Statistical Model-Based Voice Activity Detection Based on Second-Order Conditional MAP with Soft Decision

Joon-Hyuk Chang

In this paper, we propose a novel approach to statistical model-based voice activity detection (VAD) that incorporates a second-order conditional maximum a posteriori (CMAP) criterion. As a technical improvement for the first-order CMAP criterion in [1], we consider both the current observation and the voice activity decision in the previous two frames to take full consideration of the interframe correlation of voice activity. This is clearly different from the previous approach [1] in that we employ the voice activity decisions in the second-order (previous two frames) CMAP, which has quadruple thresholds with an additional degree of freedom, rather than the first-order (previous single frame). Also, a soft-decision scheme is incorporated, resulting in time-varying thresholds for further performance improvement. Experimental results show that the proposed algorithm outperforms the conventional CMAP-based VAD technique under various experimental conditions.

Keywords: Voice activity detection, second-order conditional MAP, soft decision, likelihood ratio test.

I. Introduction

A voice activity detector (VAD), referring to a mean of a finite state machine with two states, "speech absent" and "speech present," is the most important part of discontinuous transmission (DTX), especially in mobile voice over internet protocol (VoIP) systems. Most traditional algorithms are based on the linear prediction coding (LPC) parameter, cepstral feature, and the periodicity measure [2]-[4]. More recently, as a novel strategy, VADs based on the likelihood ratio test (LRT) employing a statistical model have been proposed and shown to have superior performance despite the need for optimization of a few relevant parameters [5]-[10]. Note that the traditional LRT is eventually based on the maximum a posteriori (MAP) criterion, which selects the hypothesis having the maximum probability given a current observation. The previous approach by Shin and others [1] considered the interframe correlation of the speech signal since the conventional approach based on the MAP characterizes each frame separately. Actually, this has been done by incorporating a simple conditional MAP (CMAP) criterion that chooses the hypothesis with the higher probability conditioned on the current data and the voice activity decision in the previous frame. This can be considered to be relevant because the decision threshold for the LRT has two different values adaptively according to the status of voice activity in the previous frame. On the other hand, Ramirez and others [9] proposed an algorithm to incorporate long-term speech information to the LRT method. However, this approach has an inherent delay that is not adequate for the real-time speech communication scenario.

Manuscript received June 1, 2011; revised Oct. 5, 2011; accepted Oct. 19, 2011.

This work was supported by the research fund of Hanyang University (HY-2011-201100000000210).

Joon-Hyuk Chang (phone: +82 2 2220 0355, jchang@hanyang.ac.kr) is with the Department of Electronic Engineering, Hanyang University, Seoul, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.12.0111.0344

In this paper, we propose a novel technique for the LRT-based VAD derived from the second-order CMAP, in which the LRT decision is performed in a frame using both the current observation and the voice activity of the previous two frames. This is because taking the higher order of the CMAP into consideration enables more exact detection in the VAD. As a result, the idea behind this technique results in four different thresholds rather than the two thresholds in [1], which can provide an additional freedom in decision-making. In addition, we further improve the second-order CMAP by adopting the soft decision scheme in which the four thresholds are combined with the probability of speech absence in the previous two frames, which achieves the relevant time-varying threshold. From a number of experiments on VAD, we can see that the proposed approach shows better performance than the algorithm proposed by Shin and others [1].

II. Review of CMAP-Based VAD

In the time domain, it is assumed that the noise signal $d(t)$ is added to the clean speech signal $x(t)$, with their sum being denoted by $y(t)$, which is called the noisy speech signal. They are transformed by the discrete Fourier transform (DFT) as

$$\mathbf{Y}(n) = \mathbf{X}(n) + \mathbf{D}(n), \tag{1}$$

where $\mathbf{Y}(n) = [Y(0, n), Y(1, n), ..., Y(M-1, n)]$, $\mathbf{X}(n) = [X(0,n), X(1,n), ..., X(M-1,n)]$, and $\mathbf{D}(n) = [D(0,n), D(1,n), ..., D(M-1,n)]$ at the n-th frame. Assuming that speech is degraded by uncorrelated additive noise, two hypotheses, $H_0(n)$ and $H_1(n)$, indicate speech absence and presence in the noisy spectral component

