2000年度 大韓電子工學會 夏季綜合學術大會 論文集 第 23 卷 第 1 號 2000/6

잡음 환경에서의 전송율 감소를 위한 G.723.1 VAD 성능개선에 관한 연구

김정진, 박영호, 배명진

서울특별시 동작구 상도5동 1-1
숭실대학교 정보통신공학과
mjbae@saint.soongsil.ac.kr

Tel : 02)820-0016~17, Fax : 02)820-0018

The Research of Reducing the Fixed Codebook Search Time of G.723.1 MP-MLQ

JeongJin KIM, YoungHo PARK, MyungJin BAE,

1-1 SangDo5-Dong DongJak-Ku Seoul
Dept. of Info. and Telecomm. Engr., Soongsil Univ

mjbae@saint.soongsil.ac.kr

Abstract

On CELP type Vocoders G.723.1 6.3kbps/5.3kbps Dual Rate Speech Codec, which is developed for Internet Phone and videoconferencing, uses VAD (Voice Activity Detection)/CNG (Comfort Noise Generator) in order to reduce the bit rate in a silence period. In order to reduce the bit rate effectively in this paper, we first set the boundary condition of the energy threshold to prevent the consumption of unnecessary processing time, and use three decision rules to detect an active frame by energy, pitch gain and LSP distance.

To evaluate the performance of the proposed algorithm we use silence-inserted speech data with 0, 5, 10, 20dB of SNR. As a result when SNR is over 5dB, the bit rate is reduced up to about 40% without speech degradation and the processing time is additionally decreased.

I. 서 론

CELP 계열 음성 부호화기 중에서 인터넷 폰 및 화상 통신 등을 위해 개발된 G.723.1은 묵음 구간에서의 전송률을 감소시키기 위해 VAD(Voice Activity Detection)를 사용하고 있다. VAD 판정에 의해 묵음으로 간주된 프레임에 대해서는 신호합성에 필요한 최소한의 파라미터만을 전송하게 되므로 전송률 감소뿐만 아니라 처리시간 감소 효과도 더불어 얻을 수 있다.

따라서 본 논문에서는 미리 설정된 에너지 임계값을 사용하여 아주 큰 에너지를 갖거나 아주 낮은 에너지를 갖는 경우에는 간단한 결정논리를 사용하고 현재 프레임의 에너지가 에너지 임계값을 넘는 경우에는 LSP 거리값과 피치이득 값을 이용하여 음성활동 유무판정을 하는 알고리즘을 제안한다.

II. G.723.1 VAD 알고리즘

VAD의 목적은 음성 부호화기에 의해 생성된 30ms의 각 프레임에 대해 음성의 존재 유무를 판정하는 것이다. VAD는 기본적으로 에너지 검출기이다. 역 필터링된 신호의 에너지는 문턱값과 비교되어지고 이 문턱값을 넘는 경우 그 프레임에는 유/무성음이 존재하는 것으로 판정하고 그렇지 않은 경우 잡음이 존재하는 프레임으로 판정한다. 문턱값을 계산하기 위해서는 두 과정이 필요하다. 첫째, 잡음 레벨은 이전 프레임의 잡음 레벨과 현재 프레임에서 역 필터링된 신호의 에너지에 근거하여 갱신된다. 둘째, 문턱값은 로그스케일의 잡음 레벨을 이용하여 계산한다.

II.1 Adaptation enable flag computation

현재 프레임 t에 대해 Aen_t로 표기되는 Adaptation enable flag는 VAD 잡음 레벨이 유/무성음 신호도 아니고($pc \neq 4$) 정현파도 아닌 경우($\sin D = 0$)에만 갱신되도

록 하기 위해 사용된다.

- **Adaptation enable flag 계산**

$$\begin{cases} Aen_t = Aen_{t-1} + 2 & \text{, if } pc = 4 \text{ or } SinD = 1 \\ Aen_t = Aen_{t-1} - 1 & \text{, otherwise} \end{cases} \quad (1)$$

Aen_t는 [0,6]을 경계조건으로 한다.

