LSP변화도를 이용한 G.723.1 보코더의 VAD 성능향상에 관한 연구

이 희 원, 나 덕 수, 배 명진
숭실대학교 정보통신공학과
전화 : (02) 824-0906 / 팩스 : (02) 820-0018

Improvement of VAD Performance using the LSP Variation in the G.723.1

HeeWon LEE, Ducksu NA, MyungJin BAE
Dept. of Telecomm. Engr., Soongsil Univ.
mjbae@saint.soongsil.ac.kr

Abstract

ITU-T 국제 표준화 기구에서 인터넷 폰과 화상회의를 목적으로 개발된 G.723.1 음성 부호화기는 잡음 구간에서의 전송률을 낮추기 위한 방법으로 VAD(Voice Activity Detector)와 CNG(Comfortable Noise Generator)를 사용하고 있다. 이중 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.

따라서 본 논문에서는 묵음 구간에 대해 보다 정확한 판정을 통하여 기존의 방법에 비해 전송률을 더욱 감소시킬 수 있는 방법을 제안한다. 제안한 방법은 음성신호와 잡음신호의 LSP 파라미터 간격 정보를 이용하여 음성구간을 검출한다. 묵음구간을 길게 조절한 문장을 사용하여 실험한 결과 VAD=1로 판정한 프레임수가 약 48.98% 감소하였으며 주관적인 음질평가의 경우 음질의 열하는 거의 발생하지 않았다.

Ⅰ. 서 론

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 전송률을 낮추는 방법이다. 이론적으로 디지탈 셀룰라망인 경우 보코더의 전송률이 1/2로 낮아지면 기존 대비 약 2배의 가입자를 수용할 수 있다고 알려져 있다[2][5]. 그러나 전송률 감소에 따른 음질의 저하는 통신 서비스 질에 대한 불만을 불러일으킬 수 있으므로 전송률과 음질이라는 두 지표를 적정수준으로 맞추는 것이 중요한 문제가 되고 있다.

유선망을 이용한 화상회의 및 인터넷폰을 목적으로 ITU-T에서 표준화된 G.723.1은 묵음구간에서의 전송률을 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[5].

하지만, 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 전송률을 더욱 낮추는 방법을 제안한다.

먼저 Ⅱ장에서는 G.723.1의 VAD 알고리즘을 살펴고, Ⅲ장에서는 제안한 VAD 알고리즘을 설명한다. 그 후 Ⅳ장에서는 본 알고리즘을 평가하기 위한 실험 및 결과, Ⅴ장에서 결론을 맺겠다.

Ⅱ. 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[2].

Ⅱ.1 Adaptation enable flag computation

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 $L'_{OL}, j=0,1,2,3$이라고 할 때 $L_{OL}^{\min}=Min(L'_{OL}, j=0,1,2,3)$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}이 존재하는지를 계산한다. 만약 pc가 4라면 그 신호는 유성음으로 판정된다.

- 정현파 검출

정현파 검출은 LPC 분석기 내에 포함된 $k_1[2]$가 마지막 15개 값들 중에서 최소한 14개 값이 $k_1[2]>0.95$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.

Ⅱ.2 역 필터링

입력 신호 프레임, $\{s[n]\}_{n=60..239}$는 계수, $\{a_{no}[j]\}_{j=1..10}$를 갖는 FIR 필터 $A_{no}(z)$에 의해 역 필터링된다. 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 플리핑된다.

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

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

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 VAD 알고리즘

본 논문에서 제안하는 음성구간 검출 알고리즘은 잡음신호와 음성신호의 LSP 차이를 이용하기 때문에 먼저 잡음 신호에 대한 LSP의 전반적인 정보를 알고 있어야 한다. 잡음신호에 대한 LSP 정보를 얻기 위해 화이트 잡음(White Noise)를 충분히 발생시켜 LSP의 평균 분포도를 구한다.

그림 1은 잡음신호에 대해 10차 선형예측분석을 통해 얻은 LSP 분포도이다. 10개의 LSF 분포가 일정한 간격을 유지하고 있는 것을 알 수 있다. 식 8은 10개의 LSP의 평균값을 나타낸다.