$$H_0(n): Y(k,n) = D(k,n), \tag{2}$$

$$H_1(n): Y(k,n) = X(k,n) + D(k,n). \tag{3}$$

With the Gaussian probability density function (pdf) assumption [7], the distribution of the noisy spectral components conditioned on both hypotheses are given by

$$p(Y(k,n) \mid H_0(n)) = \frac{1}{\pi\lambda_d(k,n)} \exp\{-\tfrac{|Y(k,n)|^2}{\lambda_d(k,n)}\}, \tag{4}$$

$$p(Y(k,n) \mid H_1(n)) = \frac{1}{\pi(\lambda_d(k,n) + \lambda_x(k,n))} \times \exp\{-\tfrac{|Y(k,n)|^2}{\lambda_d(k,n)+\lambda_x(k,n)}\}, \tag{5}$$

where $\lambda_d(k,n)$ and $\lambda_x(k,n)$ denote the variances of noise and speech for the individual frequency band, respectively. The likelihood ratio (LR) of the k-th frequency band is given by

$$\Lambda(k,n) \equiv \frac{p(Y(k,n) \mid H_1(n))}{p(Y(k,n) \mid H_0(n))} = \frac{1}{1+\xi(k,n)} \exp\{\tfrac{\gamma(k,n)\xi(k,n)}{1+\xi(k,n)}\}, \tag{6}$$

where $\xi(k,n) = \lambda_x(k,n)/\lambda_d(k,n)$ and $\gamma(k,n) = Y(k,n)/\lambda_d(k,n)$ denote the a priori signal-to-noise ratio (SNR) and the a posteriori SNR, respectively [7]. The a posteriori SNR $\gamma(k,n)$ is estimated using $\lambda(k,n)$, and the a priori SNR $\xi(k,n)$ is estimated by the well-known decision directed (DD) method that follows [5]:

$$\hat{\xi}(k,n) = \alpha \frac{|\hat{X}(k,n-1)|^2}{\lambda_d(k,n-1)} + (1-\alpha) P[\gamma(k,n) - 1], \tag{7}$$

where $|\hat{X}(k,n-1)|^2$ is the speech spectral amplitude estimate of the previous frame obtained using the minimum mean-square error (MMSE) estimator [7]. Also, α (=0.99) is a fixed weight [5] and the function $P[x] = x$ if $x \geq 0$ and $P[x] = 0$, otherwise. The final decision in the conventional statistical model-based VADs has been achieved by the geometric mean of the LRs computed for the individual frequency bins [5]-[10] and is obtained by

$$\Lambda(n) = \frac{1}{M} \sum_{k=0}^{M-1} \log \Lambda(k,n) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{8}$$

where an input frame is classified as speech presence if the geometric mean of the LRs is greater than a certain threshold value η; otherwise, it is classified as speech absence.

The previous CMAP-based VAD originated from the conventional MAP as the following decision rule:

$$\frac{p(H(n) = H_1(n) \mid \mathbf{Y}(n))}{p(H(n) = H_0(n) \mid \mathbf{Y}(n))} \underset{H_0}{\overset{H_1}{\gtrless}} 1, \tag{9}$$

where $H(n)$ denotes the correct hypothesis in the n-th frame. This rule is changed to the following criterion in the LRT such that

$$\frac{p(\mathbf{Y}(n) \mid H(n) = H_1(n))}{p(\mathbf{Y}(n) \mid H(n) = H_0(n))} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{p(H(n) = H_0(n))}{p(H(n) = H_1(n))}, \tag{10}$$

where $\alpha \geq 1$ [7].