II.2 잡음 레벨 계산

1) 만약 $Nlev_{t-1} > Enr_{t-1}$이면 잡음 레벨은 클리핑된다.

$$Enr_t = \frac{1}{180}\sum_{j=60}^{239} e'^2_t[n] \quad (3)$$

$$Nlev_t = \begin{cases} 0.25Nlev_{t-1} + 0.75Enr_{t-1} & \text{, if } Nlev_{t-1} > Enr_{t-1} \\ Nlev_{t-1} & \text{, otherwise} \end{cases} \quad (4)$$

2) 만약 adaptation이 활성되면 $Nlev_t$는 증가되고 그렇지 않으면 조금씩 감소된다.

$$Nlev_t = \begin{cases} 1.03125 \times Nlev_t & \text{, if } Aen_t = 0 \\ 0.9995 \times Nlev_{t-1} & \text{, otherwise} \end{cases} \quad (5)$$

$$with \begin{cases} Nlev_{\min} = 128 \\ Nlev_{\max} = 131071 \end{cases}$$

II.3 문턱값 계산 및 VAD 결정

프레임 t에서의 잡음 레벨, $Nlev_t$, 문턱값, Thr, 사이의 관계는 로그 스케일로 정의되고 다음과 같은 공식을 이용한다.

$$Thr = \begin{cases} 5.012 & \text{, if } Nlev_t = 128 \\ 10^{0.7 - 0.05\log_2 \frac{Nlev}{128}} & \text{, if } 128 < Nlev < 16384 \\ 2.239 & \text{, if } Nlev \geq 16384 \end{cases} \quad (6)$$

VAD결정은 문턱값, Thr와 현재 에너지, Enr_t의 비교에 의해 결정된다.

$$Vad_t = \begin{cases} 1 & Enr_t \geq Thr \\ 0 & Enr_t < Thr \end{cases} \quad (7)$$

III. 제안한 알고리즘

III.1 에너지, 피치 이득 및 LSP계수를 이용한 VAD 알고리즘

묵음 구간에서의 전송률을 낮추는 위한 G.723.1 VAD는 판정의 안정성과 연속성을 위해서 여러 가지 파라미터를 사용하고 있으며 SNR이 낮은 신호에 대한 정확한 판정을 위해 스펙트럼 특성을 이용하고 있다. 하지만 위와 같은 방법은 실제 임계값이 현재의 에너지 레벨에 도달할 때까지 정확한 판정을 하는 것이 불가능하며 또한 SNR이 낮은 신호가 입력으로 들어올 때 묵음이 존재하는 프레임에 대해서도 음성이 존재하는 프레임으로 판별함으로써 전송률을 낮추는데 기여하지 못하고 있다. 따라서 본 논문에서는 묵음이 존재하는 프레임을 보다 정확히 검출함으로써 전송률을 낮추는 방법에 대해 제안한다.

본 논문에 쓰이는 파라미터는 에너지, 피치 이득 및 LSP 계수이다. 입력 신호의 처음 3프레임 동안은 묵음이라는 가정하에 에너지 임계값을 구하고 묵음 구간에서의 평균 LSP 계수들을 설정하게 된다. 그런 다음 음성활동 구간 판정을 3가지 경우로 나누어 판정을 수행하게 된다. 먼저 현재 프레임의 에너지가 에너지 최대 임계값을 넘는 경우와 에너지 임계값을 넘지 않는 경우에는 각각 음성활동 구간과 음성 비활동 구간으로 판정하고 현재 프레임의 에너지가 에너지 임계값을 넘는 경우에는 SNR이 낮은 신호의 경우를 고려하기 위해 피치 이득과 LSP 파라미터를 이용하게 된다.

또한 묵음구간으로 판정된 프레임에 대해서는 합성을 위한 최소한의 파라미터만을 전송하게 되므로 파라미터 추출을 위한 계산시간을 단축할 수 있게 된다. 따라서 본 논문에서 제안한 방법을 이용할 경우 전송률 감소 효과뿐만 아니라 부가적으로 처리시간 감소 효과도 얻을 수 있게 된다.

