이동통신부호화기에서의 음성 활동 검출 장치 성능에 관한 연구

임지선*
*(주)에이치씨티
e-mail:jisun722@nate.com

A Study on Performance of Voice Activity Detector in Vocoder

Ji-Sun Lim*
*HCT. Co., Ltd.

요 약

ITU-T에서 인터넷 폰과 화상회의에 사용하기 위하여 개발된 G.723.1 음성 부호화기는 잡음 구간에서의 전송률을 낮추기 위한 방법으로 VAD(Voice Activity Detector)와 CNG(Comfort Noise Generator)를 사용하고 있다. 여기서 VAD는 최종적으로 현재 프레임의 에너지 레벨을 비교하여 음성의 활동 유무를 판정하고 있다. 하지만 G.723.1 VAD에서는 보다 안정적인 판정을 위해 음성 활동 구간 사이에 삽입되어 있는 묵음 구간에 대해서는 거의 대부분 음성이 활동하는 영역으로 판정을 하고 있다.
본 논문에서는 묵음 구간에 대해 보다 정확한 판정을 통하여 기존의 방법에 비해 전송율을 더욱 감소시킬 수 있는 방법을 제안한다. 실험에서는 묵음구간을 길게 조절한 문장을 사용하여 측정한 결과 약 50% 정도의 전송율을 감소시킬 수 있었으며, MOS 테스트 결과, 음질의 열하는 발생하지 않았다.

1. 서론

최근에 디지탈 이동통신 및 유선망을 통한 화상회의, 인터넷폰 사용자의 증가로 급증하는 통신 가입자를 보다 많이 수용하고 서비스의 질을 높이기 위해 음성부호화기에 대한 많은 연구가 진행되고 있다. 이 중 가입자의 수용율을 증가시키는데 있어서 효과적인 방법 중의 하나가 보코더의 전송율을 낮추는 방법이다. 이론적으로 디지탈 셀룰라망인 경우 보코더의 전송율이 1/2로 낮아지면 기존 대비 약 2배의 가입자를 수용할 수 있다고 알려져 있다[1][2]. 그러나 전송율 감소에 따른 음질의 저하는 통신 서비스 질에 대한 불만을 불러일으킬 수 있으므로 전송율과 음질이라는 두 지표를 적정수준으로 맞추는 것이 중요한 문제가 되고 있다.

유선망을 이용한 화상회의 및 인터넷폰을 목적으로 ITU-T에서 표준화된 G.723.1은 묵음구간에서의 전송율을 낮추기 위하여 VAD(Voice Activity Detector)와 CNG(Comfortable Noise Generator)를 사용하고 있다. 이 중 VAD는 현재 프레임의 음성 활동 유무를 판정하여 CNG 알고리즘에 정보를 제공한다. 일반적으로 VAD 알고리즘은 보다 안정적이고 연속적인 결정을 위해 지속적인 프레임의 정보를 이용하고 있다[2]. 그러나, VAD는 안정성과 연속적 판별을 위해 신호가 처음 시작되는 부분에서는 거의 모든 프레임에 대해 1로 설정을 하게 된다. 따라서 본 논문에서는 보다 효율적인 VAD 판별을 통해 묵음 구간에서의 전송율을 더욱 낮추는 방법을 제안한다.

2. G.723.1 VAD 알고리즘

VAD의 목적은 30ms의 각 프레임에 대해 음성의 존재 유무를 판정하는 것이다. VAD는 기본적으로 에너지를 이용하여 검출한다. 역 필터링된 신호의 에너지를 문턱값과 비교하고 이 문턱값을 넘는 경우 그 프레임에는 음성이 존재하는 것으로 판정하고 그렇지 않은 경우 묵음 구간으로 판정한다.

현재 프레임 t에 대해 Adaptation enable flag는 VAD 잡음 레벨이 음성 신호도 아니고 정현파도 아닌 경우에만 갱신되도록 하기 위해 사용된다.