This time, Shin and others proposed a way to incorporate the interframe correlation of the voice activity into the MAP criterion. Specifically, the a posteriori probability

$p(H(n)\,|\,\mathbf{Y}(n))$ is not only conditioned on the current observation $\mathbf{Y}(n)$, but also on the decision in the previous frame, that is, $p(H(n)\,|\,\mathbf{Y}(n), H(n-1))$. Then, it implies that

$$\frac{p(H(n)=H_1(n)\,|\,\mathbf{Y}(n), H(n-1)=H_i)}{p(H(n)=H_0(n)\,|\,\mathbf{Y}(n), H(n-1)=H_i)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha\,, i=0,1, \quad (11)$$

where α is the threshold. Upper criterion could be expressed such that [1]

$$\frac{p(\mathbf{Y}(n)\,|\,H(n)=H_1(n), H(n-1)=H_i)}{p(\mathbf{Y}(n)\,|\,H(n)=H_0(n), H(n-1)=H_i)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{p(H(n)=H_0\,|\,H(n-1)=H_i)}{p(H(n)=H_1\,|\,H(n-1)=H_i)},\ i=0,1. \quad (12)$$

Note that the likelihoods $p(\mathbf{Y}(n)\,|\,H(n)=H_1(n), H(n-1)=H_i)$ and $p(\mathbf{Y}(n)\,|\,H(n)=H_0(n), H(n-1)=H_i)$ could be simplified for the dominant contribution of the distribution of $\mathbf{Y}(n)$ in the current frame as

$$\frac{p(\mathbf{Y}(n)\,|\,H(n)=H_1(n))}{p(\mathbf{Y}(n)\,|\,H(n)=H_0(n))} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{p(H(n)=H_0\,|\,H(n-1)=H_i)}{p(H(n)=H_1\,|\,H(n-1)=H_i)},\ i=0,1. \quad (13)$$

III. Proposed Algorithm

As for the derived statistic in the method of Shin and others of the previous section, what we should note is that two separate thresholds are introduced according to the decision of the speech activity in the previous frame. Clearly, the multiple thresholds can give us an additional freedom that improves the performance of the VAD. Indeed, it is not sufficient for the single frame in the previous CMAP to express the strong correlation in the consecutive occurrences of speech frames. Accordingly, one obvious motivation based on the upper derivation is that the second-order CMAP incorporating the VAD decisions in the previous two frames could perform better than the first-order CMAP scheme proposed by Shin and others [1] and improve speech detection robustness. However, we do not consider the generalization to more than the second-order since the performance improvement considering additional computation load was limited. Based on this, we derive the second-order CMAP as

$$\frac{p(H(n)=H_1(n)\,|\,\mathbf{Y}(n), H(n-1)=H_i, H(n-2)=H_j)}{p(H(n)=H_0(n)\,|\,\mathbf{Y}(n), H(n-1)=H_i, H(n-2)=H_j)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha,\quad i=0,1 \text{ and } j=0,1. \quad (14)$$

This decision test is again changed to the following form:

$$\frac{p(\mathbf{Y}(n)\,|\,H(n)=H_1(n), H(n-1)=H_i, H(n-2)=H_j)}{p(\mathbf{Y}(n)\,|\,H(n)=H_0(n), H(n-1)=H_i, H(n-2)=H_j)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{p(H(n)=H_0\,|\,H(n-1)=H_i, H(n-2)=H_j)}{p(H(n)=H_1\,|\,H(n-1)=H_i, H(n-2)=H_j)},$$
$$i=0,1 \text{ and } j=0,1. \quad (15)$$

In a similar reason with the conventional CMAP criterion, (15) can be approximated as

$$\frac{p(\mathbf{Y}(n)\,|\,H(n)=H_1(n))}{p(\mathbf{Y}(n)\,|\,H(n)=H_0(n))} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \frac{p(H(n)=H_0\,|\,H(n-1)=H_i, H(n-2)=H_j)}{p(H(n)=H_1\,|\,H(n-1)=H_i, H(n-2)=H_j)},$$
$$i=0,1 \text{ and } j=0,1. \quad (16)$$

Therefore, (16) can be finally expressed by

$$\frac{p(\mathbf{Y}(n)\,|\,H(n)=H_1(n))}{p(\mathbf{Y}(n)\,|\,H(n)=H_0(n))} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{ij},\ i=0,1 \text{ and } j=0,1, \quad (17)$$

where $\gamma_{ij} = \alpha \frac{p(H(n)=H_0\,|\,H(n-1)=H_i, H(n-2)=H_j)}{p(H(n)=H_1\,|\,H(n-1)=H_i, H(n-2)=H_j)}$.