III.2. 에너지 임계값 및 묵음 구간의 LSP 계수 설정

본 논문에서는 처음의 입력 3프레임 동안은 묵음이라는 가정하에 3 프레임 동안 다음과 같이 평균에너지와 평균 LSP계수들을 구하게 된다.

$$Ene_i = \sum_{n=0}^{N-1} s_{i,n}^2[n]/N, \quad i = 0,1,2 \quad (8)$$

$$NLSP_k = \sum_{i=0}^{2} LSPvect_{i,k}, \quad k = 1,2,\ldots,10 \quad (9)$$

여기서 N은 240이며 $s_t[n]$은 현재 프레임 t의 입력신호이며$LSPvect$는 현재 프레임에서 구한 LSP 계수들이다. 위의 파라미터를 이용하여 다음과 같은 에너지 임계값과 묵음의 평균 LSP 계수들을 계산한다.

$$EneThr = mean(Ene) + 1.3 \times StdDev(Ene) \quad (10)$$

$$LSPave_k = NLSP_k/3, \quad k = 1,2,\ldots,10 \quad (11)$$

III.3 음성의 존재 유무 판별

본 논문에서는 음성의 존재 유무 판정을 위해 크게 3가지로 나누어 판정을 한다. i) 현재 프레임 t에서 구한

에너지가 최대 에너지 임계값을 넘는 경우 ii) 현재 프레임 t에서 구한 에너지가 에너지 임계값을 넘지 않는 경우 iii) 현재 프레임 t에서 구한 에너지가 에너지 임계값을 넘는 경우로 경우로 나누어 판정하며, i), ii)의 경우 각각 음성이 활동하는 프레임, 음성이 활동하지 않는 프레임으로 판정을 한다. 마지막 iii)의 경우 입력 신호가 낮은 SNR을 갖는 경우를 고려하기 위해 피치 이득과 LSP 거리 파라미터를 이용하여 판정을 수행한다. 즉 에너지가 임계치를 넘는다고 하더라도 SNR이 낮은 신호의 경우 묵음 구간에 존재하는 잡음에 의한 경우를 배제하기 위해 피치 이득과 LSP 거리가 미리 설정된 각각의 피치이득 임계치와 LSP 거리 임계치를 넘는 경우에만 음성이 존재하는 것으로 판정한다.

III.3.1 $Ene_t > MaxThr$인 경우

피치이득이나 LSP 거리와 상관없이 항상 음성활동 구간으로 설정한다(VAD=1). 또한 *EneThr*은 다음과 같이 갱신된다.

$$EneThr_t = EneThr \cdot (1025/1024) \tag{12}$$

III.3.2 $Ene_t < EneThr$인 경우

묵음구간으로 설정한다(VAD=0). 또한 *EneThr*은 다음과 같이 갱신된다.

$$EneThr_t = EneThr_{t-1} \cdot (31/32) \tag{13}$$

III.3.3 $Ene_t \geq EneThr$인 경우

III.3.3.1 피치 이득 계산

피치 이득은 다음과 같이 구한다.

$$\beta_t = \frac{C_{\max}}{Ene_t} \tag{14}$$

여기서 C_{max}는 다음식의 C_b를 최대로 하는 값이다.

$$C_b(j) = \frac{(Cor(j))^2}{\sum_{n=0}^{239} s_t[n-j] \cdot s_t[n-j]}, \quad 18 \leq j \leq 142 \tag{15}$$

$$Cor(j) = \sum_{n=0}^{N-1} s_t[n] \cdot s_t[n-j], \quad 18 \leq j \leq 142 \tag{16}$$

III.3.3.2 LSP 거리 계산

묵음 구간의 LSP 계수들 사이에는 일반적으로 등간격을 가지고 있지만 음성이 존재하는 경우는 포만트가 위치하는 주파수영역에 LSP 계수들이 많이 존재하는 특징이 있다. 즉 묵음구간에서 구한 LSP 계수들과 음성이 존재하는 LSP 계수들 사이의 오차를 구하면 그 값이 크게되지만 묵음구간의 LSP 계수들 사이의 오차는 상당히 적게 된다. 따라서 LSP 계수들 사이의 오차를 이용하면 음성의 존재유무를 판정할 수 있게 된다. LSP 계수들 사이의 거리는 다음과 같이 구할 수 있다.

