symmetry

MDPI

Article

Power Spectral Deviation-Based Voice Activity Detection Incorporating Teager Energy for Speech Enhancement

Sang-Kyun Kim [1], Sang-Ick Kang [1], Young-Jin Park [2], Sanghyuk Lee [3] and Sangmin Lee [1,*]

[1] Department of Electronic Engineering, Inha University, Incheon 402-751, Korea; greenwhity@nate.com (S.-K.K.); rkdtkddlr@gmail.com (S.-I.K.)
[2] Korea Electrotechnology Research Institute (KERI), 111 Hanggaul ro, Sangrok Gu, An-San shi, Kyunggi Do 426-170, Korea; yjpark@keri.re.kr
[3] Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China; Sanghyuk.Lee@xjtlu.edu.cn
* Correspondence: sanglee@inha.ac.kr; Tel.: +82-70-8251-1549

Academic Editor: Ka Lok Man
Received: 11 April 2016; Accepted: 1 July 2016; Published: 6 July 2016

Abstract: In this paper, we propose a robust voice activity detection (VAD) algorithm to effectively distinguish speech from non-speech in various noisy environments. The proposed VAD utilizes power spectral deviation (PSD), using Teager energy (TE) to provide a better representation of the PSD, resulting in improved decision performance for speech segments. In addition, the TE-based likelihood ratio and speech absence probability are derived in each frame to modify the PSD for further VAD. We evaluate the performance of the proposed VAD algorithm by objective testing in various environments and obtain better results that those attained by of the conventional methods.

Keywords: power spectral deviation; Teager energy; speech absence

1. Introduction

In speech signal processing, such as speech recognition and noise suppression (NS), the role of voice activity detection (VAD) algorithms is crucial for better performance in noisy environments [1]. To determine the presence or absence of speech, VAD algorithms are usually designed using certain decision rules that are derived from feature parameters that distinguish speech segments from other waveforms. To improve VAD performance, a feature parameter that can sufficiently specify the characteristics of speech and be robust in noisy environments is urgently needed. Traditionally, parameters that specify the characteristics of speech are based on short-time energy, spectral energy, or zero-crossing rate (ZCR). All of these parameters, however, are 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), likelihood ratio (LR), and pitch [2,3]. Although these parameters are quite effective in expressing the characteristics of a speech signal, VAD performance using such parameters remains poor in adverse environments.

In this paper, we propose a novel approach to VAD algorithm, in which the PSD based on Teager energy (TE) [4,5], is derived in order to represent the improvement difference between speech and a noise signal in temporal statistical variations, instead of conventional PSD, which is used as one of the features for VAD in the IS-127 noise suppression algorithm and is known to provide robust performance under noisy conditions [2]. It has been experimentally observed that the TE operator can enhance discriminability between speech and noise and further suppress noise components [5,6]. Additionally, in the proposed method, the LR based on TE is used as a weighting factor and the speech

absence probability (SAP) derived from the TE-based LR is adopted as a smoothing parameter for further improved PSD modification in various noisy environments [3,7]. The performance of the proposed algorithm is evaluated by an objective comparison and is demonstrated to be better than the conventional method.

2. Review of Teager Energy

In this section, we briefly review the Teager energy (TE) operator, which is used successfully in various speech applications. For a discrete speech signal $x(n)$, the TE operation is defined, as given by [4,5]:

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

In practice, the clean speech signal $x(n)$ is corrupted by the additive noise signal $d(n)$. Assuming that speech is degraded by uncorrelated additive noise, the observed noisy speech signal $y(n)$ is given by:

$$y(n) = x(n) + d(n) \tag{2}$$

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

$$\Psi_d[y(n)] = \Psi_d[x(n)] + \Psi_d[d(n)] + 2\widetilde{\Psi}_d[x(n), d(n)] \tag{3}$$

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

$$\widetilde{\Psi}_d[x(n), d(n)] = x(n)d(n) - 0.5x(n-1)d(n+1) - 0.5x(n+1)d(n-1) \tag{4}$$

Since $x(n)$ and $d(n)$ are zero-mean and independent, the expected value of the cross-TE is zero. Thus, the expected value of $\Psi_d[y(n)]$ is approximated as follows:

$$E\{\Psi_d[y(n)]\} = E\{\Psi_d[x(n)\} + E\{\Psi_d[d(n)\} \tag{5}$$

In fact, the TE of clean speech is much higher than that of noise. Therefore, $\Psi_d[d(n)]$ is negligible compared to $\Psi_d[x(n)]$ as given by [4]:

$$E\{\Psi_d[y(n)]\} \approx E\{\Psi_d[x(n)\} \tag{6}$$

For this reason, the corrupted noise signal can be suppressed by the TE operator and feature parameters that allow better discrimination between speech and noise to be obtained from the enhanced signal.