We note from the proposed test statistics that four separate thresholds are needed depending on the speech activity in the previous two frames. Specifically, for example, $\gamma_{00}\left(=\alpha \frac{p(H(n)=H_0\,|\,H(n-1)=H_0, H(n-2)=H_0)}{p(H(n)=H_1\,|\,H(n-1)=H_0, H(n-2)=H_0)}\right)$ is used if the absence of speech is detected in the previous two frames. Or, $\gamma_{10}\left(=\alpha \frac{p(H(n)=H_0\,|\,H(n-1)=H_1, H(n-2)=H_0)}{p(H(n)=H_1\,|\,H(n-1)=H_1, H(n-2)=H_0)}\right)$ is used if the presence of speech is detected in the previous first frame and the absence of speech is detected in the previous second

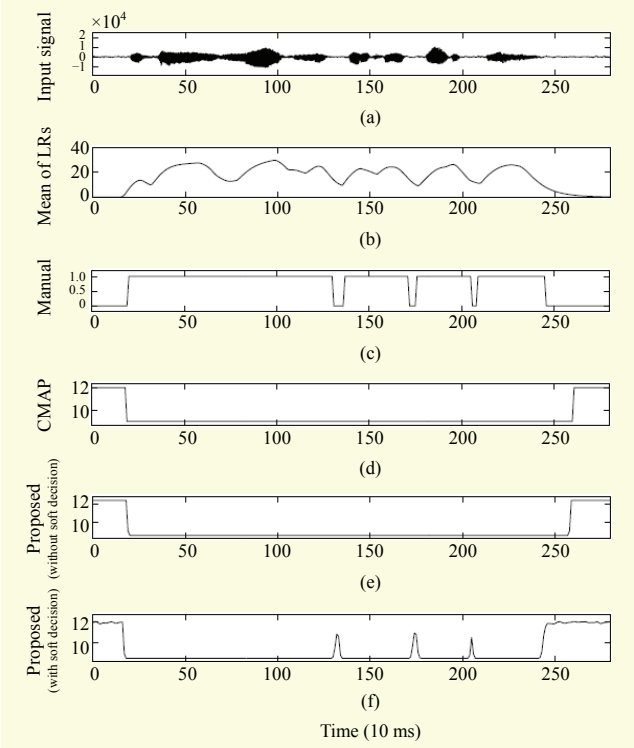

Fig. 1. (a) Waveform of test file (car noise, SNR=15 dB), (b) mean of LRs in (8), (c) manual VAD (silence=0, speech=1), (d) threshold of the first-order CMAP, (e) threshold of the second-order CMAP (without soft decision), and (f) threshold of the second-order CMAP (with soft decision).

frame. This quadruples thresholds according to the speech activities in the previous two frames and can provide more reliable statistics for testing the voice activity by considering interframe correlation successfully.

On the other hand, further improvement could be achieved by incorporating the soft decision scheme. This implies that the quadruple thresholds are combined into a single threshold with each speech absence probability (SAP) of the previous two frames as follows [11]:

$$\frac{p(\mathbf{Y}(n)\,|\,H(n)=H_1(n))}{p(\mathbf{Y}(n)\,|\,H(n)=H_0(n))} \underset{H_0}{\overset{H_1}{>}} \eta, \tag{18}$$

where $\eta(=\sum_{i=0}^{1}\sum_{j=0}^{1} w_{ij}\gamma_{ij})$ is the new threshold and $w_{ij} = p(H(n-1)=H_i\,|\,\mathbf{Y}(n-1))\cdot p(H(n-2)=H_j\,|\,\mathbf{Y}(n-2))$. In time, $p(H(n)=H_i\,|\,\mathbf{Y}(n))$ at n-th frame is the global SAP, which is obtained as in [11] such that

$$p(H(n)=H_0\,|\,\mathbf{Y}(n)) = \frac{p(Y(n)\,|\,H(n)=H_0)p(H_0)}{p(\mathbf{Y}(n))} = \frac{1}{1+\frac{p(H_1)}{p(H_0)}\prod_{k=0}^{M-1}\Lambda(k,n)}, \tag{19}$$

where $p(H_0)(=1-p(H_1))$ is the a priori probability of

Table 1. Comparison of voice activity detection P_E, P_M, and P_F among statistical model-based and CMAP-based methods and proposed technique.