- 유/무성음 검출

이전과 현재 프레임의 개회로 피치 지연을 유성음 판정을 위해 사용한다. 이 값을 $L_{OL}^{j}, j=0,1,2,3$ 이라고 할 때 $L_{OL}^{\min}=Min(L_{OL}^{j}, j=0,1,2,3)$을 먼저 계산한다. 그런 다음 계수기 $pc \in [1,2,3,4]$에서 $L_{OL}^{\min}(\pm 3)$ 배수의 주위에 얼마나 많은 지연 L_{OL}^{j}이 존재하는지를 계산한다. 만약 pc가 4라면 그 신호는 유성음으로 판정된다.

- 정현파 검출

정현파 검출은 LPC 분석기 내에 포함된 $k_t'[2]$가 마지막 15개 값들 중에서 최소한 14개 값이 $k_t'[2]>0.95$라면 정현파가 검출되는 것으로 판정한다 (SinD=1). 그렇지 않은 경우 SinD=0이 된다.

- 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.

입력 신호 프레임, $\{s[n]\}_{n=60..239}$는 계수. $\{a_{no}[j]\}_{j=1..10}$를 갖는 FIR 필터 Ano(z)에 의해 역필터링된다. 이 필터는 CNG 블록에 의해 계산되어지고 현재 프레임의 배경 잡음과 관련된 LPC 필터를 제공한다.

$$e'_t = s[n] + \sum_{j=1}^{10} a_{no}[j] \cdot s[n-j] \quad n=60 \rightarrow 239 \quad (2)$$

여기서 e'_t는 역 필터링된 신호이다.

프레임 t의 잡음레벨, $Nlev_t$는 이전의 잡음레벨과 이전의 에너지, Enr_{t-1}, 그리고 adaptation enable flag, Aen_t에 의해 갱신된다. 이런 갱신 과정은 느린 증가, 빠른 감소로 특징되어진다. 프레임 t에서의 잡음 레벨의 동적 범위는 $[Nlev_{\min}, Nlev_{\max}]$으로 제한된다.

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

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

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

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

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

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

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.

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

3. 제안한 방법에 의한 음성 활동 구간 검출

본 논문에서는 유성음 구간을 먼저 검출한 다음 유성음이 아닌 구간에 대해 무성음/묵음 구간 판정을 통하여 최종 음성 활동 구간 설정을 한다. 이를 위한 파라미터로 LSP, 에너지, b 계수, ZCR을 사용한다.

유성음 구간은 현재 프레임의 에너지 값이 에너지 문턱값보다 큰 경우 유성음 구간으로 결정하고 VAD=1로 설정하게 된다. 에너지 문턱값은 이전 5개의 유성음 프레임에서 구한 에너지 값을 이용하여 갱신된다. 본 논문에서는 Log 스케일링된 에너지 값을 사용하였다.

$$Ene_t = 10 * \log_{10}\left(\sum_{n=0}^{239} s_t^2[n]\right) \quad (8)$$

여기서 $s_t[n]$은 현재 프레임 t에서의 음성 신호이다. 만약 현재 프레임 t에서의 에너지 값이 에너지 문턱값보다 작은 경우 피치 이득 β를 이용하여 β 값이 미리 설정된 문턱값을 넘는 경우 유성음 구간

으로 결정하고 VAD=1로 설정한다. 본 논문에서는 다음과 같은 방법으로 β를 구하였다.

$$\beta = \frac{C_{\max}}{E} \tag{9}$$

여기서 $C_{\max}$은 다음 식의 C_b를 최대로 하는 값이고 E는 현재 프레임의 에너지 값이다.

$$C_b(j) = \frac{(Cor(j))^2}{\sum_{n=0}^{239} s_t[n-j] \cdot s_t[n-j]}, \quad 18 \le j \le 142 \tag{10}$$

$$E = \sum_{n=0}^{239} s_t^2[n] \tag{11}$$

$$Cor(j) = \sum_{n=0}^{239} s_t[n] \cdot s_t[n-j], \quad 18 \le j \le 142 \tag{12}$$

위의 두 파라미터를 사용하여 유성음이 아니라고 판정이 된 경우 ZCR과 LSP 거리를 이용하여 무성음과 묵음 구간을 검출하게 된다.

