A Statistical Model-Based Voice Activity Detection

Jongseo Sohn, *Student Member, IEEE,* Nam Soo Kim, *Member, IEEE,* and Wonyong Sung

***Abstract*— In this letter, we develop a robust voice activity detector (VAD) for the application to variable-rate speech coding. The developed VAD employs the decision-directed parameter estimation method for the likelihood ratio test. In addition, we propose an effective hang-over scheme which considers the previous observations by a first-order Markov process modeling of speech occurrences. According to our simulation results, the proposed VAD shows significantly better performances than the G.729B VAD in low signal-to-noise ratio (SNR) and vehicular noise environments.**

***Index Terms*— Decision-directed estimation, hidden Markov model, likelihood ratio test, voice activity detection.**

I. Introduction

As THE demand for variable-rate speech coding applications increases, the role of voice activity detector (VAD) becomes crucial for the efficient bandwidth reduction. Most of the conventional VAD algorithms assume that the background noise statistics are stationary over a longer period of time than those of speech, which makes it possible to estimate the time varying noise statistics in spite of the occasional presence of speech [1]. To determine the presence or absence of speech, the observed signal statistics in the current frame are compared with the estimated noise statistics according to some decision rules. Moreover, this initial decision is modified by a hang-over scheme to minimize misdetections at weak speech tails.

Traditional VAD algorithms are usually designed using heuristics, which makes it difficult to optimize the relevant parameters. Recently, an effort to optimize a VAD by applying a statistical model has been made in [2], where the decision rule is derived from the likelihood ratio test (LRT) by estimating unknown parameters using the maximum likelihood (ML) criterion. In this letter, we further optimize the decision rule by employing the decision-directed (DD) method for the estimation of the unknown parameters [3]. We also propose an effective hang-over scheme based on the hidden Markov model (HMM).

II. Decision Rules Based on LRT

Assuming that speech is degraded by uncorrelated additive noise, two hypotheses for a VAD to consider for each frame are

$$H_0\text{: speech absent: } \mathbf{X} = \mathbf{N}$$

$$H_1\text{: speech present: } \mathbf{X} = \mathbf{N} + \mathbf{S}$$

where $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{X}$ are L dimensional discrete Fourier transform (DFT) coefficient vectors of speech, noise, and noisy speech with their kth elements S_k, N_k, and X_k, respectively. We adopt the Gaussian statistical model that the DFT coefficients of each process are asymptotically independent Gaussian random variables [3]. Then the probability density functions conditioned on H_0 and H_1 are given by

$$p(\mathbf{X}|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)}\right\} \tag{1}$$

$$p(\mathbf{X}|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k)+\lambda_S(k)]} \cdot \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)+\lambda_S(k)}\right\} \tag{2}$$

where $\lambda_N(k)$ and $\lambda_S(k)$ denote the variances of N_k and S_k, respectively. The likelihood ratio for the kth frequency band is

$$\Lambda_k \triangleq \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\} \tag{3}$$

where $\xi_k \triangleq \lambda_S(k)/\lambda_N(k)$ and $\gamma_k \triangleq |X_k|^2/\lambda_N(k)$, and they are called the *a priori* and *a posteriori* signal-to-noise ratios (SNR's), respectively [3]. The decision rule is established from the geometric mean of the likelihood ratios for the individual frequency bands, which is given by

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{4}$$

We assume that $\lambda_N(k)$'s are already known through the noise statistic estimation procedure, and have to estimate the unknown parameters, ξ_k's.

The ML estimator for ξ_k can easily be derived as follows:

$$\hat{\xi}_k^{(\mathrm{ML})} = \gamma_k - 1. \tag{5}$$

Substituting (5) into (4) and applying the LRT yields the Itakura–Saito distortion (ISD) based decision rule [2], i.e.,

$$\log \hat{\Lambda}^{(\mathrm{ML})} = \frac{1}{L} \sum_{k=0}^{L-1} \{\gamma_k - \log \gamma_k - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{6}$$

Note that the left-hand side of (6) can not be smaller than zero, which is the well-known property of ISD and implies that the likelihood ratio is biased to H_1.

