD) refers to the system of distinguishing active speech from non-speech frames. It has been used for various real-time applications like speech coding and recognition. Recently many attractive statistical model-based VADs using the likelihood ratio test (LRT) have been developed [1]. The statistical methods using hidden Markov models (HMMs) as a hangover scheme have made significant contributions to voice activity detection progress [2, 3]. However one major drawback of HMMs is that they might not provide an adequate representation of the temporal structure of speech that is usually unaffected by noises [3, 4].

To alleviate this limitation and to improve the noise-robustness, we concentrate on speech durations. We view the signal composed of speech and noise as a time duration hidden markov chain with two states (speech and non-speech) and propose a novel VAD algorithm based on the hidden semi-Markov model (HSMM) [5] to explicitly model the state duration invariant to noises. Statistical observations on TIMIT and IEEE sentence database [6] show that the state duration follows the Weibull distribution. For final VAD decisions, we adopt likelihood ratio test (LRT) combining both state prior knowledge and probabilities of observation sequence with either state (named modified forward variables). The proposed method derives an efficient way similar to Vertibi of HMMs to recursively calculate the modified forward variables, and dynamically adjusts parameters to improve robustness to the varying noisy environment.

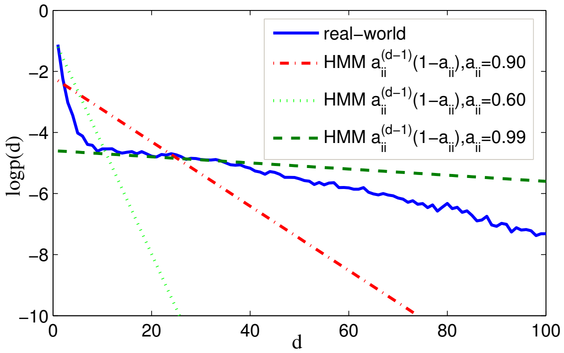

Figure 1. Duration distribution of a TIMIT subset compared with HMMs.

This paper is organized as follows. In Section 2 we present the segmental HSMM-based modeling framework for VAD. Section 3 gives implementation details including dynamical parameter adjustment and modified forward variables calculation. Experiments with HSMM-based VAD are provided in Section 4. Finally, conclusions are drawn in Section 5.

2 HSMM-Based Voice Activity Detection

Modern VAD algorithms' hangover schemes using HMMs implicitly describe the state duration effect on likelihood of state transition. Their geometric state du-

*Supported by grants from the National High Technology Research and Development Program of China (No.2009AA043303) and the Major State Basic Research Development Program of China (No.2005CB321903)

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.28

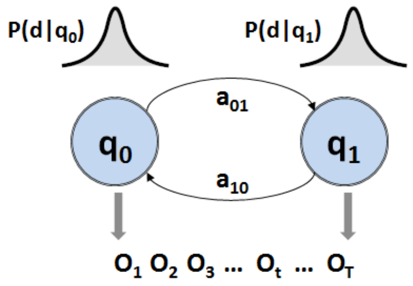

Figure 2. A hidden semi-Markov model with two states.

ration probabilities $P(d|q_i) = a_{ii}^{d-1}(1 - a_{ii})$ [4] are inappropriate for most real-life applications [4]. Fig. 1 shows state duration distribution of a TIMIT subset compared with that of HMMs. It can be concluded that real-world duration distribution (the solid curve) differs from the geometrical function (dot lines).

In this paper, signals composed of speech and noise are regarded as a time duration hidden markov chain with two states (speech and non-speech) that can be modeled by the HSMM $\lambda = (A, B, \tau, \pi)$. Fig. 2 shows that there exist two states in voice activity detection, named non-speech (or noise) q_0 and speech q_1. Signal frames remaining at the same state can be 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 one varies over its duration which can be modeled by certain distribution $\tau = \{P(d|q_i)\}$ or implicit techniques [3], therefore either state duration reflects the temporal dependence of the state. Next we will present components of the segmental HSMM-based modeling framework λ for VAD.

2.1 The Hidden Layer

Here, the transition matrix $A = (a_{ij})$ is given by

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

To obtain duration distribution τ, one way is to estimate it from the training samples. Statistics of subsets of both TIMIT database and the IEEE sentence database [6] show that the durations of noise and speech states share similar properties and fit the Weibull distribution well. The Weibull distribution, known as the life distribution, is usually employed to analyze life data. The fact indicates that duration of speech can be viewed as a lifetime of vocalization subject to cyclic patterns of stress and strain, which is consistent with the mechanism of human vocalization.