$$LSPDist = \sqrt{\sum_{k=1}^{10} \{LSP_t(k) - LSPave(k)\}^2} \tag{17}$$

III.3.3.3 Ene_t가 *EneThr*보다 크거나 같은 경우의 음성활동 검출

위에서 구한 피치이득 값과 *LSPdist* 값이 미리 설정된 각각의 임계값보다 작은 경우 묵음 구간으로 그렇지 않은 음성활동 구간으로 설정하게 된다.

$$VAD = \begin{cases} 0, & \text{if } b < bthr \text{ and } LSPdist < LSPThr \\ 1, & otherwise \end{cases} \tag{18}$$

III.4 헹오버

판정의 지속성을 위해 본 논문에서는 다음과 같은 기능을 추가하였다.

$$Vcnt = \begin{cases} Vcnt+2, & \text{if } Ene_t \geq EneThr \\ Vcnt-1, & Ene_t < EneThr \end{cases} \tag{19}$$

비록 제안한 알고리즘이 묵음구간으로 판정하더라도 판정의 급격한 변화를 방지하기 위해 Vcnt가 0보다 큰 경우에는 음성활동 구간으로 설정하게 된다.

$$VAD = \begin{cases} 1, & \text{if } Vcnt > 0 \\ 0, & \text{if } Vcnt = 0 \end{cases} \tag{20}$$

IV. 실험 및 결과

제안한 알고리즘을 실험하기 위한 장비는 IBM-PC 586(333MHz)에 상용화된 AD/DA 컨버터를 인터페이스한 시스템을 사용하였다. 입력신호는 G.723.1의 입력신호와 같이 8kHz로 표본화하고 16bit로 양자화한 음성을 입력으로 하며 각 음성시료에 대해 한 프레임의 길이를 240표본, 부프레임의 길이를 60샘플로 하여 처리하였다. 처리결과의 성능을 측정하기 위해 다음의 음성 시료에 대해 AWGN과 실제 실험실 환경의 잡음을 첨가하여 사용하였다. 본 논문에서는 음성시료의 SNR이 각각 -5, 0, 5, 10, 20dB가 되도록 AWGN을 Clean Speech에 첨가한 음성과 실제 실험실 환경에서 접할 수 있는 잡음을 마이크를 통하여 입력 받은 다음 Clean Speech에 더한 음성시료를 사용하였다. 또한 묵음구간의 검출 결과를 알아보기 위해 각 어절마다 묵음구간을 길게 삽입하여 묵음의 검출여부를 관찰하였으며 대표적인 문장을 연령층이 다른 남녀화자가 발성한 음성을 음성시료로 사용하였다.

발성1: /인수네 꼬마는 천재소년을 좋아한다./
발성2: /예수님께서 천지창조의 교훈을 말씀하셨다./
발성3: /창공을 헤쳐 나가는 인간의 도전은 끝이 없다./
발성4: /숭실대학교 정보통신과 음성통신 연구팀이다./
발성5: /공일이삼사오육칠팔구/

제안한 알고리즘의 시뮬레이션은 C-언어로 구현하여

수행하였다. 성능 비교는 G.723.1 Annex A를 통과한 음성과 제안한 알고리즘을 통과한 음성을 비교하였다. 전송율 감소량을 측정하기 위해 전체 프레임 중에서 VAD=1로 판정한 프레임 수를 비교하였으며 처리시간 감소량 측정은 C 알고리즘이 제공하는 clock 함수로 측정하였다. 실험 환경은 MP-MLQ 모드에서 실험실 환경의 잡음이 첨가된 음성을 사용하여 5번 측정한 다음 평균값을 사용하였다. 음질 측면에서는 MOS test를 사용하였다.

V. 결 론

CELP 계열 음성 부호화기 중에서 인터넷 폰 및 화상통신을 위해 개발된 G.723.1 6.3 Kbps/5.3 Kbps 이중 전송률 음성 코덱은 묵음 구간에서의 전송률을 감소시키기 위해 VAD(Voice Activity Detection)를 사용하고 있으며, VAD 판정에 의해 묵음으로 간주된 프레임에 대해서는 신호합성에 필요한 최소한의 파라미터만을 전송하게 되므로 전송률 감소뿐만 아니라 실시간 구현을 위한 처리시간 감소 효과도 더불어 얻을 수 있다.