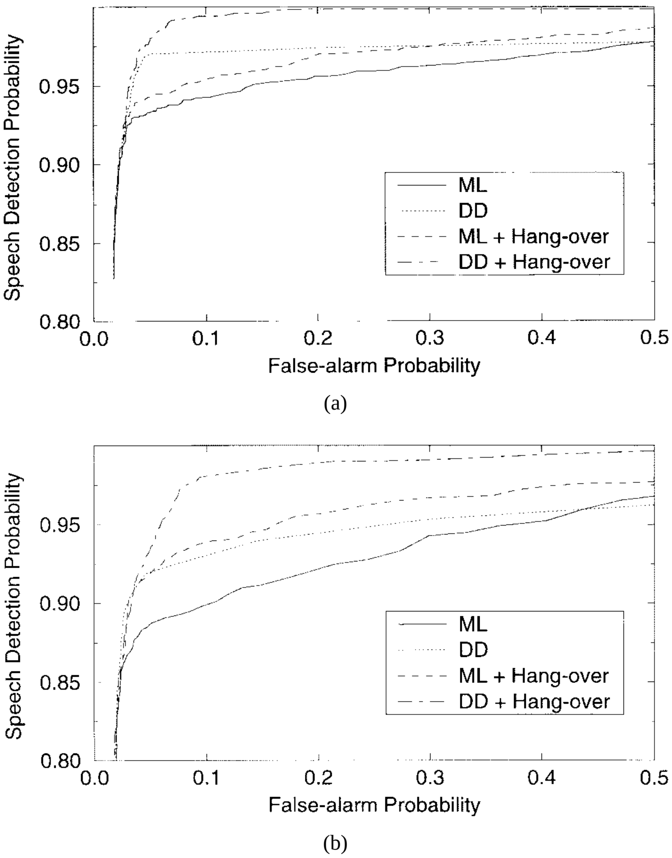
In order to reduce this bias, we apply the DD *a priori* SNR estimation method [3]:

$$\hat{\xi}_k(n)^{(\mathrm{DD})} = \alpha \frac{\hat{A}_k^2(n-1)}{\lambda_N(k, n-1)} + (1-\alpha)P[\gamma_k(n) - 1] \tag{7}$$

Manuscript received July 17, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. K. Paliwal.

The authors are with the School of Electrical Engineering, Seoul National University, Seoul 151-742, Korea (e-mail: wysung@dsp.snu.ac.kr).

Publisher Item Identifier S 1070-9908(99)00783-X.



(a)

(b)

Fig. 1. Receiver operating characteristics of ML and DD based decision rules with and without hang-over at 5 dB SNR. (a) Vehicular noise. (b) White noise.

where n is the frame index, $P[x] = x$ if $x \geq 0$, and $P[x] = 0$ otherwise, and $\hat{A}_k(n-1)$'s are the signal amplitude estimates of the previous frame, for which we use the minimum mean-square error (MMSE) estimator [3]. The DD method of (7) provides smoother estimates of the *a priori* SNR than the ML method [4], and consequently reduces the fluctuation of the estimated likelihood ratios during noise-only periods.

III. HMM-Based Hang-Over Scheme

In practical VAD's, the initial decision is modified to prevent clipping of weak speech tails, by considering the previous decision results. Conventional hang-over algorithms usually adopt a scheme that delays the transition from H_1 to H_0, which reduces the misdetection of speech at the cost of the increased false-alarm rate.

Actually, the hang-over is based on the idea that there is a strong correlation in the consecutive occurrences of speech frames. To express this property explicitly, we model the sequence of frame states as a first-order Markov process. Since the Markov process assumes that the current state only depends on the previous state, the correlative characteristic of speech occurrence can be represented by $P(q_n = H_1 | q_{n-1} = H_1)$ with the following constraint:

$$P(q_n = H_1 | q_{n-1} = H_1) > P(q_n = H_1) \qquad (8)$$

where q_n denotes the state of the nth frame and is either H_0 or H_1.