무성음의 경우 ZCR은 유성음이나 묵음 보다 큰 값을 갖게 되므로 현재 프레임의 ZCR이 미리 설정된 ZCR 문턱값을 넘는 경우 무성음으로 결정하고 VAD=1로 설정한다. 문턱값은 이전 5프레임의 무성음 구간에서 계산된 ZCR을 이용하여 갱신한다.

$$Zcr = \frac{\sum_{n=1}^{239} |sgn[s(n)] - sgn[s(n-1)]|}{2} \tag{13}$$

$$\begin{aligned} sgn[s(n)] &= 1, \quad \text{if } f(n) \ge 0 \\ &= -1, \quad \text{if } f(n) < 0 \end{aligned} \tag{14}$$

만약 위의 모든 경우에 해당되지 않는 경우 최종적으로 현재 프레임의 각 LSP 계수들과 묵음 구간에서 구한 LSP 계수들 사이의 거리를 측정하여 무성음/묵음 판정을 하게 된다.

LSP 계수들은 음성 구간과 묵음 구간에서 서로 다른 분포특성을 가지고 있으므로 이를 이용하여 음성 구간과 묵음 구간을 판정하게 된다[3]. 이를 위해 먼저 묵음으로 판정된 구간에서 구한 LSP 계수들의 평균 값들과 현재 프레임에서 구한 LSP 계수들과의 거리 측정을 하여 그 값이 미리 설정된 문턱값을 넘는 경우 무성음으로 판정하고 VAD=1로 설정한다.

초기화 과정으로 모든 음성신호의 첫 번째 프레임은 묵음으로 가정하고 여기서 구한 LSP 계수 값들을 LSP_{ave} 값으로 한다. 거리 측정은 다음과 같이 한다.

$$Dist = \sqrt{\sum_{i=1}^{10} \{LSP_t(i) - LSP_{ave}(i)\}^2} \tag{15}$$

여기서 $LSP_t(i)$는 현재 프레임 t에서 구한 i번째 LSP 계수이고 $LSP_{ave}(i)$는 묵음 구간에서 구한 LSP 계수들의 평균 값이다. LSP_{ave}는 현재 프레임이 묵음으로 판정될 경우 갱신된다. 제안한 알고리즘에 사용된 각 파라미터는 현재 프레임에 대한 VAD 결정에 따라 각각 다르게 갱신된다. 갱신되는 파라미터는 에너지 문턱값, ZCR 문턱값, LSP_{ave}이다.

현재 프레임이 유성음으로 판정된 경우 에너지 문턱값은 다음과 같이 갱신된다.

$$EneThr = Mean(Ene(i)) - StdDev(Ene(i)) * 0.25 \quad 0 \le i \le 4 \tag{16}$$

여기서 $Ene(i)$는 과거 유성음 구간에서 계산된 Ene 값이다. $Mean$과 $StdDev$는 각각 평균과 표준편차이다.

이렇게 구해진 $EneThr$은 과거의 $EneThr_{old}$와의 조합에 의해 계산된다.

$$EneThr_{new} = 0.85 EneThr_{old} + 0.15 EneThr \tag{17}$$

현재 프레임이 무성음으로 판정된 경우 ZCR 문턱값은 다음과 같이 갱신된다.

$$ZCRThr = Mean(ZCR(i)) - StdDev(ZCR(i)) * 0.25 \quad 0 \le i \le 4 \tag{18}$$

여기서 $ZCR(i)$는 과거 무성음 구간에서 계산된 ZCR 값이다. 이렇게 구해진 ZCR은 에너지 문턱값 계산에서와 동일하게 최종 ZCR 문턱값을 계산한다.

