An Improved Voice Activity Detection Algorithm Employing Speech Enhancement Preprocessing

Yoon-Chang Lee, Sang-Sik Ahn
Dept. of Electronics and Information Engineering
Korea University, Seoul, Korea
Tel:+82-415-860-1792, Fax:+82-415-860-1356
E-mail: ychlee@hard.korea.ac.kr , sahn@tiger.korea.ac.kr

Abstract: In this paper we derive a new VAD algorithm, which combines the preprocessing algorithm and the optimum decision rule. To improve the performance of the VAD algorithm we employ the speech enhancement algorithm and then apply the maximal ratio combining technique in the preprocessing procedure, which leads to maximized output SNR. Moreover, we also perform extensive computer simulations to demonstrate the performance improvement of the proposed algorithm under various background noise environments.

1. Introduction

In many speech coding applications such as CDMA wireless networks and packetized communication systems variable rate transmission is preferred. To reduce the average bit rate and the required transmission bandwidth by assigning less or no bits in the absence of speech, the speech coders in such systems usually employ a voice activity detector (VAD). In a typical conversation, each speaker talks for less than 40 % of the time and the remaining is silence. When the VAD algorithm is employed and discontinuous transmission is in operation, the transmitter is switched off if silence is detected. Utilizing this technique, mobile communication systems can increase the system capacity by reducing the required bandwidth and the transmitting power.

In systems where the background noise level is very low, a simple signal energy level detecting algorithm can be used to detect the silence regions. The most well known VAD algorithm of this kind is the G.729B VAD [1]. On the other hand, in systems where a large and varying background noise is present, it is impossible to distinguish noisy speech from background noise by using the simple energy level detection, thus a more intelligent algorithm is required. To this ends, a statistical model based algorithm has been proposed [2] and is further optimized by employing the decision-directed method for the estimation of the unknown parameters [3]. They show significantly better performance then that of the G. 729B in low signal-to-noise ratio (SNR) and various noise environments. However, they have following limitations and problems:

1) Performance of [2] is limited due to the bias
2) Decision-directed method for the estimation of the *a priori* SNR demands a lot of computations

In this paper we derive an improved VAD algorithm, which combines the preprocessing algorithm and the optimum decision rule to alleviate the aforementioned limitations and the problems. In the preprocessing procedure we employ the speech enhancement algorithm and apply the maximal ratio combining technique. The enhanced signal can be obtained by using the simple power subtraction method, and the maximal ratio combining technique [4] is utilized to maximize the output SNR, which improves the performance of the VAD.

This paper is organized as follows. In section 2, we start with an energy based VAD and move on a statistical model based VAD. We then propose an improved VAD algorithm employing the new speech enhancement preprocessing in section 3, and summarize the simulation results in section 4 to demonstrate the performance improvement of the proposed VAD algorithm under various background noise conditions. Finally, we discuss and conclude in section 5.

2. VAD Algorithms

A VAD algorithm in a mobile environment must detect a speech signal in the presence of a wide range of diverse background noises. In the very low SNR environment, it is impossible to distinguish speech from noise using the simple energy level based detection techniques when parts of the speech utterance are buried below the noise. The distinction under these conditions can be practical only by considering the spectral characteristics of the input signal. If the background noise is stationary over a relatively long period, the spectral characteristics of the noise will be very similar from frame to frame and then, in principle, it is possible to detect the presence of speech by monitoring deviation from the spectral characteristics of the background noise.

The VAD used for the discontinuous transmission of the GSM system incorporates an inverse filter, and the filter coefficients are derived during noise only periods [5]. When a noisy speech signal is applied, the noise signal is attenuated by the filter and the most residual energy is caused by the spectral characteristics of the speech. The operation of the VAD is to compare the average residual energy of the inverse filtered signal with the threshold value and can be formulated as follows [2].