By assuming that the Markov process is time invariant, we can use the notation, $a_{ij} \triangleq P(q_n = H_j | q_{n-1} = H_i)$. We further assume the stationarity of the process to have $P(q_n = H_i) = P(H_i)$, where $P(H_0)$ and $P(H_1)$ are the steady-state probabilities obtained from $a_{01}P(H_0) = a_{10}P(H_1)$ and $P(H_0) + P(H_1) = 1$. Thus, the overall process can be characterized by only two parameters, a_{01} and a_{10}.

In this Markovian frame state model, the current state depends on the previous observations as well as the current one, which is reflected on the decision rule in the following way:

$$\begin{aligned}\mathcal{L}(n) &\triangleq \frac{p(\mathcal{X}_n | q_n = H_1)}{p(\mathcal{X}_n | q_n = H_0)} \\ &= \frac{P(H_0)}{P(H_1)} \frac{P(q_n = H_1 | \mathcal{X}_n)}{P(q_n = H_0 | \mathcal{X}_n)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \end{aligned} \qquad (9)$$

where $\mathcal{X}_n = \{\mathbf{X}(n), \mathbf{X}(n-1), \cdots, \mathbf{X}(1)\}$ represents the set of observations up to the current frame n. For the efficient computation of the *a posteriori* probability ratio in (9), $\Gamma(n) \triangleq P(q_n = H_1 | \mathcal{X}_n)/P(q_n = H_0 | \mathcal{X}_n)$, we define the forward variable as $\alpha_n(i) \triangleq p(q_n = H_i, \mathcal{X}_n)$. By using the forward procedure [5], we can solve for $\alpha_n(i)$ as follows:

$$\alpha_n(i) = \begin{cases} P(H_i)p(\mathbf{X}(1)|q_1 = H_i), & \text{if } n = 1 \\ (\alpha_{n-1}(0)a_{0j} + \alpha_{n-1}(1)a_{1j}) & \\ \quad \cdot p(\mathbf{X}(n)|q_n = H_i), & \text{if } n \geq 2. \end{cases} \qquad (10)$$

Based on the above formulations, a recursive formula for $\Gamma(n)$ is obtained as

$$\Gamma(n) = \frac{\alpha_n(1)}{\alpha_n(0)} = \frac{a_{01} + a_{11}\Gamma(n-1)}{a_{00} + a_{10}\Gamma(n-1)} \Lambda(n) \qquad (11)$$

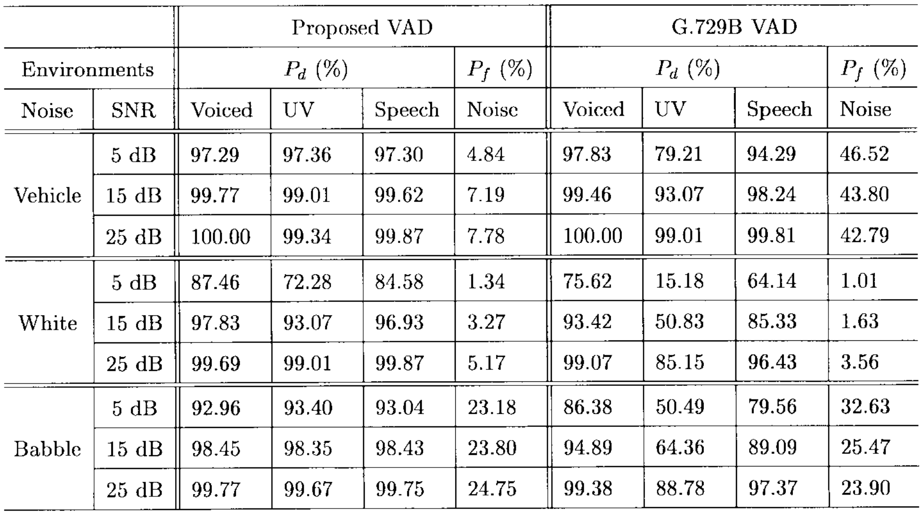
where $\Lambda(n)$ denotes the likelihood ratio in (4) at nth frame. Consequently, the final decision statistic is obtained by $\mathcal{L}(n) = [P(H_0)/P(H_1)]\Gamma(n)$.