3. Proposed VAD Algorithm Based on the Power Spectral Deviation of Teager Energy

In the previous section, we noted that the TE operator provides better discriminability of speech from noise compared with other methods. Based on this, we propose a novel VAD algorithm using PSD based on TE. We consider TE-based PSD (TE-PSD) for the feature parameter of VAD, in which the TE-based LR and SAP are introduced to modify the proposed PSD. Figure 1 presents an overall block diagram of the proposed VAD algorithm. For this, the proposed TE-PSD $\Delta_{TE}(i)$ is derived by:

$$\Delta_{TE}(i) = \log_{10}\left(\frac{\beta(i)}{M}\sum_{k=1}^{M}\left|Y_{TE}(i,k) - \overline{Y}_{TE}(i,k)\right|\right) \tag{7}$$

where M (=16) denotes the total band size of each frame and $Y_{TE}(i,k)$ with time index i and frequency index k represents the estimated power spectrum $|\Psi[Y(i,k)]|^2$ of noisy speech based on TE. Here, $\Psi[Y(i,k)]$ denotes an estimate of the Fourier spectrum of noisy speech based on TE, compared to

the conventional Fourier spectrum $Y(i,k)$ in the discrete Fourier transform domain. In Equation (7), $\beta(i)$ is the weighting factor used to achieve improved performance on the PSD and is derived from $\Psi[Y(i,k)]$ as follows [8]:

$$\beta(i) = \prod_{k=1}^{M} \Lambda(\Psi[Y(i,k)]) \tag{8}$$

where $\Lambda(\Psi[Y(i,k)])$ is the TE-based LR computed in the kth frequency bin under the assumption that noise and speech are statistically independent and characterized by zero-mean complex Gaussian distributions as given by [3,8]:

$$\Lambda(\Psi[Y(i,k)]) \frac{p(\Psi[Y(i,k)]|H_1)}{p(\Psi[Y(i,k)]|H_0)} = \frac{1}{1+\varsigma(i,k)} \exp\left[\frac{\eta(i,k)\varsigma(i,k)}{1+\varsigma(i,k)}\right] \tag{9}$$

where the TE-based *a posteriori* signal-to-noise ratio (SNR) $\eta(i,k)$ and the TE-based a priori SNR $\varsigma(i,k)$ are defined by [3,8]:

$$\eta(i,k) \equiv \frac{|\Psi[Y(i,k)]|^2}{\sigma_d(i,k)}, \varsigma(i,k) \equiv \frac{\sigma_x(i,k)}{\sigma_d(i,k)} \tag{10}$$

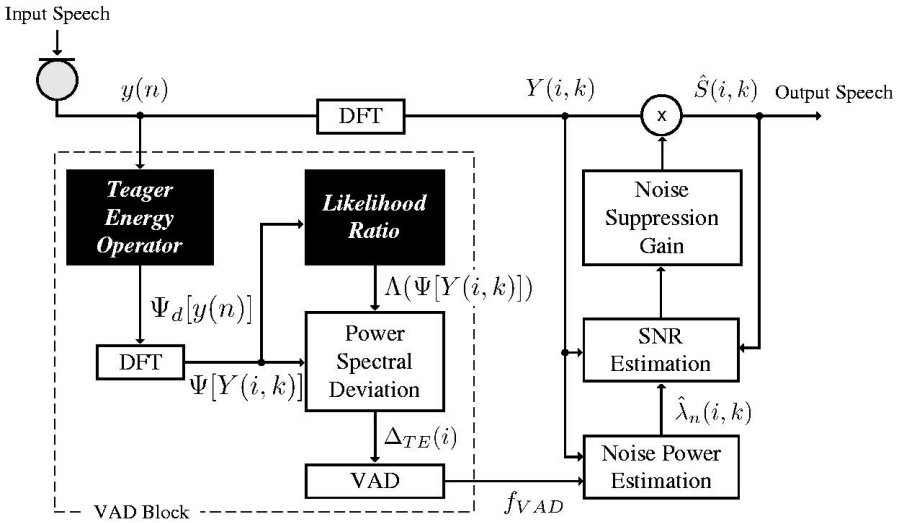

Figure 1. Block diagram of the proposed voice activity detection (VAD) algorithm.