Thus, we choose the Weibull distribution to model state duration of speech and non-speech:

$$P(d|q_i) = \frac{k_i}{\omega_i}(\frac{d}{\omega_i})^{k_i - 1} e^{(\frac{d}{\omega_i})^{k_i}}, \tag{2}$$

where $d > 0$ is the length of duration in state q_i, $k_i > 0$ is the shape parameter and $\omega_i > 0$ is the scale parameter of the distribution.Duration statistics of a TIMIT subset are fitted to Weibull distributions well according to Pearson test. We expect that the HSMM with proper duration distributions will be able to model the signal sequence well.

2.2 The Observation Layer

In the observation layer, the likelihood $B = \{b_i(O_t)\}$ of the observation O_t, namely the frame at moment t, being a speech or a noise frame is concerned. We choose GD to model the spectra for noise and LD for clean speech [2]. The observation O_t consists of K independent DCT coefficients $o_i, (0 < i \leq K)$ whose probability density function conditioned on state q_i are respectively given by

$$\begin{aligned} b_0(o_i) &= p(o_i|S_t = q_0) \\ &= \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(o_i - \mu_i^G)^2}{2\sigma_i^2}}, \end{aligned} \tag{3}$$

$$\begin{aligned} b_1(o_i) &= p(o_i|S_t = q_1) \\ &= \frac{1}{4l_i} e^{\frac{\sigma_i^2}{2l_i^2}} [e^{\frac{o_i'}{l_i}} erfc(\frac{l_i o_i' + \sigma_i^2}{\sqrt{2} l_i \sigma_i}) \\ &\quad + e^{-\frac{o_i'}{l_i}} erfc(\frac{-l_i o_i' + \sigma_i^2}{\sqrt{2} l_i \sigma_i})], \end{aligned} \tag{4}$$

where $o_i' = o_i - \mu_i^G - \mu_i^L$. μ_i^G and σ_i in GD are the mean and variance of o_i, and μ_i^L and l_i in LD are the location parameter and scale parameter. Then given q_i the join distributionfor the observation is $b_i(O_t) = \Pi_{s=1}^{K} b_i(o_s)$.

2.3 Likelihood Ratio Test

For each frame, there are two hypotheses H_0 and H_1 corresponding to either state q_0 and q_1. To detect which hypothesis holds in real time, a decision of hidden state can be derived from likelihood ratio test:

$$\begin{aligned} LRT(t) &= \ln \frac{P(O_1^t|S_t = q_1, \lambda)}{P(O_1^t|S_t = q_0, \lambda)} \\ &= \ln \frac{P(S_t = q_0|\lambda)}{P(S_t = q_1|\lambda)} \frac{P(O_1^t, S_t = q_1|\lambda)}{P(O_1^t, S_t = q_0|\lambda)} \end{aligned} \tag{5}$$

$\frac{P(S_t = q_0|\lambda)}{P(S_t = q_1|\lambda)}$ is the prior probability ratio. We assume the stationarity of the HSMM to have $P(S_t = q_i|\lambda) =$

$P(H_i)$ where $p(H_0) + p(H_1) = 1$. Define modified forward variables $\alpha_t(i) = P(O_1^t, S_t = q_i|\lambda)$ and then:

$$LRT(t) = \ln \frac{P(H_0)}{P(H_1)} \frac{\alpha_t(1)}{\alpha_t(0)} \qquad (6)$$

Then a VAD decision can be made based on LRT. If $LRT(t) \geq \eta$ where η is a dynamic threshold, then H_1 holds; otherwise H_0 holds. Details about how to effectively calculate $\alpha_t(i)$ and how to adjust η adaptive to different environment will be presented in next section.

3 Implementation

This section will discuss details about parameter estimation and modified forward variables calculation.

3.1 Parameter Estimation and Adjustment

To improve the performance of HSMM-based VAD, we need estimate parameters of HSMM λ and LRT threshold η accurately.

1) According to properties of the on-off process $\pi = \{\pi_i\}$ can be estimated by $\frac{\omega_i \Gamma(1+\frac{1}{k_i})}{\sum_j \omega_j \Gamma(1+\frac{1}{k_j})}$ initially.

2) The maximum likelihood estimators for k_i and ω_i of Weibull distributions are achieved shown in Table 1.

