Robust Voice Activity Detector under the Noise Environment in G.723.1 Vocoder

KYUNGA JANG*, SOYEON MIN**, MYUNGJIN BAE*

*Department of Information & Telecommunication Engineering
** Department of Electronic Engineering
University of Soongsil, KOREA

ABSTRACT

Generally the one of problems in Voice Activity Detection (VAD) is speech region detection in noise environment. Therefore, this paper proposes the new method using the energy, lsp varation. Processing time and speech quality of the proposed algorithm as a result, the processing time is reduced due to the accurate detection of inactive period, and there is almost no difference in the subjective speech quality test. Proposed algorithm measures the number of VAD=1 and the results shows predominant reduction of bit rate as SNR of the noisy speech is low (about 5~10dB).

Key-Words : Voice Activity Detection (VAD), Linear Spectrum Pairs(LSP)

I. INTRODUCTION

Information technology area, two fields such like multimedia and mobile communication, have improved dramatically and widely at a point of social view up to recently. These areas have a purpose that the data is transmitted under approximate condition with various ways. Recently, remarkable improving a digital mobile communication and personal communication network makes that the speech coding technique has studied for applying optimum services in both side of amount, quality and increasing more users. In order to increase the users in communication networks areas, one of the effective methods is the decrement of the transmission rate of vocoder. It is possible to allow double users when the transmission rate of vocoder is decreased half in digital cellular network. Specially, the increment of the quantity of internet phone and PCS(Personal Communication Service) consumed are supposed to have more attention to the CELP (Code Excited Linear Prediction) vocoder. One of CELP vocoder, G.723.1 coder, is designed to the internet phone and net meeting which has two-bit rate associated with it, 5.3 and 6.3kbit/s depending on the transmission circumstance. It provides system designers with additional flexibility. The excitation signal for the high rate coder is MP-MLQ (Multipulse Maximum Likelihood Quantization) and for the low rate coder is ACELP (Algebraic-Code-Excited Linear-Prediction)[3]. G.723.1 and G.729 standardized in 1996 for using digital cellular service at home recently has been used the VAD(Voice Activity Detection) for decreasing the transmission rate during silent intervals of speech. Especially, VAD in G 723.1 is to indicate whether each 30 msec frame produced by the speech encoder contains speech or not and uses the energy as a parameter for detecting. Thus the energy threshold is decided depending on the pitch value of constant frames and detecting the sine waveform. Generally, The purpose of the VAD is to reliably detect the presence or absence of speech in any other background noise and characteristic of the spectrum and periodicity of input speech signal is used for the processing. However, this parameter doesn't give a influence directly to decide that the present period in speech signal is voiced signal or not so it is the much difficult for making the accurate decision about low SNR (Signal to Noise Rate) signal. The reminder of this paper is organized as follows. The problem of VAD in G.723.1 annex A dual rate vocoder is briefly described before. In Section2, the simple decision theory as an algorithm is proposed to detect the robust voice activity period with parameters such as the energy, LSP coefficients and the average of AWGN (A White Gaussian Noise) in this paper. The simulation result of the proposed algorithm showed with figures and tables in Section 3 and concluded in Section 4.

2 ROBUST VOICE DECTECTOR ALGORITHM IN THE NOISE ENVIRONMENT

2.1 VAD algorithm using energy, LSP(Line spectrum pair) and average of the Noise.

Voice activity detector for decreasing the bit rate in silence period uses many parameters for the safety and the continuity in decision as explained above. Beside it uses spectrum