어떠한 배경잡음에 대해서도 음성신호를 판별하기 위해 G.723.1에서는 스펙트럼 특성과 입력신호의 주기성을 이용하고 있다. 하지만 이런 파라미터는 판정에 직접적으로 영향을 미치는 것이 아니기 때문에 이런 과정을 거치는 것으로 SNR이 낮은 신호에 대해 정확한 판정을 한다는 상당히 어렵다.

따라서 본 논문에서는 처리시간의 불필요한 소모를 방지하기 위해 미리 설정된 에너지 임계값을 사용하여 아주 큰 에너지를 갖거나 아주 낮은 에너지를 갖는 경우에는 간단한 결정논리를 사용하고 현재 프레임의 입력신호의 에너지가 경계조건 안에 해당하는 경우 잡음 구간에서 구한 LSP 계수와 입력신호에서 구한 LSP 계수 사이의 거리값을 이용하여 스펙트럼 특성을 고려하고 피치이득 값을 이용하여 주기성을 고려하여 음성활동 유무판정을 하는 알고리즘을 제안한다.

실험 결과 SNR이 5dB에서 10dB 사이의 음성신호의 경우 최적의 전송율 감소 효과를 얻을 수 있었으며 처리시간의 비교 결과 평균 7% 정도의 처리시간 감소효과를 얻을 수 있었다. 주관적 음질 평가의 결과 음질열하는 거의 발생하지 않았다.

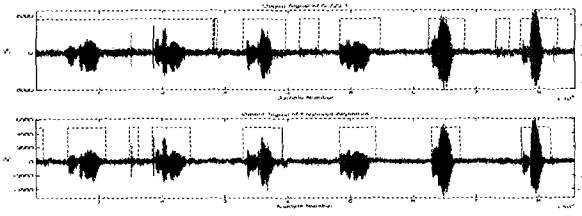

그림 1. Clean Speech + 실험실 환경의 잡음
(SNR 8dB, 남성화자, 음성시료 1)
(a) G.723.1 알고리즘 (b) 제안한 알고리즘

표 1. VAD=1로 판정한 프레임의 수
(전체 358 Frame, 음성 시료 1, 남성화자)

	Clean	-5dB	0dB	5dB	10dB	20dB	실험실 잡음
G.723.1 VAD 알고리즘	169	358	250	231	203	171	245
제안한 VAD 알고리즘	166	358	358	139	144	150	149
감소율 (%)	1.7	0	-43.2	39.8	29.1	12.3	39.2

표 2. 처리시간의 비교(Sec)

	발성 1	발성 2	발성 3	발성 4	발성 5
G.723.1 VAD 알고리즘	8.90	10.26	10.69	10.22	11.93
제안한 VAD 알고리즘	7.89	9.59	9.83	9.32	11.80
감소율 (%)	11.3	6.5	8.1	8.8	1.1

표 3. 음질의 비교 (실험실 잡음을 첨가한 경우)

	발성 1	발성 2	발성 3	발성 4	발성 5	평균
G.723.1	3.7	3.8	3.7	3.9	3.8	3.78
제안한 알고리즘	3.7	3.7	3.7	3.9	3.8	3.76

참고문헌

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

[2] A.M. Kondoz, *"Digital Speech"*,, John Wiley & Sons, 1994.

[3] 나덕수, 정찬중, 박영호, 배명진, "LSP를 이용한 음성신호의 성분분리에 의한 CELP 보코더의 전송률 감소에 관한 연구", 한국음향학회, 학술발표대회논문집, 1999, 8월

[4] N. S. Jayant and P. Noll, Digital Coding of Waveform-Principles and Applications to Speech and Video, pp.220-221, Prentice-Hall, 1978.

[5] W. B. Klejin et. al, *"Speech Coding and Synthesis"*, Elsevier Science B.V., 1995.

[6] 배 명진, "디지탈 음성부호화," 동영출판사, 1997.