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

H_0 : speech absent : $x_n = v_n$

H_1 : speech present : $x_n = s_n + v_n$

where n, s_n, and v_n represent a frame index, clean speech, and noise signal vectors, respectively. Then, the decision rule of this type of algorithm is given by

$$\frac{1}{M}\sum_{k=0}^{M-1}\frac{|X_k|^2}{|\hat{V}_k|^2} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \tag{1}$$

where $|\hat{V}_k|^2$ is the estimated noise power and X_k is a $k-th$ element of M dimensional discrete Fourier transform (DFT) coefficient vector of noisy speech.

The decision statistic of Eq. (1) can be identified as the average of M frequency bin SNRs. This is almost the same as the statistic proposed by Yang [6]. He used the average subband *a priori* SNR by power subtraction method. Note from Eq. (1) that we need to increase the SNR to improve the performance of the VAD algorithm based on the energy level.

On the other hand, in systems where a large and varying background noise is present, it is impossible to distinguish noisy speech from background noise by using the simple energy level detection, thus a more intelligent algorithm is required. To this ends, a statistical model based algorithm has been proposed [7] and derived as follows.

Assuming that the DFT coefficients of each process are asymptotically independent Gaussian random variables, joint probability density functions on two hypotheses H_0 and H_1 are given by [8]

$$P(X \mid H_0) = \prod_{k=0}^{M-1}\frac{1}{\pi|\hat{V}_k|^2}\exp\left\{-\frac{|X_k|^2}{|\hat{V}_k|^2}\right\}, \tag{2}$$

$$P(X \mid H_1, |\hat{S}_k|^2) = \prod_{k=0}^{M-1}\frac{1}{\pi(|\hat{S}_k|^2+|\hat{V}_k|^2)}\exp\left\{\frac{-|X_k|^2}{|\hat{S}_k|^2+|\hat{V}_k|^2}\right\}. \tag{3}$$

Now, applying the maximum likelihood estimation for $|\hat{S}_k|^2$, we get $|\hat{S}_k|^2 = |X_k|^2 - |\hat{V}_k|^2$, then the generalized log likelihood ratio test can be formulated as

$$\Lambda = \frac{1}{M}\log\frac{p(X \mid H_1, |\hat{S}_k|^2)}{p(X \mid H_0)} = \frac{1}{M}\sum_{k=0}^{M-1}\left\{\frac{|X_k|^2}{|\hat{V}_k|^2} - \log\frac{|X_k|^2}{|\hat{V}_k|^2} - 1\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda. \tag{4}$$

Again, note here that the performance of the VAD algorithm can be improved by increasing the SNR.

3. Proposed VAD Algorithm

One way of enhancing the speech signal in an additive acoustic noise environments is to perform a spectral decomposition of a frame of noisy speech and to attenuate a particular spectral line, depending on the power difference between the noisy speech and the noise. Under the same assumptions as in section 2, we perform the spectral decomposition using the DFT and enhance the speech signal by employing the power subtraction method [9]. That is,

$$\hat{s}_n = \frac{1}{M}\sum_{k=0}^{M-1}\hat{S}_k^0 \exp\left\{j\frac{2\pi}{M}kn\right\}, \tag{5}$$

where

$$\hat{S}_k^0 = \sqrt{|X_k|^2 - |\hat{V}_k|^2}\,\frac{X_k}{|X_k|}. \tag{6}$$

Now, the signals $\hat{S}_k^0$ from each of M frequency bins are individually weighted to provide the maximum SNR using the maximal ratio combining technique. If the signal from each frequency bin has been weighted by a_k then the resulting signal applied to the detector is

$$\zeta_M = \sum_{k=0}^{M-1} a_k \hat{S}_k^0 \tag{7}$$

and the magnitude of the $\hat{S}_k^0$ can be approximated as follows:

$$\sqrt{|X_k|^2 - |\hat{V}_k|^2} = \sqrt{|S_k|^2 + |V_k|^2 - |\hat{V}_k|^2} \approx |S_k| + w_k, \tag{8}$$