characteristic for making the accurate decision about low SNR signal. However this method hasn't only a difficulty in exact detecting until the real threshold value reaches the current energy level but also doesn't contribute to decrease the bit rate that low SNR signal as input is detected as the voiced frames even the silence frames[3][6]. Therefore, this paper proposed a method of decreasing the bit rate by more exact detection of the silence frames with maintaining the routine of the conventional method. The parameters used in this paper are the energy, the pitch gain and the average of signal's magnitude. The energy threshold is estimated that we regard the first three frames in the input signal as the silence period and the variance threshold is estimated by using the energy and the variance of silence period. After that the rate is estimated between the energy of a frame and the energy of the signal which of being over an even magnitude. If the ratio is over the energy threshold, the period is the voice activity region otherwise the variance of the noise and LSP uses for the detection as regards that period again. If the value of analysis period is over the variance threshold value as compared the variance threshold value with the value of analysis period, the analysis period is detected as the voice reigon. Otherwise the average of analysis period's magnitude is over the average of silence period as the threshold value estimated. The average of silence period sets as the threshold because of the fact that the energy of white gaussian noise is zero. If the analysis period is decided as the silence period that the variance threshold updates in order to estimate adaptively the variation of LSP and repeats the above process again. So, the computation time of finding parameters can be reduced. Therefore if the proposed method is used, we can reduce not only transmit rate but also processing times

2.2 Voice activity detection.

2.2.1 The case of energy ratio over energy threshold.

The energy ration is as follows.

$$EneRatio = \frac{\sum_{n=0}^{N-1} s_L^2(n)}{\sum_{n=0}^{N-1} s^2(n)}, \quad s_L(n) = \begin{cases} s(n), & s(n) > EneTh \\ 0, & otherwise \end{cases} \qquad (1)$$

Where $s(n)$ is the speech signal and $s_L(n)$ denotes like above. *EneTh* denotes the energy threshold. is the speech signal and It is possible to reduce the error that the noise region is regarded as a voice activity region in low SNR signal if the ratio uses for the detection of voice activity region. In case of energy of ratio in analysis period over the energy threshold VAD sets 1 (one).

2.2.2 The case of the energy ratio not over the energy threshold.

If the energy ratio of analysis period isn't over the energy threshold we consider two cases. Variance threshold updates if current frame is the first part of frame or the end point of the speech signal and voice activity region is detected by using the LSP variance and the average of signal if the current frame includes voiced/unvoiced region.

2.2.2.1 The case of the first part of frame or the end point of speech signal.

In order to set the threshold by using the variance and the energy of noise first we estimate the average distribution of LSP about A white gaussian noise.

Figure 1 shows the LSP distribution table by 10^{th} linear prediction analysis about noise signal. A number of 10^{th} LSP distribution have a regular interval in figure 1. Average of 10^{th} LSP is shown in table 1. First, average threshold of LSP estimates as regarding with the first three frames of input speech signal. We regard the first three frames as the noise. The difference between LSP of the current input speech signal and average value of LSP is estimated and added to the average value of LSP. This processing reflects the characteristic of current noise signal and equation (2) denotes as follows. This is the updated average LSP value of the noise.

$$T_{LSP}(i) = Ave_{LSP}(i) + \frac{1}{K}\sum_{k=0}^{K-1}(V_{LSP}(k,i) - Ave_{LSP}(i)) \qquad (2)$$

V_{LSP}: currentframeLSP, $K=3$, $i=0,1,\cdots,P-1$(P:$LPCorder$)

$Ave_{LSP} = [360, 740, 1080, 1460, 1820, 2200, 2540, 2920, 3280, 3660]$

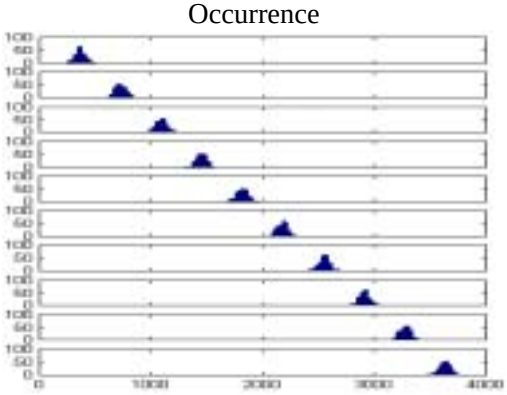

Fig.1 10^{th} LSP distribution of a white gaussian noise

Table 1. 10^{th} average LSP of A white gaussian noise

LSPth	1	2	3	4	5	6	7	8	9	10
Frequency (Hz)	360	740	1080	1460	1820	2200	2540	2920	3280	3660

$$\text{var}_{LSP} = \frac{1}{P}\sum_{i=0}^{P-1}(T_{LSP}(i) - V_{LSP}(i))^2 \qquad (3)$$