In Equation (10), $\sigma_x(i,k)$ and $\sigma_d(i,k)$ are the variance of the speech and estimated noise, respectively, based on TE. Additionally, $\overline{Y}_{TE}(i,k)$ in Equation (7) is the long-term smoothed power spectral estimate calculated during the previous sub-band. The proposed VAD algorithm is further improved by a smoothing parameter that uses SAP $p(H_0|\Psi[Y(i)])$ based on TE for modifying $\overline{Y}_{TE}(i,k)$ as follows:

$$\overline{Y}_{TE}(i,k) = (1 - p(H_0|\Psi[Y(i)]))\overline{Y}_{TE}(i-1,k) + p(H_0|\Psi[Y(i)])Y_{TE}(i,k) \tag{11}$$

where $H_0 (= 1 - H_1)$ indicates speech absence and $p(H_0|\Psi[Y(i)])$, which can be derived from Bayes' rule is simply obtained by using the previous estimated weighting factor $\beta(i)$, such that [8]:

$$\begin{aligned} & p(H_0|\Psi[Y(i)]) \\ &= \frac{p(H_0)\prod_{k=1}^{M} p(\Psi[Y(i,k)]|H_0)}{p(H_0)\prod_{k=1}^{M} p(\Psi[Y(i,k)]|H_0) + p(H_1)\prod_{k=1}^{M} p(\Psi[Y(i,k)]|H_1)} \\ &= \frac{1}{1+q\prod_{k=1}^{M}\Lambda(\Psi[Y(i,k)])} = \frac{1}{1+q\beta(i)} \end{aligned} \tag{12}$$

in which $q = p(H_1)/p(H_0)$ is set to 0.0625, which is the appropriate value chosen from experiments in order to obtain good performance in various noisy environments [8].

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

$$f_{VAD} = \begin{cases} speech, & if\ \Delta_{TE}(i) > T \\ nonspeech, & otherwise \end{cases} \quad (13)$$

In Figure 2, we give an example of the feature parameters estimated by the conventional method and by the proposed TE-PSD method. From Figure 2d,e, we can see that the feature parameters of the conventional scheme insufficiently discriminate between speech and noise, since the conventional LR [3] and PSD [2] are sensitive to noise. In contrast, it is evident that in noisy conditions, the proposed TE-based method yielded a better representation of the VAD decision compared with the conventional methods. Based on these results, from Figure 2f, it can be seen that the proposed VAD method performs well by taking advantage of the TE-based approach and allows the noise power estimate $\hat{\lambda}_n(i,k)$ to be updated by the decision rule in Equation (13) during non-speech with the following averaging rule:

$$\hat{\lambda}_n(i,k) = \alpha_n \hat{\lambda}_n(i-1,k) + (1-\alpha_n)|Y(i,k)|^2 \quad (14)$$

in which the smoothing parameter α_n is set to 0.9.

