IEICE TRANS. INF. & SYST., VOL.E95–D, NO.10 OCTOBER 2012

LETTER

Voice Activity Detection Using Global Speech Absence Probability Based on Teager Energy for Speech Enhancement

Yun-Sik PARK[†], *Nonmember and* **Sangmin LEE**[†a)], *Member*

SUMMARY In this paper, we propose a novel voice activity detection (VAD) algorithm using global speech absence probability (GSAP) based on Teager energy (TE) for speech enhancement. The proposed method provides a better representation of GSAP, resulting in improved decision performance for speech and noise segments by the use of a TE operator which is employed to suppress the influence of noise signals. The performance of our approach is evaluated by objective tests under various environments, and it is found that the suggested method yields better results than conventional schemes.

key words: *voice activity detection, speech absence probability, teager energy*

1. Introduction

In a variety of speech procedures, such as speech recognition and speech enhancement, the voice activity detection (VAD) algorithm is indispensable because the performance of these speech processing procedures depends critically on the result of VAD. To determine the presence or absence of speech by VAD algorithms, various feature parameters to distinguish speech segments from other waveforms have been adopted. Traditionally, the parameters that can specify the characteristics of speech have been based on short-time energy or spectral energy and zero-crossing rate (ZCR). All of these parameters, however, are rather sensitive to noise and cannot fully specify the characteristics of a speech signal. Therefore, several other parameters have also been proposed, including power spectral deviation (PSD), linear prediction coefficients (LPCs) and likelihood ratio (LR) based on statistical models [1]–[4]. Although these parameters are quite effective in expressing the characteristics of a speech signal, the performance of the VAD using such parameters remains poor in adverse environments. Therefore, a feature parameter that can sufficiently specify the characteristics of speech and be robust in noisy environments is urgently needed to improve the performance of the VAD algorithm.

In this letter, we propose a novel approach to the VAD algorithm in which global speech absence probability (GSAP) [3], based on Teager energy (TE) [5], [6], is derived to improve the performance of VAD in various noisy environments. Statistical model-based GSAP is one of the feature parameters which is widely adopted in the decision rule for VAD, and it is used as the smoothing parameter for updating the noise signal in the speech enhancement algorithm. In addition, a TE operator which can provide better characteristics of speech from noise is employed to suppress the influence of noise signals. In practice, it has been experimentally observed that the TE operator can enhance the ability to discriminate between speech and noise and further suppress the noise components. Therefore, we utilize TE-based GSAP as a feature parameter for VAD to derive better performance of the speech enhancement algorithm by the proposed VAD method in noisy environments. The performance of the proposed algorithm is evaluated by an objective comparison and it is consequently demonstrated to be better than those of conventional methods.

Manuscript received March 12, 2012.
Manuscript revised June 28, 2012.

[†]The authors are with the Department of Electronic Engineering, Inha University, Incheon, Korea.
a) E-mail: sanglee@inha.ac.kr

DOI: 10.1587/transinf.E95.D.2568

2. Review of the Teager Energy Operator

In this section, we briefly review the Teager energy (TE) operator, which is employed to suppress the influence of noise signals. In practice, since corrupted noise is effectively suppressed by the TE operator, the TE operator can provide better ability to discriminate speech characteristics from noise. The TE operator is easily implemented through the time domain and is defined as given by [5], [6]:

$$\Psi_c[s(t)] = [\dot{s}(t)]^2 - s(t)\ddot{s}(t) \tag{1}$$

where $s(t)$ is a continuous-time signal, and $\dot{s} = ds/dt$. In discrete-time, the TE operator can be approximated by

$$\Psi[s(n)] = s(n)^2 - s(n+1)s(n-1) \tag{2}$$

where $s(n)$ is a discrete-time signal. In practice, a clean speech signal $s(n)$ is corrupted by the additive noise signal $d(n)$. Assuming that speech is degraded by an uncorrelated additive noise, the observed noisy speech signal $y(n)$ is given by

$$y(n) = s(n) + d(n) \tag{3}$$

where $s(n)$ and $d(n)$ are zero mean and independent. Based on this, the TE of $y(n)$ is obtained by

$$\Psi[y(n)] = \Psi[s(n)] + \Psi[d(n)] + 2\tilde{\Psi}[s(n), d(n)] \tag{4}$$

where, $\Psi[y(n)]$, $\Psi[s(n)]$ and $\Psi[d(n)]$ are the TE of noisy speech, clean speech and additive noise, respectively. Also, the cross-TE $\tilde{\Psi}[s(n), d(n)]$ of $s(n)$ and $d(n)$ can be computed as follows:

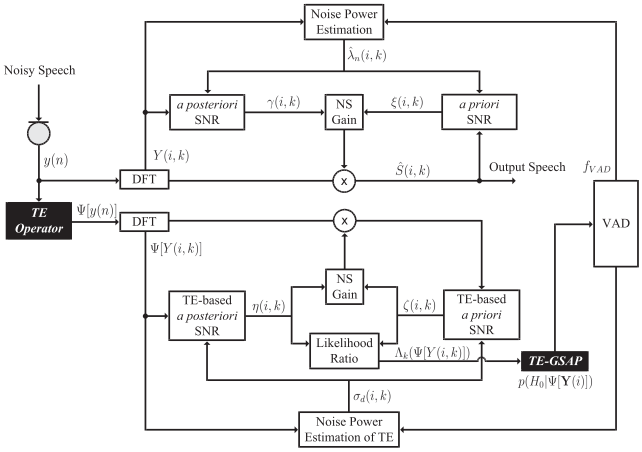

Fig. 1 Block diagram of the proposed VAD algorithm.

$$\tilde{\Psi}[s(n),d(n)] = s(n)d(n) - 0.5s(n-1)d(n+1) - 0.5s(n+1)d(n-1). \quad (5)$$

Since $s(n)$ and $d(n)$ are zero mean and independent, the expected value of the cross-TE is equal to zero. Thus, the expected value of $\Psi[y(n)]$ is approximated as

$$E\{\Psi[y(n)]\} = E\{\Psi[s(n)]\} + \{\Psi[d(n)]\}. \quad (6)$$

In fact, the TE of clean speech is much higher than that of noise. Therefore, $\Psi[d(n)]$ is negligible compared to $\Psi[s(n)]$ as given by [5], [6]

$$E\{\Psi[y(n)]\} \approx E\{\Psi[s(n)]\}. \quad (7)$$

For this reason, the TE operator can suppress the noise signal and provide better discrimination between speech and noise. Therefore, TE-based feature parameters can enhance the ability to discriminate speech and noise for effective VAD in noisy environments.

3. Proposed VAD Using Global Speech Absence Probability Based on Teager Energy

In the previous section, it was noted that the TE operator provides better ability to discriminate between speech and noise by suppressing the noise signal. Based on this, we propose a novel VAD algorithm using GSAP based on the TE. GSAP based on the likelihood ratio employing statistical models has been shown as a good feature parameter for detecting the presence of speech in noisy environment [3]. However, the performance of VAD algorithms using GSAP as the feature parameter remains poor in the adverse noise conditions. Therefore, in the proposed method, we derive an improved feature parameter based on GSAP by taking advantage of the TE, in contrast with the conventional GSAP-based method. Figure 1 presents an overall block diagram of the proposed VAD algorithm which utilizes the proposed TE-based GSAP (TE-GSAP). For this, we first assume that the following two hypotheses, H_0 and H_1, indicate speech absence and presence from the signal derived from the TE operator [3]:

$$H_0 : \text{speech absent} : \Psi[\mathbf{Y}(i)] = \Psi[\mathbf{D}(i)] \quad (8)$$

$$H_1 : \text{speech present} : \Psi[\mathbf{Y}(i)] = \Psi[\mathbf{D}(i)] + \Psi[\mathbf{S}(i)] \quad (9)$$

where $\Psi[\mathbf{Y}(i)] = [\Psi[Y(i,1)], \Psi[Y(i,2)], \ldots, \Psi[Y(i,M)]]$ represents the Fourier domain spectra of noisy speech based on the TE with a frame index i, and M(=16) is the total band size of each frame. Here, $\Psi[Y(i,k)]$ denotes an estimate of the Fourier spectrum of noisy speech based on the TE compared to the conventional Fourier spectrum $Y(i,k)$ with a time index i and frequency index k in the discrete Fourier transform (DFT) domain. Also, $\Psi[\mathbf{D}(i)] = [\Psi[D(i,1)], \Psi[D(i,2)], \ldots, \Psi[D(i,M)]]$ and $\Psi[\mathbf{S}(i)] = [\Psi[S(i,1)], \Psi[S(i,2)], \ldots, \Psi[S(i,M)]]$, respectively, represent the Fourier spectra derived from the TE of the noise and clean speech signal. Under the assumption that $\Psi[D(i,k)]$ and $\Psi[S(i,k)]$ are statistically independent, $\Psi[D(i,k)]$ and $\Psi[S(i,k)]$ are characterized by zero-mean complex Gaussian distributions such that [3]:

$$p(\Psi[Y(i,k)]|H_0) = \frac{1}{\pi\sigma_d(i,k)} \exp\Big[- \frac{|\Psi[Y(i,k)]|^2}{\sigma_d(i,k)} \Big] \quad (10)$$

$$p(\Psi[Y(i,k)]|H_1) = \frac{1}{\pi(\sigma_s(i,k) + \sigma_d(i,k))} \exp\Big[- \frac{|\Psi[Y(i,k)]|^2}{\sigma_s(i,k) + \sigma_d(i,k)} \Big] \quad (11)$$

where $\sigma_s(i,k)$ and $\sigma_d(i,k)$ are the variance of the speech and estimated noise based on the TE, respectively. Accordingly, the TE-GSAP $p(H_0|\Psi[\mathbf{Y}(i)])$ is derived from Bayes' rule, such that [3]

$$p(H_0|\Psi[\mathbf{Y}(i)]) = \frac{p(\Psi[\mathbf{Y}(i)]|H_0)p(H_0)}{p(\Psi[\mathbf{Y}(i)]|H_0)p(H_0) + p(\Psi[\mathbf{Y}(i)]|H_1)p(H_1)} \quad (12)$$

where $p(H_0)(= 1 - p(H_1))$ represents the *a priori* probability of speech absence. Since the spectral component in each frequency bin is assumed to be statistically independent, (12) can be rewritten as

$$p(H_0|\Psi[\mathbf{Y}(i)]) = \frac{p(H_0)\prod_{k=1}^{M} p(\Psi[Y(i,k)]|H_0)}{p(H_0)\prod_{k=1}^{M} p(\Psi[Y(i,k)]|H_0) + p(H_1)\prod_{k=1}^{M} p(\Psi[Y(i,k)]|H_1)} = \frac{1}{1 + q\prod_{k=1}^{M} \Lambda_k(\Psi[Y(i,k)])} \quad (13)$$

in which $q = p(H_1)/p(H_0)$ is set to 0.0625 [3], and $\Lambda_k(\Psi[Y(i,k)])$ is the likelihood ratio computed in the kth frequency bin, as given by [4]:

$$\Lambda_k(\Psi[Y(i,k)]) = \frac{p(\Psi[Y(i,k)]|H_1)}{p(\Psi[Y(i,k)]|H_0)} = \frac{1}{1+\zeta(i,k)} \exp\Big[\frac{\eta(i,k)\zeta(i,k)}{1+\zeta(i,k)}\Big] \quad (14)$$

Table 1 Comparison of total error rate (TER), false rejection rate (FRR), and false acceptance rate (FAR) among the methods of the conventional and the proposed technique.

Environments		I.Y. Soon			J. Sohn			GSAP			Proposed		
Noise	SNR (dB)	TER	FRR	FAR	TER	FRR	FAR	TER	FRR	FAR	TER	FRR	FAR
Wihte	0	39.36	38.21	40.95	41.15	35.63	48.85	42.71	36.85	50.86	41.34	36.26	48.42
	5	32.47	26.35	40.97	33.29	26.36	42.93	32.21	25.58	41.43	31.96	25.64	40.76
	10	27.87	18.83	40.46	28.93	19.19	42.48	24.40	17.93	33.41	22.08	17.82	28.01
	15	19.96	14.97	26.91	23.67	14.4	36.58	18.93	12.78	27.49	16.83	9.8	26.62
Babble	0	37.14	27.41	50.68	37.86	27.89	51.73	38.87	31.28	49.43	34.88	24.54	49.28
	5	30.82	19.85	46.09	32.08	19.49	49.6	32.2	21.54	47.05	25.41	18.93	34.42
	10	25.48	12.98	42.89	27.14	12.83	47.08	23.85	13.24	38.62	19.57	11.01	31.5
	15	21.57	8.44	39.84	22.43	8.6	41.68	17.74	7.27	32.32	16.43	5.99	30.97
Vehicle	0	9.29	8.25	10.74	9.37	8.35	10.79	8.34	5.02	12.96	9.39	8.38	10.8
	5	6.97	5.52	8.99	6.93	5.62	8.75	6.61	3.1	11.51	6.75	5.38	8.67
	10	4.68	3.73	6.01	4.95	3.71	6.68	3.17	2.58	3.99	4.95	3.71	6.67
	15	3.47	3.15	3.92	3.48	2.85	4.36	3.06	2.5	3.83	4.19	2.84	6.06

where the TE-based *a posteriori* signal-to-noise ratio (SNR) $\eta(i,k)$ and the TE-based *a priori* SNR $\zeta(i,k)$ are defined by

$$\eta(i,k) \equiv \frac{|\Psi[Y(i,k)]|^2}{\sigma_d(i,k)}, \qquad \zeta(i,k) \equiv \frac{\sigma_s(i,k)}{\sigma_d(i,k)} \tag{15}$$

in which $\zeta(i,k)$ is estimated by the well-known decision-directed approach [1], [7].