V_{LSP} : *current LSP*

var_{LSP} is the variance threshold and the parameter for detecting the voice activity region. This case represents the variance of a white gaussian noise. $T\,\mathrm{var}_{LSP}$ is estimated by using var_{LSP} and the average of first three frames. This value makes the decision adaptively about the noise as its updating every time the silence period starts[4-5].

$$T\ \mathrm{var}_{LSP} = \frac{\frac{1}{K}\sum_{i=0}^{K-1} \mathrm{var}_{LSP}(i)}{E} + B$$
$$E = R \times \frac{1}{N}\sum_{j=0}^{N-1} s(j)^2 \qquad (4)$$

2.2.2.2 Maintaining the continuance of decision.

The average of signal's magnitude in analysis period estimates and if its value is over the threshold *MThr* , which is closed to 0 (zero), the analysis period is decided as the silence region. Otherwise var_{LSP} estimates each frame and compares with $T\,\mathrm{var}_{LSP}$. If var_{LSP} is over the $T\,\mathrm{var}_{LSP}$, the analysis period is decided as the voice activity region. In order to maintain the continuance of decision we use a VCnt (counter) as a variable.

$$VAD = \begin{cases} 1, & , \ Vcnt > 0 \\ 0, & , \ Vcnt = 0 \end{cases} \qquad (5)$$

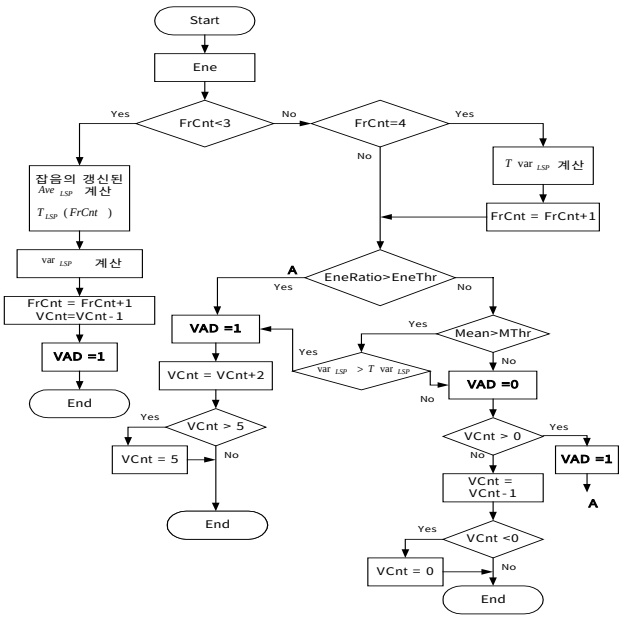

Fig. 2 Blockdiagrame of proposed algorithm

Figure 2 descirbes the blockdiagrame of proposed algorithm.

3. Experiment and result.

We used the system interfaced AD/DA convertor used the general-purposed processor in common for simulating the algorithm. The sampling rate of input speech signal is 8kHz and quantization rate is on 16 bits. A frame length is 240 samples on each speech signal and sub-frame length is 60 samples. To test the performance of operating result, AWGN (A white gaussian noise) and the real noise in laboratory is added to the utterances.

In this paper, we used those utterances to make the SNR of the utterances is each -5, 0, 10, 20dB that AWGN and the noise got through the mike in real laboratory, both of two added to clean speech respectably. Moreover the silent period is put longer in each syllabol on purpose and we see the result of detecting in silent period. Utterances are gotten by male and female speakers from various ages.

Utterance1: /Insune komaneun cheonjae sonyuneul joahanda/
Utterance2: /Yesunimkkesuh chunjichangchoeu keohuneul malssumhasuhtda./
Utterance3: /Changongul haecheo nagaun inganu dojuneun kkuteeupda./
Utterance4:/Soongsildahakgyojungbotonshingwaumsungtong-shin teameeda./
Utterance5: /Gongileesamsaohyukchipalgu/

C languages simulated the proposed algorithm. The result of efficiency is compared to speech signals got from G.723.1 Annex A and from proposed algorithm.