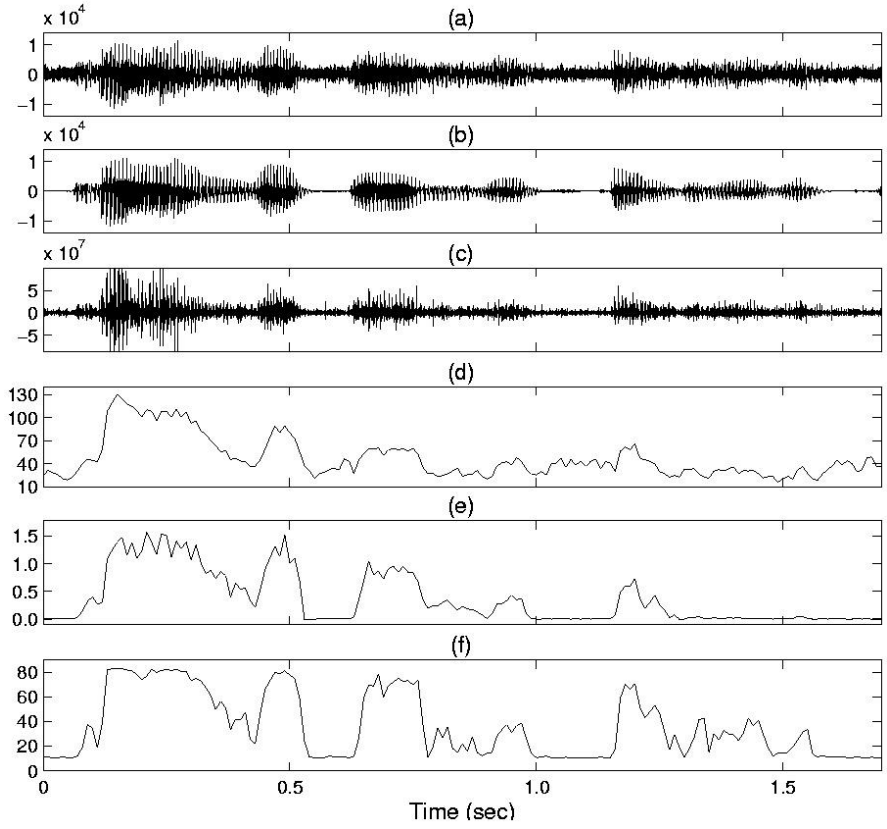

Figure 2. (**a**) Noisy speech waveform (office noise, signal-to-noise ratio (SNR) = 0 dB); (**b**) Clean speech waveform; (**c**) Teager energy (TE) waveform; (**d**) VAD based on power spectral deviation (PSD); (**e**) VAD based on likelihood ratio (LR); (**f**) VAD based on the proposed method.

4. Experiments and Results

The proposed VAD method was adopted for NS algorithms using suppression gain based on minimum mean square error estimation [9] and was evaluated by objective comparison experiments under various noise conditions. For the test material [10], 456 s of speech data were recorded by four males and four females and were sampled at 8 kHz. To evaluate VAD performance, we first made reference decisions on the clean speech material by labeling it manually at each 10 ms frame. The proportion of hand-marked speech frames was 57.1% and consisted of 44.0% voiced sounds and 13.1% unvoiced sounds. In addition, to consider various noise environments, three types of noise sources (babble, car, and office noises) were added to the clean speech waveform at SNRs of 0, 5, 10, and 15 dB.

Table 1 shows comparative results, including total error rate (TER), false rejection rate (FRR) and false acceptance rate (FAR), for the proposed method and the CMAP-SG algorithm [11] which is an LR-based VAD algorithm that incorporates spectral variation. In addition, to show that the performance of the proposed method is acceptable in practice, the result of the well-known standard VAD algorithm, ITU-T G.729B Appendix III, is included for each condition [12]. From these results, it is evident that the proposed VAD algorithm outperformed or was comparable to conventional methods in terms of overall detection accuracy under the given noise conditions. This fact could be confirmed by Figures 3–5, showing the receiver operating characteristics (ROC) that are insensitive to parameter tuning, since they are a trade-off between detection rate (100-FRR) and FAR [10]. Based on these characteristics, we can see the overall performance differences of the previously discussed methods. From the Figures 3–5, the proposed TE-based VAD yielded higher or equivalent performance compared with the conventional method.