음성구간 검출을 위해서 먼저 위에서 구한 평균 LSF

를 이용하여 입력신호의 처음 90msec동안의 신호를 잡음으로 간주하고 LSP를 구하여 평균 LSP값을 갱신한다. 현재 입력신호의 LSP값과 평균 LSP값의 차이를 구하여 그 값을 평균 LSP값에 더한다. 이렇게 함으로써 현재 잡음신호의 특징을 반영한다.

$$Ave_{LSP}=[360,740,1080,1460,1820,2200,2540,2920,3280,3660] \quad (8)$$

$$T_{LSP}(i)=Ave_{LSP}(i)+\sum_{k=0}^{K-1}(V_{LSP}(k,i)-Ave_{LSP}(i)) \quad i=0,1,2,\cdots,P-1 \quad (9)$$

T_{LSP}는 잡음의 갱신된 평균 LSP값이고, V_{LSP}는 현재 프레임의 LSP 값이다. K는 90msec 동안의 프레임 수를 나타내고 k는 현재 처리되는 프레임을, P는 LSP 차수를 나타낸다.

$$var_{LSP}=\frac{1}{P}\sum_{i=0}^{P-1}(T_{LSP}(i)-V_{LSP}(i))^2 \quad (10)$$

var_{LSP}는 음성구간 검출에 사용되는 값으로 평균 잡음신호의 LSP와 현재 프레임의 LSP와의 차이를 나타낸다. 여기서 P는 LSP의 차수이다. 본 논문에서는 P를 10차로 사용하였다. 문턱값은 90msec동안의 에너지와 이때의 var_{LSP}의 평균을 사용하여 다음과 같이 결정하였다.

$$Tvar_{LSP}=300+\frac{\frac{1}{K}\sum_{i=0}^{K-1}var_{LSP}(i)}{E} \quad (11)$$

$$E=R\times\frac{1}{K}\sum_{i=0}^{K-1}(\frac{1}{N-1}\sum_{j=0}^{N-1}s(j)^2)$$

식 11은 실험적으로 얻어진 식이다. $Tvar_{LSP}$는 문턱값이고, K는 90msec 동안의 프레임 수이다. E는 프레임 평균 에너지이고 R은 상수로 0.07을 사용하였다.

그림 2는 제안한 LSP를 이용한 음성구간 검출 방법에 대한 순서도이다. 먼저 8kHz 샘플링에 16비트로 양자화된 입력신호를 240샘플씩 프레임 처리한다. 입력음성은 먼저 LP 분석을 통해 LSP 값으로 계산하게 되고, 처음 3프레임 동안은 잡음신호로 간주하고 잡음신호의 평균 LSP값을 갱신하게 된다. 이 때 이후 프레임은 앞에서 구한 잡음신호의 평균 LSP 값과의 분산을 계산하여 문턱값을 넘을 경우 음성의 시작점을 결정하고 Flag를 1로 변경한다. 시작점이 결정된 이후 VAD 상태를 1로 유지하다가 현재프레임의 파라미터 값이 문턱값을 연속 6번이상 넘으면 VAD를 0으로 변경하여 음성의 끝부분을 결정한다[1].