The number of VAD=1 detected speech signal among all frames of input signal is compared to measure the decrement of transmission rate and clock function provided in C language is used to measure the decrement of processing time.

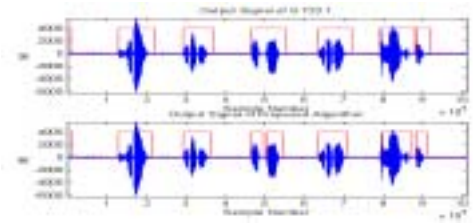
Fig.3 Clean speech (male speaker);
/Insune komaneuncheonjae sonyuneul joahanda/
(a) G.723.1 algorithm, (b) Proposed algorithm.

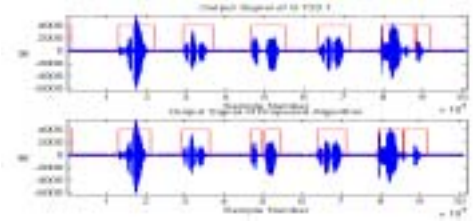
Fig. 4 Clean speech (male speaker) + AWGN(20dB);
/Insune komaneun cheonjae sonyuneul joahanda/
(a) G.723.1 algorithm, (b) Proposed algorithm.

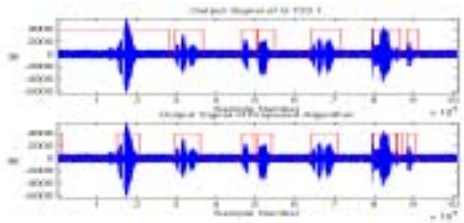

Fig. 5 Clean speech (male speaker) + AWGN(5dB); /Insune komaneun cheonjae sonyuneul joahanda/ (a) G.723.1 algorithm, (b) Proposed algorithm.

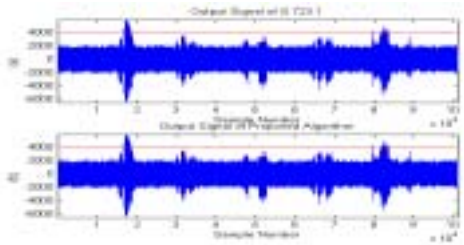

Fig. 6 Clean speech (male speaker) + AWGN(-5dB); /Insune komaneun cheonjae sonyuneul joahanda/ (a) G.723.1 algorithm, (b) Proposed algorithm.

The processing time decrement estimated with clock function from C algorithm. The part of dotted line in Fig 3~6 presents the part of detecting the voice activity period. After performing VAD the period got 0 (zero) which means non-speech period and dotted-line also represents 0 as well but in case of got 1 (one) in period, it means there's speech period. In case of clean speech got from male's uttered the utterance 1, conventional G.723.1 and proposed algorithm, both of two doesn't have the difference. In time of 0dB speech, transmission rate of proposed algorithm is increased but in case of synthesizing with using CNG (Comfort Noise Generator) about the large noise level in silent period, much distortion of speech signal is caused. So, determining the silent with large noise as voice activity period can be avoided dramatic distortion of speech signal. Otherwise, the transmission rate is decreased depending on the level of noise. In fig 5, proposed algorithm can be detected as a result that the voice activity period and large noise period with clean speech added the noise which got real laboratory. On the other hand, conventional algorithm can't be detected. Table 1 represented the comparison of the number of the frames which detected VAD=1 as voice activity period. As the energy of noise over SNR 5dB speech signal is large, the efficiency of decreasing the transmission rate is much. Table 2 and 3 describe the comparison of the processing time that in case of detecting the voice activity algorithm region (VAD=1). The reduction efficiency of processing time is higher as much as the noise in speech signal over 5dB. In case of the speech got SNR about 5dB~20dB, proposed algorithm has better result than conventional G.723.1 VAD algorithm in the side of bit rate reduction as well as processing time decrease. Also the quality of speech after performed by proposed algorithm is the same as much as one by conventional algorithm in subjective evaluation of speech quality but on the other hand, bit rate is decreased.