where w_k is a residual noise. We assume that, after speech enhancement, the residual noise power σ_w^2 in each frequency bin is the same. Therefore, the total noise power $\sigma_{w_T}^2$ applied to the detector is simply the weighted sum of the noise power in each frequency bin, i.e.,

$$\sigma_{w_T}^2 = \sigma_w^2 \sum_{k=0}^{M-1}|a_k|^2. \tag{9}$$

Then the output SNR γ_M applied to the detector becomes

$$\gamma_M = \frac{\zeta_M^2}{\sigma_{w_T}^2}. \tag{10}$$

Now, by Schwartz inequality, γ_M is maximized when $a_k = \hat{S}_k^{0*}/\sigma_w$ and reduced to

$$\gamma_M = \sum_{k=0}^{M-1}\frac{|\hat{S}_k^0|^2}{\sigma_w^2} = \sum_{k=0}^{M-1}\gamma_k. \tag{11}$$

Thus the output SNR of the maximal ratio combiner is the sum of the SNR in each frequency bin. Recall from section 2 that, without the maximal ratio combining, the output SNR is simply the average of M frequency bin SNRs. Then, the enhanced and weighted signal in each

frequency bin S_k^1 becomes

$$|\hat{S}_k^1| = |X_k|^2 - |\hat{V}_k|^2 . \quad (12)$$

Next, in order to maintain an identity in the absence of noise (in this case $a_k = 1$ for all k), the input phase is appended and then the speech-enhanced signal $\hat{S}_k^2$ becomes

$$\hat{S}_k^2 = \left(|X_k|^2 - |\hat{V}_k|^2 \right) \frac{X_k}{|X_k|} . \quad (13)$$

Fig. 1 shows the performance of conventional speech enhancement algorithm and the proposed algorithm under babble noise conditions.

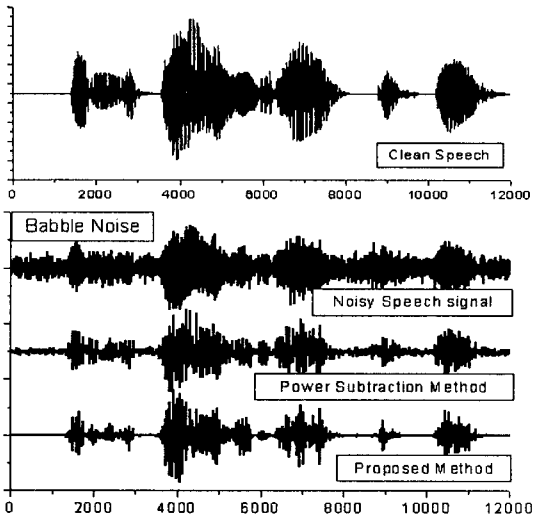

Fig. 1. Speech Enhanced Signal

Finally, by applying the suggested speech enhancement preprocessing to the statistical model based likelihood ratio test we propose a new VAD algorithm, which is summarized in Fig. 2.

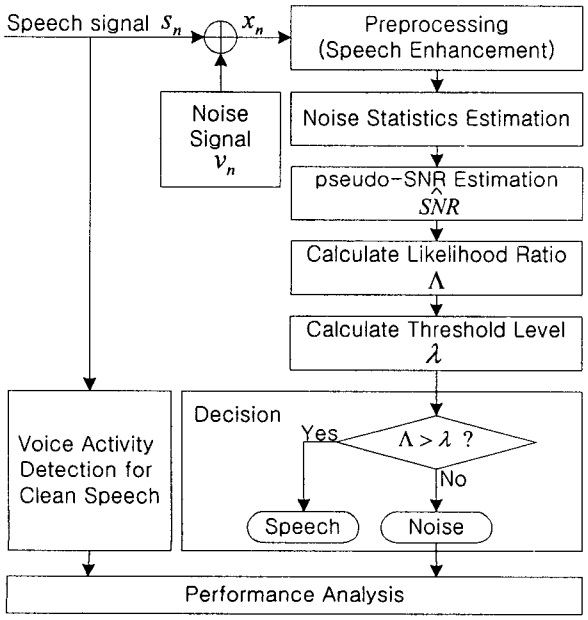

Fig. 2. Block Diagram of the Proposed VAD Algorithm

In the above algorithm we used the following moving average technique to estimate the noise statistics;