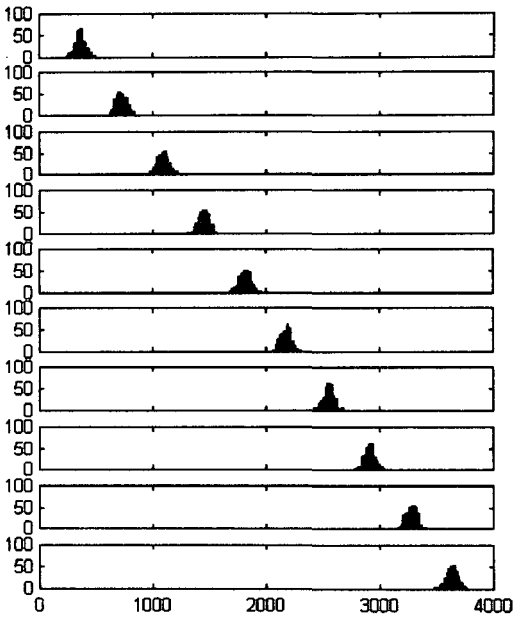

그림 1. 잡음(White Noise)의 10차 LSP 분포도

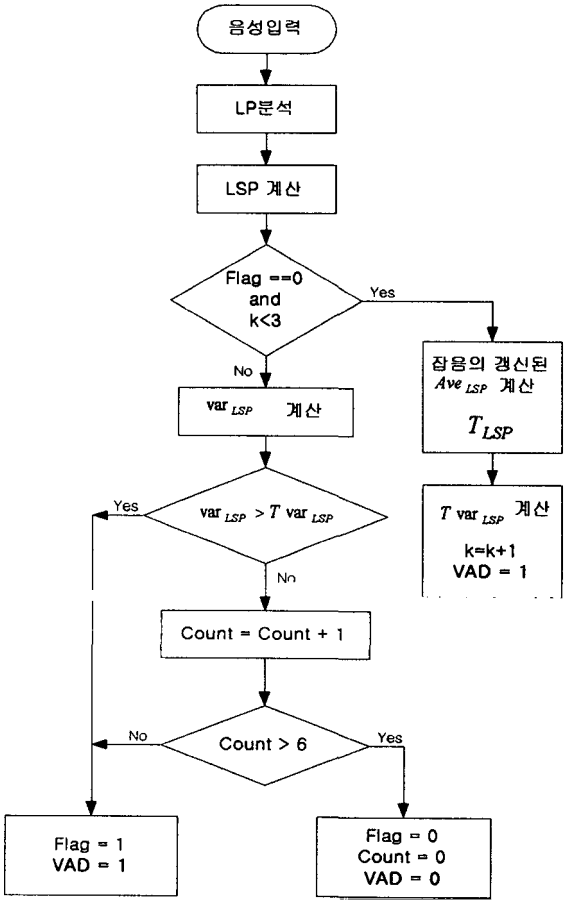

그림 2. 제안한 알고리즘의 순서도

IV. 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. 음성 시료는 각각 묵음구간의 길이를 조절하여 긴 묵음 구간을 갖는 음성으로 SNR이 30dB인 환경하에서 녹음하였다. 음성 시료는 다음과 같다.

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 음성의 VAD=1로 판정한 프레임 수를 비교하였으며 음질 측면에서는 MOS test를 사용하였다. 그림 3은 발성 1을 입력 음성신호로 사용한 경우 묵음구간을 고의적으로 길게 조절한 것에 대한 그림이다. 표 1에서는 각각의 음성에 대해 VAD=1로 판정한 프레임 수를 나타내고 있으며 표 2에서는 MOS score를 비교한 결과를 나타내고 있다. 실험결과 VAD=1로 판정한 프레임수가 약 48.98% 감소하는 효과를 얻을 수 있었다. 주관적 음질평가의 경우·음질 열하는 거의 없었다.

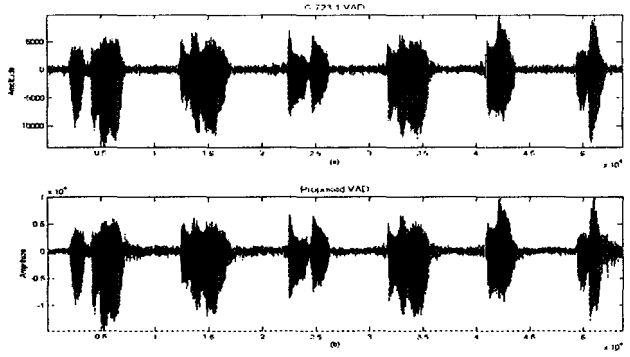

그림 3. 발성 1에 대해 VAD를 통과시킨 음성 신호
(a) G.723.1 VAD를 통과한 음성
(b) 제안한 방법의 VAD를 통과한 음성

V. 결 론

G.723은 잡음구간에서의 전송률을 낮추기 위하여 VAD(Voice Activity 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.

제안하는 음성구간 검출 방법은 잡음신호의 LSF 간격정보와 입력신호의 LSF 간격정보를 비교하여 음성구간인지 결정하게 된다. 잡음신호의 LSF 간격은 비교적 일정한데 반해 음성신호에서는 공명특징으로 인해 LSF의 간격이 음소에 따라 변화하게 된다. 실험 결과 묵음구간을 고의적으로 길게 발성한 문장을 사용한 경우 실험 결과 VAD=1로 판정한 프레임수가 약 48.98% 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 (223)	223	101	54.7
발성 2 (332)	307	158	50.8
발성 3 (326)	226	156	27.4
발성 4 (374)	322	141	56.2
발성 5 (362)	251	125	55.8
Total	48.98		

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] 나덕수, 강은영, 배명진, "LSF를 이용한 음성신호의 끝점 검출 방법", 한국음향학회, 제17회 음성통신 및 신호처리 학술대회 논문집, 2000년.
[2] ITU-T Recommendation G.723.1, March, 1996.
[3] 민 병준, 강 병준, "EVRC패킷에서 LSP거리를 이용한 음성 끝점 검출", 한국 음향학회지 18권 6호, 1999, 8월.
[4] 나덕수, 정찬중, 박영호, 배명진, "LSP를 이용한 음성신호 성분분리에 의한 CELP 보코더의 전송률 감소에 관한 연구", 한국음향학회, 학술발표대회논문집, 1999, 8월.
[5] A.M. Kondoz, "Digital speech", John Wiley & Sons, 1994