3) In order to make the algorithm more robust and adaptive to the varying noise environment, we devise a dynamical parameter adjusting mechanism. The update procedure is summarized in Table 2.

Table 1. Parameter Estimation

i	π_i	k_i	ω_i	ρ_i
0	0.478	11.958	0.679	0.99
1	0.522	12.292	0.677	0.79

Table 2. Parameter adjustment

```
$t > P, i = 1, 2, \ldots, K$
if $S(t) = q_0$ then
    $\mu_i^G = \rho_0 \mu_i^G + (1 - \rho_0) o_i$
    $\sigma_i = \rho_0 \sigma_i + (1 - \rho_0)(o_i - \mu_i^G)(o_i - \mu_i^G)^T$
    $\eta = \rho_0 \eta + (1 - \rho_0) LRT(t - 1)$
else
    $\mu_i^L = \rho_1 \mu_i^L + (1 - \rho_1) o_i$
    $l_i = \rho_1 l_i + (1 - \rho_1)|o_i - \mu_i^L|$
    $\eta = \rho_1 \eta + (1 - \rho_1) LRT(t - 1)$
end
```

Initially $\mu_i^G = \mu_i^L$, σ_i, l_i and η can be estimated from the first P frames. In our evaluation, smaller ρ_1 (see Table 1) is chosen for speech state, because speech can be assumed to be more stationary than noise.

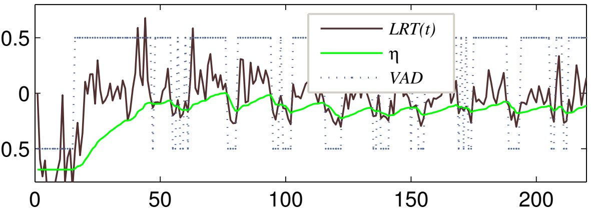

Figure 3. Likelihood ratio test on noisy speech (Babble 15dB).

3.2 Forward Variables

Similar to forward variables of HMMs, modified forward variables $\alpha_t(i) = P(O_1^t, S_t = q_i|\lambda)$ are given by:

$$\alpha_t(i) = \sum_{d=1}^{D} \sum_{d'=0}^{d} \sum_{j \neq i} \alpha_{t-d'}^*(j) a_{ji} P(d|q_i) \Pi_{s=t-d'+1}^{t} b_i(O_s), \qquad (7)$$

where $t = 1, 2, \ldots, T$, $i = 0, 1$, and $\alpha_t^* = P(O_1^t, S_t = q_i | S_{t+1} \neq q_i, \lambda)$ [5] that means at moment t duration in q_i will finish:

$$\alpha_t^*(i) = \sum_{d=1}^{D} \sum_{j \neq i} \alpha_{t-d}^*(j) a_{ji} P(d|q_i) \Pi_{s=t-d+1}^{t} b_i(O_s) \qquad (8)$$

where initially $\alpha_1^*(i) = \pi_i P(d = 1|q_i) b_i(O_1)$.

Since α_t^* can be easily obtained by iteration according to (8), $\alpha_t(i)$ will be be efficiently calculated. Theoretically, by limiting the maximum duration to D, the time complexity will be $O(N^2 DT)$ [7]. We can see this modified variable brings no much more complexity than inference of HMMs.

4 Experiments

A noisy speech corpus (NOIZEUS) [6] is used to evaluate the proposed algorithm. NOIZEUS contains thirty 80kHz voice streams (three male and three female speakers, all phonemes in the American English language) corrupted by different real-world noises (from the AURORA database) including train, babble, car, and street noise at different SNR levels between 15 dB and 0 dB. We made reference decisions for a clean speech material by labeling manually at every 10 ms frame. The percentage of the hand-marked speech frames is 75.13%. In our experiments, the G.729B [8] encoder performs on 80 samples/frame. For Adaptive MultiRate VAD phase 2 (AMR2) [9] 160 samples/frame is used and the VAD option 2 is selected since it performs better than option 1 in our experiment.