$$\hat{V}_k(n) = \alpha \hat{V}_k(n-1) + (1-\alpha)|X_k(n)| , \quad 0 \le \alpha < 1 \quad (14)$$

where $|X_k(n)|$ and $\hat{V}_k(n)$ denote the spectral magnitude of the measured signal and the estimated noise at frame n in frequency bin k, respectively, and α is a forgetting factor. If the statistical characteristics of the noise change slower than those of the speech, then the forgetting factor will be close to '1', otherwise away from '1'. We need a secondary VAD since the noise statistics are updated during noise only periods. It detects the noise only periods using the energy level in each frame. If the energy level of the current frame is comparable with that of the previous frame, the secondary VAD decides that this frame is the noise only frame, otherwise the speech frame. In practice, it is a difficult to make a correct decision especially when the SNR is low. But this problem can be overcome by adjusting the forgetting factor properly.

On the other hand, we compute the threshold level λ in the decision rule using the mean ($\overline{\Lambda}$) and standard deviation (σ_Λ) of Λ by assuming that speech is absent during the initial period as follows:

$$\lambda = \overline{\Lambda} + \gamma \, \sigma_\Lambda \quad (15)$$

where γ is a parameter that affects the performance. The clipping probability is direct proportional and the false alarm probability is inverse proportional to γ. Therefore, γ may be adapted in inverse proportion to the SNR to get optimum performance.

4. Computer Simulation

We performed extensive computer simulations to evaluate the performance of the proposed algorithm under various background noise environments. Noise and speech samples were drawn from NOISEX-92 dB [10] and Si Pro Laboratory [11], respectively. The speech signals were digitized at 8 kHz sampling frequency with 16 bits resolution and the noise was added to the clean speech signal to produce a noisy signal at the specific SNR. For computational savings, we took 128 points fast Fourier transforms with 30 ms-long Hamming windowed signals.

It is common in modern VAD algorithms to use a hangover period of a few frames to delay any pre-mature transitions from speech to noise. This is to minimize the clipping probability especially for a low-level unvoiced speech signal. The basic function of the hangover is to delay the decision by a few frames by monitoring the VAD decisions in neighboring frames. We assigned four frames to the hangover period.

The proposed VAD algorithm in Fig. 2 works on a frame-by-frame basis. The VAD accepts 128 frames of the input speech signal to initialize the VAD and

estimates the noise statistics. We assume these initial frames as noise only frames even if they actually contain speech signals. With Table 1 and 2, we demonstrate that the proposed algorithm has a better performance than those of [1]-[3] in a wide SNR range and various noise environments. We obtained the results in Table 1 with the estimated noise statistics and those in Table 2 with the assumption that the noise statistics were known.

Table 1. VAD results with noise estimation
P_D: Detection Probability, P_T: Total Error Probability

Environment		Proposed VAD		G.729B [1]	
Noise	SNR	P_D[%]	P_T[%]	P_D[%]	P_T[%]
White	-5 dB	91.98	13.68	48.20	56.00
	0 dB	94.50	11.05	62.50	32.80
	5 dB	95.75	7.80	64.14	34.87
	15 dB	99.20	5.95	85.33	16.30
	25 dB	99.31	5.50	96.43	7.13
Babble	-5 dB	98.40	19.82	56.50	61.50
	0 dB	98.98	18.84	62.80	48.20
	5 dB	99.30	18.20	79.56	53.07
	15 dB	99.83	17.48	89.09	36.38
	25 dB	100.00	16.00	97.37	26.53
Vehicular	-5 dB	92.70	24.70	87.25	56.50
	0 dB	95.60	19.80	91.50	51.50
	5 dB	97.30	19.60	94.29	52.23
	15 dB	99.15	18.85	98.24	25.56
	25 dB	99.90	15.40	99.81	42.98