Table 2. The number of active speech frame (VAD=1) (Utterance 2. male speaker, total frames=367, voice activity region frame = 152)

	Clean	-5dB	0dB	5dB	10dB	20dB	Laboratory noise
G.723.1 VAD Algorithm	170	358	253	235	213	173	232
Proposed VAD Algorithm	166	358	358	137	154	152	150
Decrement (%)	2.35	0	-41.5	41.7	27.7	12.1	35.3

Table 3. The number of active speech frame (VAD=1) (Utterance 3. female speaker, total frames=424, voice activity region frame = 202)

	Clean	-5dB	0dB	5dB	10dB	20dB	Laboratory noise
G.723.1 VAD Algorithm	231	263	252	239	243	252	296
Proposed VAD algorithm	235	207	167	187	197	225	193
Decrement (%)	-1.73	21.3	33.7	17.3	18.9	10.7	34.8

Table 4. Comparison of the processing time, U=Utterance (sec, in case of adding the background of laboratory noise)

	U 1	U 2	U 3	U 4	U 5
G.723.1 VAD Algorithm	8.92	10.36	10.57	10.32	11.93
Proposed VAD algorithm	7.89	9.69	9.68	9.37	10.79
Decrement (%)	6.89	6.46	8.58	9.21	9.55

Table 5. Comparison of objective speech quality, U=Utterance (MOS, in case of adding the background of laboratory noise)

	U 1	U 2	U 3	U 4	U 5	Average
G.723.1	3.7	3.8	3.7	3.9	3.8	3.78
Proposed algorithm	3.7	3.7	3.7	3.9	3.8	3.76

4. CONCLUSIONS

G.723.1 6.3/5.3Kbps dual rate speech codec developed for internet phone and net meeting among CELP(Code-Excited Linear Prediction) series speech vocoders has used VAD (Voice Activity Detection) for decreasing the bit rate during the silent period. For synthesizing the speech signal only the minimum parameters transmits when the frames is detected as silent period through voice activity detecting so the reduction of bit rate and of processing time for the real time can be followed. This paper proposed a method of decreasing the bit rate by more exact detection of the silence frames with maintaining the routine of the conventional method. The parameters used in this paper are the energy, the pitch gain and the average of signal’s magnitude. The energy threshold is estimated that we regard the first three frames in the input

signal as the silence period and the variance threshold is estimated by using the energy and the variance of silence period. After that the rate is estimated between the energy of a frame and the energy of the signal which of being over an even magnitude. If the ratio is over the energy threshold, the period is the voice activity region otherwise the variance of the noise and LSP uses for the detection as regards that period again. If the value of analysis period is over the variance threshold value as compared the variance threshold value with the value of analysis period, the analysis period is detected as the voice reigon. Otherwise the average of analysis period's magnitude is over the average of silence period as the threshold value estimated. The average of silence period sets as the threshold because of the fact that the energy of white gaussian noise is zero. If the analysis period is decided as the silence period that the variance threshold updates in order to estimate adaptively the variation of LSP and repeats the above process again. So, the computation time of finding parameters can be reduced. Therefore if the proposed method is used, we can reduce not only transmit rate but also processing times. As a experiment result, the optimum efficiency about bit rate reduction is obtained when SNR of the speech signal is about 5~10dB and about the processing time proposed algorithm has more efficient about average 8% than conventional one. On subjective quality test there is almost no difference compared with the G.723.1 dual rate vocoder.

References:

[1] A.M. Kondoz, "Digital Speech-Coding for Low Bit Rate Communication System, 1994

[2] W.B. Kleijn and K.K. Paliwal, "Speech Coding and Synthesis", ELSEVIER SCIENCE B.V., pp. 6-10, 1995

[3] ITU-T Recommendation G.723.1, March, 1996

[4] B.J.Min, B.J.Kang, "End point detection using by the distance of LSP in EVRC packet", *The jounal of the Acouastic society of Korea*. vloume8, Oct. 1999.

[5] JJ.Kim, KA.Jang, MJ.Bae, "G.723.1 Voice Activity Detector using by LSP coefficients and energy", Signal processing workshop. Oct.,1999,

[6] DS.Na, MJ.Bae, "Transition period detection by pitch synchronization" *The Acouastic Society of Korea*, Speech comm., and signal processing workshop. Vol.15, No.1, PP.454-459, 1998.