현재 프레임이 묵음으로 판정된 경우 LSP_{ave}는 과거의 묵음 5개 프레임에서 구한 LSP 계수들의 평균 값을 취한다.

$$LSP_{ave}(i) = Mean\left\{ \sum_{j=0}^{4} LSP_j(i) \right\}, \quad 1 \le i \le 10 \tag{19}$$

4. 실험 및 결과

컴퓨터 시뮬레이션에 이용한 장비는 IBM-PC 586(233MHz)에 상용화된 AD/DA 컨버터를 인터페이스한 시스템이다. G.723.1에서는 8kHz로 음성을 표본화한 음성을 입력으로 하며 각 시료에 대해 한 프레임의 길이를 240표본으로 하여 처리하였다. 처리결과의 성능을 측정하기 위해 다음의 대표적인 문장을 연령층이 다양한 남녀 5명의 화자가 각 5번씩 발성하여 시료로 사용하였다.

제안한 알고리즘의 시뮬레이션은 C-언어로 구현하여 수행하였다. 성능 비교는 G.723.1 Annex A를 통과한 음성과 제안한 알고리즘을 통과한 음성의 VAD=1로 판정한 프레임 수를 비교하였으며 음질 측면에서는 MOS test를사용하였다. 표 1에서는 각각의 음성에 대해 VAD=1로 판정한 프레임 수를 나타내고 있으며 표 2에서는 MOS score를 비교한 결과를 나타내고 있다. 실험 결과 약 46.8%의 전송율 감소효과를 얻을 수 있었다. 주관적 음질평가의 경우 음질 열하는 거의 없었다.

5. 결론

G.723은 잡음구간에서의 전송율을 낮추기 위하여 VAD (VoiceActivity Detector) / CNG(Comfortable Noise Generator)를 사용하고 있다. 이 중 VAD는 현재 프레임의 음성 활동 구간 및 묵음 구간을 판정하여 CNG알고리즘에 정보를 제공하는 역할을 하고 있다. VAD의 가장 큰 문제점은 SNR이 아주 낮은 신호에서도 음성 신호의 존재 유무를 정확히 판정해야만 한다는 것이고 이런 문제점을 해결하는 방안으로 스펙트럼상의 특징을 고려하고 있다. 하지만, VAD는 판별의 안정성과 연속성을 위해 음성이 존재하는 구간사이에 삽입된 잡음 구간에 대해서는 거의 모든 경우 1로 설정을 하게 된다. 따라서 본 논문에서는 안정성을 해치지 않는 범위 안에서 음성 활동 구간 및 묵음 구간을 판별하는 알고리즘을 제안하였다. 유성음 판정을 위해서 에너지와 피치 이득 β를 사용하였고 묵음 판정을 위해서는 ZCR과 LSP 파라미터를 사용하였다. 실험 결과 묵음 구간을 고의적으로 길게 발성한 문장을 사용한 경우 약 46.8%의 전송율이 감소하였으며 주관적 음질 평가의 경우 음질 열하는 거의 없었다.

[표 1] VAD=1로 판정한 프레임의 수
(( )안의 수는 전체 프레임 수)

	G.723.1	제안한 알고리즘	감소율(%)
발성 1 (224)	224	100	55.4
발성 2 (322)	297	146	50.8
발성 3 (326)	226	164	27.4
발성 4 (384)	326	181	44.5
발성 5 (374)	281	124	55.9
Total	46.8%		

[표 2] MOS test 결과

	발성1	발성2	발성3	발성4	발성5	평균
G.723.1	3.7	3.63	3.87	3.83	3.75	3.76
제안한 알고리즘	3.65	3.61	3.86	3.80	3.71	3.72

참고 문헌

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

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

[3] 민병준, 강병준, "EVRC패킷에서 LSP거리를 이용한 음성 끝점 검출", 한국 음향학회지 18권 6호, 1999, 8월.

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

[5] L.R.Rabiner, R.W.Schafer, "Digital Processing of Speech Signal", Prentice Hall, 1978.