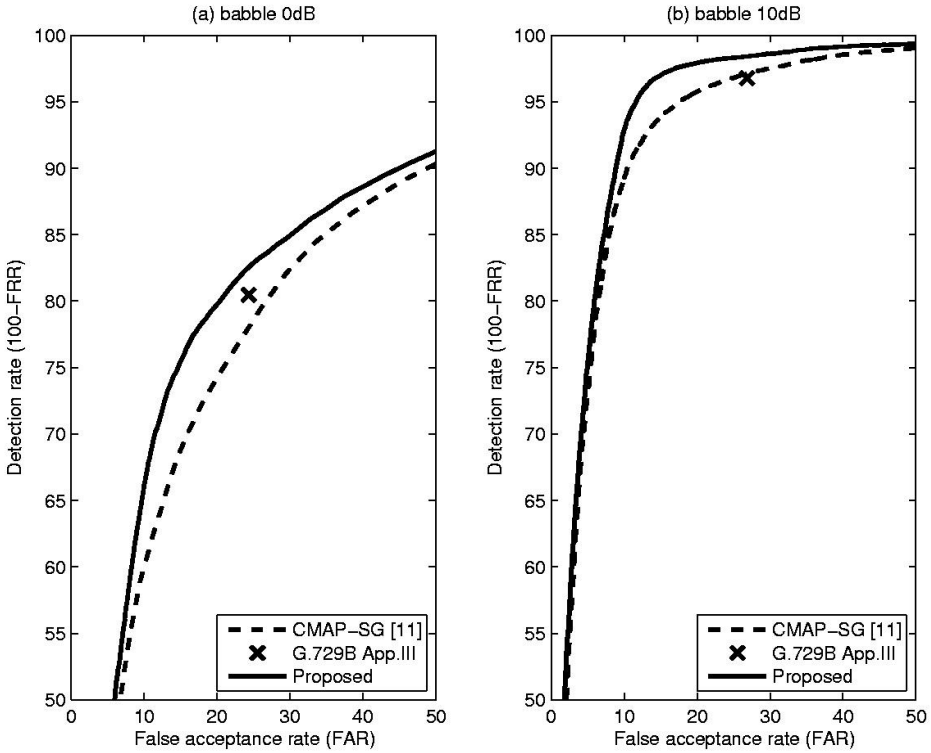

Figure 3. (**a**) Receiver operating characteristics (ROC) curve for the babble at 0 dB SNR; (**b**) ROC curve for the babble at 10 dB SNR.

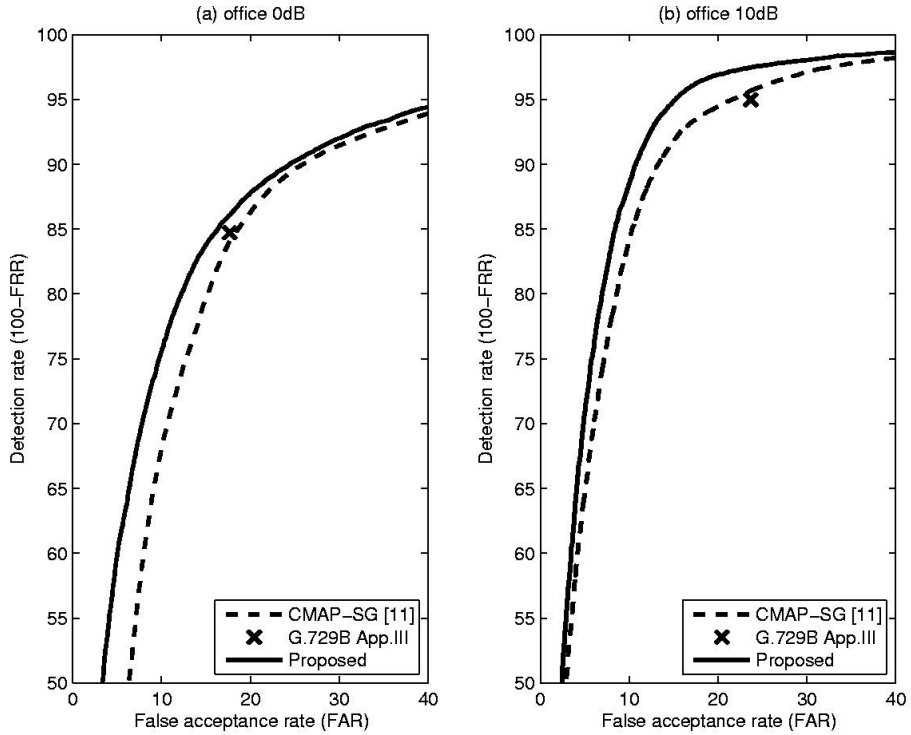

Figure 4. (**a**) ROC curve for the office at 0 dB SNR; (**b**) ROC curve for the office at 10 dB SNR.