Table 3. Performance of the Proposed VAD compared with G.729B and AMR2

Environment		G.729B		AMR2		Proposed	
Noise	SNR	P_c	P_e	P_c	P_e	P_c	P_e
Babble	0	3.8	91.4	41.6	37.2	4.3	69.6
	5	3.0	90.6	64.8	64.8	5.1	69.9
	10	2.8	90.5	5.7	71.9	5.9	68.3
	15	1.7	91.1	1.8	80.0	6.2	68.6
Car	0	13.6	68.6	22.1	55.5	3.7	70.3
	5	8.7	71.8	7.6	76.8	4.3	70.9
	10	5.7	72.9	1.1	84.4	4.6	68.9
	15	5.4	70.5	0.4	89.3	4.8	68.2
Street	0	8.2	81.1	9.3	83.8	4.3	71.4
	5	5.8	82.0	8.9	77.6	4.8	70.1
	10	2.5	83.0	0.1	99.1	4.4	69.5
	15	3.5	83.5	0.5	93.2	5.1	69.7
Train	0	6.2	83.9	32.6	51.2	3.5	71.0
	5	4.8	87.3	19.4	54.9	4.4	70.2
	10	2.7	84.8	6.3	69.9	4.7	68.7
	15	2.2	82.4	1.2	81.6	4.7	68.9

[1] P_c: the ratio of speech frames classified as noise to total speech frames; P_e: the ratio of noise frames classified as speech to total noise frames.

As shown in Fig. 3, the LRT threshold dynamically varies as LRT value changes. Also under different noisy environment, we found that the adaptive threshold can be able to capture the variation of LRT values and tunes its value to approximate a better decision over time.

Table 3 shows the performance of the proposed HSMM-based VAD compared with AMR2 and ITU G.729B VAD. On average G.729B VAD provides the lowest clipping rate over AMR VAD and the proposed VAD. This happens at the cost of higher false detection rate (most over 80% and 84.80% on average). In contrast, the proposed VAD achieves the very closed clipping rate with much lower false detection rate. The proposed VAD owns a lower clipping rate below 4.83% than G.729B VAD in most situations except for Babble noise, and achieves the lowest false detection rate below 70% among the three algorithms. On the whole, the proposed HSMM-based VAD provides a balanced performance: (1) a significant improvement to G.729B and AMR2 for false detection rate; (2) a low clipping rate in most situations especially with low SNR levels.

Furthermore it can be noted that the proposed VAD performs almost equally well under different noises and different SNR levels (about 4.68% clipping rate and 69.63% false detection rate). We believe it is partially due to that the explicit state duration models of the HSMM estimated properly from a TIMIT subset approximates the true speech/non-speech duration distribution invariant to noises in real world. Therefore with such a powerful probabilistic framework it can capture the duration variation in different noisy environments and thus performs robustly. Also the adaptive ability deriving from the dynamic adjusting mechanism contributes to the performance. Therefore, it can be safely concluded that the proposed HSMM-based VAD is robust and adaptive to varying noise environment.

5 Conclusion

In this paper we concentrate on noise-invariant properties of speech like state durations and first propose a novel VAD algorithm based on a HSMM using Weibull distributions. To detect voice activity in real time, a decision rule is made by LRT combining state prior probabilities and modified forward variables which is designed to be recursively calculated. The proposed VAD performs robustly and provides an improvement to standard ITU-T G.729B VAD and AMR2 on the noise data. Further work can concentrate on the duration distribution mechanism and theoretically optimal thresholds.

References

[1] Jongseo Sohn, Nam Soo Kim and Wonyong Sung, ”A Statistical Model-Based Voice Activity Detection,” *IEEE Signal Processing Letters*, Vol. 6, No. 1, Jan. 1999.

[2] S. Gazor and W. Zhang, ”A soft voice activity detector based on a laplacian-gaussian model,” *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 498C505, September 2003

[3] H. Othman and T. Aboulnasr, ”A Semi-Continuous State Transition Probability HMM-Based Voice Activity Detection,” *ICASSP*, 2004

[4] L. R. Rabiner, ”A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition,” *IEEE Proc.*, vol. 77, pp. 257-286. Feb. 1989.

[5] K. P. Murphy, ”Hidden semi-Markov models (HSMMs),” *unpublished notes*, Nov. 2002.

[6] Hu, Y. and Loizou, P., ”Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, 16(1), 229-238, 2008.

[7] S-Z. Yu and H. Kobayashi, ”An Efficient Forward-Backward Algorithm for an Explicit Duration Hidden Markov Model,” *IEEE Signal Processing Letters*, Vol. 10, no. 1, pp. 11-14, January 2003.

[8] Adil Benyassine, Eyal Shlomot, and Huan-Yu Su, ”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.*, pp. 64-73, September 1997.

[9] ETSI, ”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 Release 1998.