Finally, in the proposed VAD algorithm, speech segments are decided by the decision rule as follows:

$$f_{VAD} = \begin{cases} \text{speech}, & \text{if } p(H_0|\Psi[\mathbf{Y}(i)]) < T \\ \text{nonspeech}, & \text{otherwise} \end{cases} \tag{16}$$

where the threshold value T is experimentally determined to 0.3 based on a large number of noisy speech data samples which contain a variety of noises and SNR conditions. Figure 2 (d) shows the estimates of GSAP obtained by the conventional method and by the proposed TE-GSAP method. The conventional method based-GSAP represented by the dashed line is derived from a noisy speech signal as shown in Fig. 2 (a), and the proposed TE-GSAP represented by the solid line in Fig. 2 (d) is derived by employing the enhanced noisy speech based on the TE as shown in Fig. 2 (c). Figure 2 (d) clearly shows the difference between the conventional GSAP and the proposed GSAP derived from the signal enhanced by the TE operator. From Fig. 2 (d), we can see that the GSAP estimation of the conventional scheme insufficiently discriminates between speech and noise since the conventional GSAP estimate is sensitive to noise. On the contrary, it can be seen that in given noisy conditions, TE-GSAP estimated in the proposed method performs well by taking advantage of the better characteristic of speech against noise through the TE operator.

Based on this, the noise power estimate $\hat{\lambda}_n(i,k)$ can be updated during nonspeech with the following averaging rule:

$$\hat{\lambda}_n(i,k) = \alpha_n \hat{\lambda}_n(i-1,k) + (1-\alpha_n)|Y(i,k)|^2 \tag{17}$$

in which the smoothing parameter α_n is set at 0.9 and $\sigma_d(i,k)$ in (15) is also derived based on $\Psi[Y(i,k)]$ by utilizing the proposed update routine.

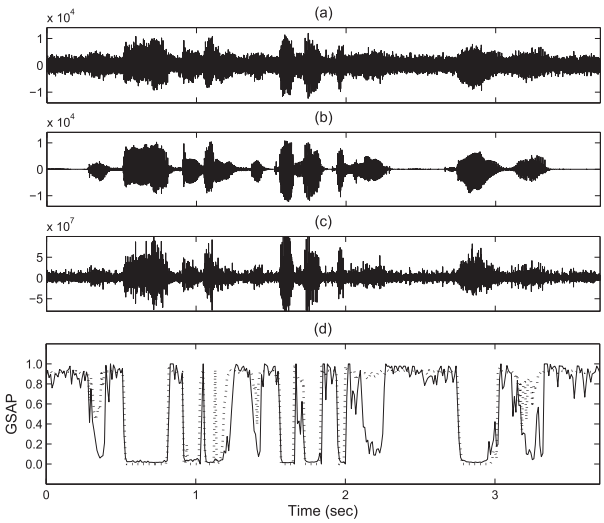

Fig. 2 Comparison of GSAP (white noise, SNR=0 dB) (a) Noisy speech waveform (b) Clean speech waveform (c) TE waveform (d) GSAP: the conventional method (dashed line), the proposed TE-based method (solid line).

4. Experimental Results

The proposed VAD method was adopted for the noise suppression algorithm using the suppression gain based on MMSE (minimum mean square error) estimation [7] and was evaluated with objective comparison experiments under various noise conditions.

For the test material in terms of detection accuracy (%) [8], we formed 456 s speech data 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. Also, to consider various noise environments, three types of noise sources white, babble, and vehicle noise from the NOISEX-92 database were added to the clean speech waveform at SNRs of 0, 5, 10 and 15 dB. Table 1 including TER (total error rate), FRR (false rejection rate), and FAR (false acceptance rate) shows comparative results for the soft decision-based approach that represents the probability of speech absence in speech enhancement method by I.Y. Soon *et al.* [2], the LR-based method by J.