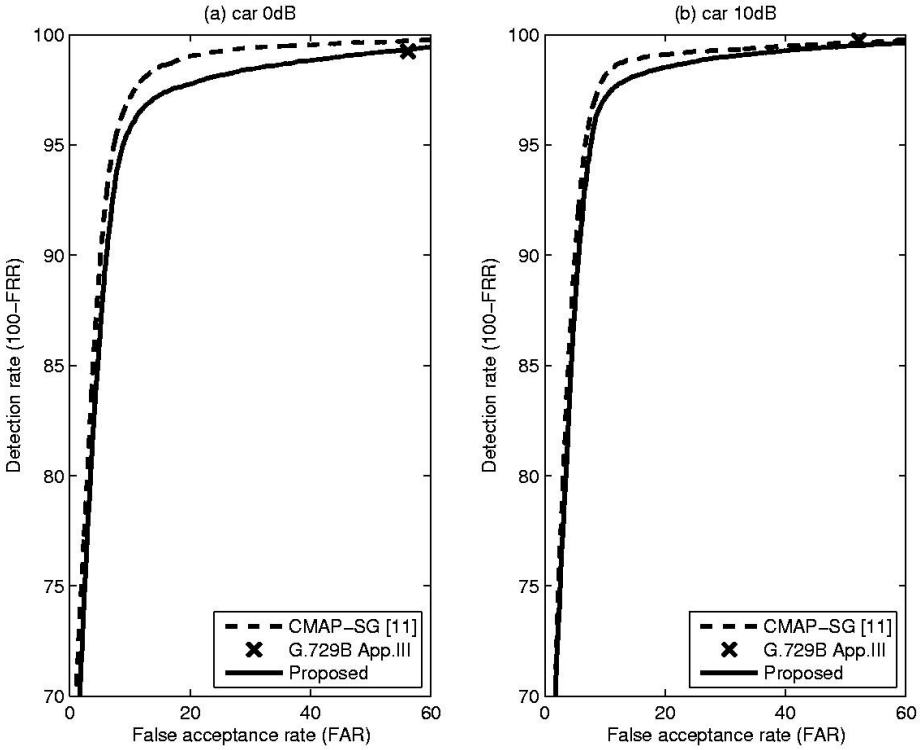

Figure 5. (**a**) ROC curve for the car at 0 dB SNR; (**b**) ROC curve for the car at 10 dB SNR.

Table 1. Comparison of total error rate (TER), false rejection rate (FRR) and false acceptance rate (FAR) among the method of the CMAP-SG, the G.729B App. III and the proposed TE-PSD technique.

Environments		CMAP-SG			G.729B App. III			Proposed		
Noise	**SNR (dB)**	**TER**	**FRR**	**FAR**	**TER**	**FRR**	**FAR**	**TER**	**FRR**	**FAR**
Babble	0	22.76	27.70	17.95	21.55	19.54	24.34	19.52	22.58	16.69
	5	16.86	18.15	15.02	16.24	7.43	28.50	15.04	17.03	14.50
	10	10.09	8.97	11.28	13.11	3.23	26.86	8.30	4.89	11.79
	15	8.21	7.16	10.42	12.45	1.67	27.47	7.53	4.51	10.24
Office	0	16.94	13.68	21.47	16.28	15.28	17.66	15.32	15.90	15.27
	5	15.78	11.93	21.13	14.87	9.13	22.87	12.07	9.63	14.53
	10	10.84	6.79	16.48	12.84	5.03	23.70	9.77	6.87	12.84
	15	8.16	2.81	15.57	12.11	2.61	25.35	8.02	5.36	11.05
Car	0	6.02	1.83	11.85	23.94	0.75	56.23	6.82	5.07	8.78
	5	5.62	1.78	10.94	22.44	0.56	52.91	6.53	4.39	8.65
	10	5.51	1.69	10.82	21.97	0.25	52.21	6.11	2.96	9.84
	15	5.34	1.60	10.53	20.34	0.17	48.41	5.86	2.84	8.94

In addition, for an objective comparison of speech quality, we evaluated the objective quality of the output signal as obtained by the NS algorithm from which the VAD algorithms based on the conventional and proposed scheme are adopted. For the test material, 90 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 speech quality, we utilized the perceptual evaluation of speech quality (PESQ, ITUT P.862) [13], which is a worldwide applied industry standard for objective speech quality testing. The results of the PESQ scores for the evaluated methods are presented in Table 2. Table 2 illustrates that the proposed approach performed comparably to the conventional methods under the given noise conditions and achieved a meaningful performance improvement over conventional methods, especially for low SNRs.