		G.729 B			Sohn [7]			CMAP [1]			Proposed		
Noise	SNR (dB)	P_E	P_M	P_F	P_E	P_M	P_F	P_E	P_M	P_F	P_E	P_M	P_F
Car	0	31.63	13.51	56.87	7.45	6.58	8.65	7.20	6.24	8.53	6.97	5.91	8.46
	5	27.52	13.00	47.72	7.21	6.75	8.30	6.85	6.52	8.18	6.62	5.83	8.07
	10	23.48	10.03	42.20	5.68	1.87	11.07	5.65	1.92	10.82	5.63	1.91	10.81
	15	19.79	7.14	37.40	5.43	1.63	10.86	5.36	1.58	10.83	5.32	1.57	10.80
Street	0	23.53	24.35	22.38	10.90	8.51	14.04	10.85	7.61	15.16	10.74	8.04	14.37
	5	23.03	23.74	22.05	10.33	7.40	14.41	10.29	7.88	13.65	10.15	7.83	13.38
	10	19.52	19.09	20.11	9.34	1.97	19.60	8.93	2.17	18.34	8.79	2.32	17.79
	15	15.73	14.05	18.07	9.00	0.81	20.40	8.84	0.85	19.97	8.72	0.89	19.63
Office	0	28.51	38.15	15.10	17.87	13.45	24.02	17.50	10.02	27.91	17.29	10.71	26.46
	5	26.45	28.67	23.35	17.85	13.61	23.75	16.80	12.92	22.20	16.17	12.50	21.28
	10	22.74	22.29	23.37	13.46	4.08	26.52	12.85	3.80	25.44	12.48	3.73	24.67
	15	19.28	17.30	22.04	11.25	1.35	25.03	10.84	1.18	24.29	10.65	1.18	23.81
Babble	0	34.48	41.67	24.48	22.51	19.14	27.21	22.08	17.17	28.91	21.45	15.81	29.30
	5	27.81	30.04	24.72	13.64	10.91	17.43	13.37	9.01	19.21	13.01	8.09	19.86
	10	22.65	22.24	23.22	8.25	5.64	11.87	7.99	5.53	11.41	7.85	5.54	11.06
	15	19.62	13.91	25.17	6.27	3.04	10.61	6.08	2.53	10.83	5.93	2.49	10.72

the speech absence. Note that it is desirable to modify the quadruple threshold γ_{ij} like (18) if we are provided with the knowledge on speech absence or presence of the previous two frames because speech might not be present at all frequencies and at all times [12]. For example, if the previous two frames are not certain about either speech absence or presence such as $p(H(n-1)=H_0 \mid \mathbf{Y}(n-1))=0.5$ and $p(H(n-2)=H_0 \mid \mathbf{Y}(n-2))=0.5$, η results in $0.25\gamma_{00}+0.25\gamma_{01}+0.25\gamma_{10}+0.25\gamma_{11}$. Based on this, it is observed that we can adjust the threshold by γ_{ij} and the SAPs in the previous two frames as shown in Fig. 1. From the figure, it can be seen that the threshold in the proposed method performs well by taking advantage of the soft decision scheme. Specifically, the new threshold η being adjusted by the SAPs show soft values along time, which yields better performance taking into account the manual VAD in Fig. 1(c).

Note that this method inherently incorporates the interframe correlation in determining the detection thresholds so that it can be combined with other hangover schemes such as the HMM [7] and the smoothing of the LRs [6].