Table 2. VAD results with known noise statistics
P_D: Detection Probability, P_T: Total Error Probability

Environment		Proposed VAD		Sohn et al. [3]	
Noise	SNR	P_D[%]	P_T[%]	P_D[%]	P_T[%]
White	-5 dB	94.50	9.50	N/A[1]	N/A
	0 dB	97.80	7.50	N/A	N/A
	5 dB	98.91	1.00	84.58	16.76
	15 dB	99.97	0.40	96.93	6.34
	25 dB	100.0	0.20	99.87	5.30
Babble	-5 dB	95.00	8.50	N/A	N/A
	0 dB	96.50	6.50	N/A	N/A
	5 dB	98.63	1.50	93.04	30.14
	15 dB	99.50	0.40	98.43	25.37
	25 dB	100.0	0.10	99.75	25.00
Vehicular	-5 dB	97.80	5.50	N/A	N/A
	0 dB	99.00	3.20	N/A	N/A
	5 dB	98.80	1.50	97.30	7.54
	15 dB	99.00	1.00	99.62	7.57
	25 dB	100.0	0.80	99.87	5.30

5. Conclusions

We derived a new VAD algorithm, which combines the preprocessing algorithm and the optimum decision rule to alleviate the reported limitations and the problems. In the preprocessing procedure we employed the speech enhancement algorithm and applied the maximal ratio combining technique. The enhanced signal is obtained by using the simple power subtraction method, and the maximal ratio combining is employed to improve the performance of the VAD algorithm further by maximizing the output SNR. Then we conducted extensive computer simulations. The simulation results summarized in Table 1 and Table 2 show that the proposed VAD algorithm outperforms the G. 729B and [3] over the wide range of SNR under various background noise environments. Moreover, the presented VAD algorithm requires comparatively low computational loads. Thus it may be implemented by consuming a portion of the computational power of a single digital signal processor for mobile and portable wireless communication systems. Furthermore, the suggested VAD algorithm can be applied to the internet telephony or teleconference environments, since the common noises in a house or office are white or babble noises.

Future studies may include the application of the adaptive idea to adjust the threshold level to changing environments for a further performance improvement.

References

[1] ITU-T Recommendation, "G.729 Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70", Nov. 1996.

[2] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation", Proc. Int. Conf. ASSP, 1998, pp 365-368.

[3] J. Sohn and W. Sung, "A statistical model-based voice activity detection", IEEE signal processing letters, Vol 6, No 1, Jan. 1999.

[4] T. S. Rappaport, "Wireless Communications," Prentice Hall, 1996, pp 328-330.

[5] P. A. Barrett, "Information tone handling in the half-rate GSM voice activity detector" Communications, 1995. ICC '95 Seattle, 'Gateway to Globalization', 1995 IEEE International Conference on Volume: 1, 1995 , Page(s): 72 -76 vol.1.

[6] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems", in Proc. Int. Conf. ASSP, 1993, pp 363-366.

[7] R. J. Mcaulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, pp. 137 – 145, April 1980.

[8] W.A. Pearlman and R.M. Gray, "Source coding of the discrete Fourier transform," IEEE Trans. Inform., Theory, vol. IT-23, pp. 683-692, Nov. 1978.

[9] 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

[10] http://spin.rice.edu, Rice Univ. DSP Group.

[11] ftp://ftp.sipro.com, Si Pro Lab. Telecom Inc.

[1] N/A : Not Available