Table 2. Perceptual evaluation of speech quality (PESQ) scores obtained from the proposed VAD algorithm based on proposed TE-PSD with those yielded by the conventional method under various noise environments.

Environments		PESQ	
Noise	**SNR (dB)**	**CMAP-SG [11]**	**Proposed**
Office	0	1.848 ± 0.013	1.891 ± 0.016
	5	2.193 ± 0.009	2.225 ± 0.012
	10	2.556 ± 0.006	2.583 ± 0.008
	15	2.841 ± 0.004	2.859 ± 0.006
Babble	0	1.982 ± 0.018	2.055 ± 0.023
	5	2.362 ± 0.016	2.425 ± 0.022
	10	2.674 ± 0.009	2.711 ± 0.011
	15	2.965 ± 0.008	2.997 ± 0.009
Car	0	3.152 ± 0.002	3.151 ± 0.002
	5	3.443 ± 0.001	3.439 ± 0.002
	10	3.694 ± 0.001	3.692 ± 0.001
	15	3.944 ± 0.001	3.941 ± 0.001

5. Conclusions

In this paper, we have proposed a novel VAD algorithm using TE-based PSD. Furthermore, to improve the performance of the proposed algorithm, TE-based LR and SAP are applied as modification of TE-based PSD. Compared to conventional VAD algorithms (CMAP-SG [11] and G.729B App III),

the performance of the proposed technique under various noise environments was superior in objective tests of speech detection accuracy, ROCs, and PESQ.

Acknowledgments: This research was supported by the KERI Primary Research Program through the Korea Research Council for Industrial Science & Technology funded by the Ministry of Science, Information & Communication Technology and Future Planning (No. 16-12-N0101-44).

Author Contributions: Sang-Kyun Kim wrote this manuscript; Sang-Ick Kang, Young-Jin Park, Sanghyuk Lee and Sangmin Lee contributed to the writing, direction and content and also revised the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Karray, L.; Mokbel, C.; Monne, J. Solutions for robust speech/non-speech detection in wireless environment. In Proceedings of the IEEE 4th Workshop, Interactive Voice Technology for Telecommunications Applications IVITA'98, Torino, Italy, 29–30 September 1998; pp. 166–170.
2. TIA/EIA/IS-127. *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, 1996.
3. Sohn, J.; Kim, N.S.; Sung, W. A statistical model-based voice activity detection. *IEEE Signal Process. Lett.* **1999**, *6*, 1–3. [CrossRef]
4. Jabloun, F.; Cetin, A.E.; Erzin, E. Teager energy based feature parameters for speech recognition in car noise. *IEEE Signal Process. Lett.* **1999**, *6*, 259–261. [CrossRef]
5. Chen, S.-H.; Wu, H.-T.; Chang, Y.; Truong, T.K. Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator. *Pattern Recognit. Lett.* **2007**, *28*, 1327–1332. [CrossRef]
6. Evangelopoulos, G.; Maragos, P. Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. ASLP* **2006**, *14*, 2024–2038. [CrossRef]
7. McAualy, R.J.; Malpass, M.L. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 137–145. [CrossRef]
8. Kim, N.S.; Chang, J.-H. Spectral enhancement based on global soft decision. *IEEE Signal Process. Lett.* **2000**, *7*, 108–110.
9. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1109–1121. [CrossRef]
10. Ramirez, J.; Segura, J.C. An effective subband OSF-based VAD with noise reduction for robust speech recognition. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 1119–1129. [CrossRef]
11. Kim, S.K.; Chang, J.H. Voice activity detection based on conditional MAP criterion incorporating the spectral gradient. *Signal Process.* **2012**, *92*, 1699–1705. [CrossRef]
12. ITU-T. *Appendix III: G.729 Annex B Enhancement in Voice-Over-IP Applications-Option 2*; 2005.
13. ITU-T. *Recommendation P.862, Perceptual Evaluatioon of Speech Quality (PESQ), an Objective Method for end-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*; 2001.