IV. Experiments and Results

Conventional methods and the proposed method were evaluated in a quantitative comparison under various noise environments. For the test material, 456 s of speech data was recorded by four males and four females, and then it was sampled at 8 kHz. To evaluate the performance, we first made reference decisions on the clean speech material by labeling it manually at every 10 ms frame. The proportion of hand-marked speech frames was 57.1% and that consisted of 44.0% voiced sounds and 13.1% unvoiced sounds. In all cases, regardless of noise type, voice activity detection was conducted with $\gamma_{00}=12.5$, $\gamma_{01}=10$, $\gamma_{10}=9.5$, and $\gamma_{11}=8.5$, which are experimentally selected based on the correct speech/silence transcription on the 230-s length different speech material. Also, the initial value η was set to 12.5 because we assumed that only noise exists on the initial several frames. Note that the threshold parameters are not sensitive to the type of and intensity of environmental noise if we see the LRs depending on the hypotheses obtained from several noises. In this regard, to consider various noise environments, car, street, and office noises were added to the clean speech data by varying SNR such as 0 dB, 5 dB, 10 dB, and 15 dB. Also, to simulate non-stationary noise in real environments, the babble noise is included at the same SNRs.

Table 1 including probability of error P_E, probability of miss P_M, and false alarm probability P_F shows comparative results for the Sohn's VAD (without hangover), the first-order CMAP-based method and the proposed approach with either soft decision or not. In particular, to help someone to repeat the results, the standardized VAD such as ITU-T G.729 Annex B [13] is included. From the results, it is evident that the proposed VAD algorithm shows better performance than the previously reported VAD methods including the first-order CMAP [1] in most of the environmental conditions. This fact could be confirmed by Figs. 2 and 3 showing the receiver operating characteristics (ROC), which are insensitive to parameter tuning since it is a trade-off between $(1-P_M)$ and P_F. Based on this, we can see the overall performance differences of the aforementioned systems. From the figures, it can be seen that the proposed second-order CMAP-based VAD yielded consistently higher performance than the previous CMAP-based method. In particular, the proposed approach with soft decision has been found to be the best among the given methods, which shows the advantage of the soft decision

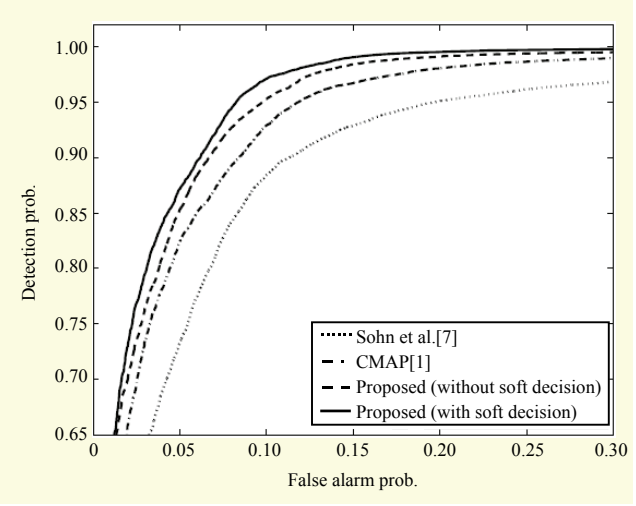

Fig. 2. ROC curve for office noise at 5 dB SNR.

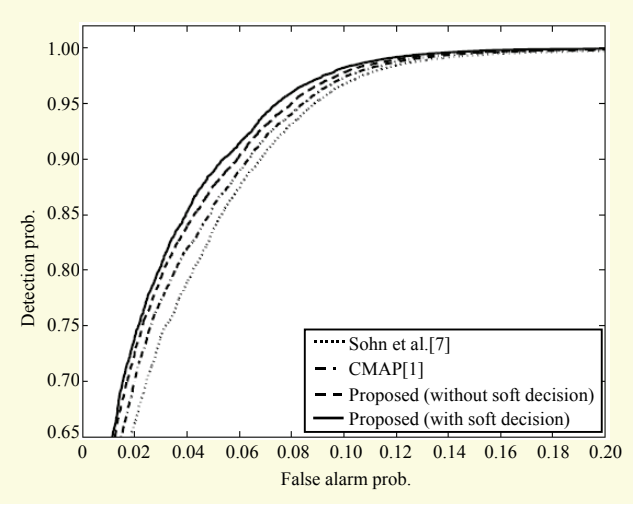

Fig. 3. ROC curve for street noise at 15 dB SNR.

Table 2. Computational cost comparison per single frame (common parts like the FFT are excluded).