IV. Experimental Results

To verify the effectiveness of the proposed algorithms, we compared the speech detection and false-alarm probabilities (P_d and P_f) of the ML and DD based decision rules with and without hang-over. For the frame state model, we used $a_{01} = 0.2$ and $a_{10} = 0.1$. To obtain P_d and P_f, we made reference decisions for a clean speech material of 46 s long by labeling manually at every 10 ms frame. The percentage of the hand-marked speech frames is 34.27%, which consists of 27.76% voiced and 6.51% unvoiced frames. We defined P_d as the ratio of correct speech decisions to the hand-marked speech frames, while P_f as that of false speech decisions to the hand-marked noise frames.

The receiver operating characteristics (ROC's), which shows the trade-off characteristic between P_d and P_f, of the four decision rules are shown in Fig. 1, where the VAD algorithms were applied to the noisy speech samples corrupted by the vehicular and white noise sources from NOISEX-92 data base at 5 dB SNR. As shown in Fig. 1, the DD-based decision rule performs superior to the ML-based one unless the allowed P_f is impractically very small or very large.

TABLE I
P_d's AND P_f's OF THE PROPOSED AND G.729B VAD's FOR VARIOUS ENVIRONMENTAL CONDITIONS



		Proposed VAD				G.729B VAD			
Environments		P_d (%)			P_f (%)	P_d (%)			P_f (%)
Noise	SNR	Voiced	UV	Speech	Noise	Voiced	UV	Speech	Noise
Vehicle	5 dB	97.29	97.36	97.30	4.84	97.83	79.21	94.29	46.52
	15 dB	99.77	99.01	99.62	7.19	99.46	93.07	98.24	43.80
	25 dB	100.00	99.34	99.87	7.78	100.00	99.01	99.81	42.79
White	5 dB	87.46	72.28	84.58	1.34	75.62	15.18	64.14	1.01
	15 dB	97.83	93.07	96.93	3.27	93.42	50.83	85.33	1.63
	25 dB	99.69	99.01	99.87	5.17	99.07	85.15	96.43	3.56
Babble	5 dB	92.96	93.40	93.04	23.18	86.38	50.49	79.56	32.63
	15 dB	98.45	98.35	98.43	23.80	94.89	64.36	89.09	25.47
	25 dB	99.77	99.67	99.75	24.75	99.38	88.78	97.37	23.90

Note that the increase of P_d due to the proposed hang-over scheme is also quite noticeable in both the ML and DD based decision rules.

Finally, we developed a robust VAD by combining the DD based decision rule and the proposed hang-over scheme with the soft-decision based noise spectrum adaptation algorithm described in [2]. To evaluate the performance of the proposed VAD, we measured the P_d and P_f in various environmental conditions using the aforementioned speech material and reference decisions. And we compared the proposed scheme with the VAD specified in the ITU standard G.729 Annex B [6], in terms of P_d and P_f. The results are summarized in Table I, where the hand-marked speech frames are subdivided into the voiced and unvoiced regions. In all the testing conditions, the proposed VAD significantly outperformed or at least was comparable to the G.729B VAD.

V. CONCLUSIONS

In this study, the decision-directed parameter estimation method is applied to the likelihood ratio test, which improves the performance of the VAD by yielding smoother estimates of the *a priori* SNR. The HMM-based hang-over scheme also increases the speech detection probability for a given false-alarm rate. The proposed VAD shows better performances in various environmental conditions, while requires only a few parameters to optimize when compared with the G.729B VAD.

REFERENCES

[1] K. Srinivasan and A. Gersho, “Voice activity detection for cellular networks,” in *Proc. IEEE Speech Coding Workshop,* Oct. 1993, pp. 85–86.

[2] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing,* 1998, pp. 365–368.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[4] O. Cappé, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Trans. Speech Audio Processing,* vol. 2, pp. 345–349, Apr. 1994.

[5] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[6] ITU-T Rec. G.729, Annex B, *A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70.*