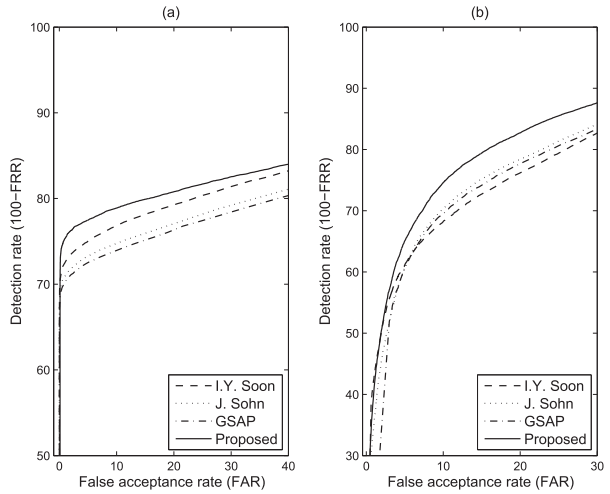

Fig. 3 (a) ROC curve for the white noise at 10 dB SNR (b) ROC curve for the babble noise at 5 dB SNR.

Table 2 PESQ scores obtained from the proposed VAD algorithm based on proposed TE-GSAP with those yielded by the conventional methods under various noise environments.

Environments		PESQ			
Noise	SNR (dB)	I.Y. Soon	J. Sohn	GSAP	Proposed
White	0	1.611	1.635	1.621	1.649
	5	2.110	2.109	2.103	2.112
	10	2.450	2.448	2.447	2.452
	15	2.752	2.753	2.753	2.755
Babble	0	1.944	1.950	1.949	1.972
	5	2.334	2.337	2.337	2.353
	10	2.670	2.661	2.663	2.667
	15	2.962	2.954	2.956	2.956
Vehicle	0	3.136	3.135	3.134	3.136
	5	3.433	3.432	3.432	3.434
	10	3.687	3.687	3.687	3.691
	15	3.944	3.944	3.944	3.945

Sohn *et al.* [1], GASP [3] and the proposed approach. From the results, it is evident that the proposed VAD algorithm outperformed or at least was comparable to the conventional methods in terms of overall detection accuracy under the given noise conditions. This fact could be confirmed by Fig. 3 showing the receiver operating characteristics (ROC) which are insensitive to parameter tuning since it is a trade-off between detection rate (100-FRR) and FAR [8]. Based on this, we can see the overall performance differences of the aforementioned methods. From the figure, it can be seen that the proposed TE-based VAD yielded overall higher performance than the conventional method.

Also, for the comparison of an objective speech quality, we evaluated the objective quality of the output signal as obtained by the NS algorithm in which the VAD algorithms based on the conventional and proposed scheme are adopted. For the test material, ninety test phrases with a sampling rate of 8 kHz were used as the experimental data. Each phrase consisted of two different meaningful sentences and lasted 8 sec. In order to evaluate the speech quality, we adopted the perceptual evaluation of speech quality (PESQ, ITU-T P.862) which is a worldwide applied industry standard for objective speech quality testing [9]. The results of the PESQ scores for the evaluated methods are presented in Table 2. Table 2 illustrates that the proposed approach outperformed comparable to the conventional methods under the given noise conditions and achieves a meaningful performance improvement over the conventional methods especially at low SNRs.

5. Conclusion

In this paper, we have proposed a novel VAD algorithm using TE-based GSAP. The GSAP estimate derived from the enhanced input noisy signal by the TE operator is applied to the VAD algorithm as a robust feature parameter. The performance of the proposed algorithm has been found to be superior to that of the conventional technique through objective evaluation tests.

Acknowledgement

This work was supported by grant No. SS100022 by Seoul R&BD Program and Key Research Institute Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0005858).

References

[1] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Process. Lett., vol.6, no.1, pp.1–3, Jan. 1999.
[2] I.Y. Soon, S.N. Koh, and C. k. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," Signal Process., vol.75, pp.151–159, 1999.
[3] N.S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," IEEE Signal Process. Lett., vol.7, no.5, pp.108–110, May 2000.
[4] M. Fujimoto, K. Ishizuka, and T. Nakatani, "Study of integration of statistical model-based voice activity detection and noise suppression," Proc. Interspeech '08, pp.2008–2011, 2008.
[5] F. Jabloun, A.E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," IEEE Signal Process. Lett., vol.6, no.10, pp.259–261, 1999.
[6] K.C. Wang and Y.H. Tsai, "Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy," Second International Symposium on Universal Communication 2008, pp.423–428, Dec. 2008.
[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, no.6, pp.1109–1121, Dec. 1984.
[8] J. Ramirez, J.C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," IEEE Trans. Speech Audio Process., vol.13, no.6, pp.1119–1129, Nov. 2005.
[9] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.