	Sohn [7]	CMAP [1]	Proposed
Computational cost	8,977	8,979	9,009

scheme.

Even though the superiority of the proposed technique in terms of detection accuracy is clear, is computational complexity should be issued for fair comparison. For this reason, we investigated the computational cost based on the cost of the operation such that add=1, multiplication=1, division=5, and exponent=10. Then, the computational costs of the previous approaches (Sohn and CMAP) and the proposed approach were compared as shown in Table 2. As shown in Table 2, an improvement of the detection accuracy was achieved with little burden regarding the additional computational load involved in the proposed second-order CMAP with soft decision.

V. Conclusion

In this paper, we proposed a novel VAD technique based on the second-order CMAP algorithm in which is incorporated the speech presence or absence probability of the previous two frames for a robust VAD decision. The proposed approach yields better performance than the conventional method in various noise environments.

References

[1] J.W. Shin et al., "Voice Activity Detection Based on Conditional MAP Criterion," *IEEE Signal Proc. Lett.*, vol. 15, Feb. 2008, pp. 257-260.

[2] L.R. Rabiner and M.R. Sambur, "Voiced-Unvoiced-Silence Detection Using the Itakura LPC Distance Measure," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, May 1977, pp. 323-326.

[3] J.A. Haigh and J.S. Mason, "Robust Voice Activity Detection Using Cepstral Features," *Proc. IEEE TENCON*, vol. 3, Oct. 1993, pp. 321-324.

[4] K. Srinivasant and A. Gersho, "Voice Activity Detection for Cellular Networks," *Proc. IEEE Works. Speech Coding Telecommu.,* Oct. 1993, pp. 85-86.

[5] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-32, no. 6, Dec. 1984, pp. 1109-1121.

[6] Y.D. Cho, K. Al-Naimi, and A. Kondoz, "Improved Voice Activity Detection Based on a Smoothed Statistical Likelihood Ratio," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, vol. 2, May 2001, pp. 737-740.

[7] J. Sohn, N.S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Proc. Lett.*, vol. 6, no. 1, Jan. 1999, pp. 1-3.

[8] J.-H. Chang, N.S. Kim, and S.K. Mitra, "Voice Activity Detection Based on Multiple Statistical Models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, June 2006, pp. 1965-1976.

[9] J. Ramirez et al, "Statistical Voice Activity Detection Using a Multiple Observation Likelihood Ratio Test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, Oct. 2005, pp. 689-692.

[10] J.-H. Chang, J.W. Shin, and N.S. Kim, "Likelihood Ratio Test with Complex Laplacian Model for Voice Activity Detection," *Proc. Eurospeech*, Aug. 2003, pp. 1065-1068.

[11] J.-H. Chang et al., "Global Soft Decision Employing Support Vector Machine for Speech Enhancement," *IEEE Signal Proc. Lett.*, vol. 16, no. 1, Jan. 2009, pp. 57-60.

[12] P.C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[13] ITU-T, "A Silence Compression Scheme for G.729 Optimised for Terminals Conforming to Recommendation V.70," ITU-T Rec. G.729, Annex B, 1996.

Joon-Hyuk Chang received his BS in electronics engineering from Kyungpook National University, Daegu, Rep. of Korea, in 1998, and his MS and PhD in electrical engineering from Seoul National University, Rep. of Korea, in 2000 and 2004, respectively. From March 2000 to April 2005, he was with Netdus Corp., Seoul, as a chief engineer. From May 2004 to April 2005, he was with the University of California, Santa Barbara, in a postdoctoral position to work on adaptive signal processing and audio coding. In May 2005, he joined the Korea Institute of Science and Technology, Seoul, as a research scientist to work on speech recognition. From August 2005 to February 2011, he was an assistant professor in the School of Electronic Engineering at Inha University, Incheon, Rep. of Korea. Currently, he is an associate professor in the School of Electronic Engineering at Hanyang University, Seoul, Rep. of Korea. His research interests are in speech coding, speech enhancement, speech recognition, audio coding, and adaptive signal processing. He is a winner of the IEEE/IEEK young IT engineer of the year 2011.